Sous la direction de Jean-Pierre Ramis et André Warusfel

Xavier Buff · Josselin Garnier · Emmanuel Halberstadt · François Moulin ·
Monique Ramis · Jacques Sauloy

Mathématiques

Tout-en-un pour la Licence 1

4e édition

DUNOD

Jean-Pierre Ramis, ancien élève de l'École Normale Supérieure de la rue d'Ulm, membre de l'Institut (Académie des Sciences), membre de l'Institut Universitaire de France, membre de l'Académie des Sciences, Inscriptions et Belles-Lettres de Toulouse, professeur émérite à l'Institut de Mathématique de Toulouse (Université Paul Sabatier), a été directeur de l'Institut de Recherches Mathématiques Avancées de Strasbourg et de l'Institut de Mathématiques de Toulouse.

André Warusfel, ancien élève de l'École Normale Supérieure de la rue d'Ulm, a été professeur de mathématiques spéciales au Lycée Louis-le-Grand à Paris et inspecteur général de Mathématiques.

Xavier Buff, ancien élève de l'École Normale Supérieure de la rue d'Ulm, professeur à l'Institut de Mathématiques de Toulouse, ancien directeur de l'Institut de Recherches sur l'Enseignement des Mathématiques de Toulouse.

Josselin Garnier, ancien élève de l'École normale supérieure de la rue d'Ulm, professeur à l'École Polytechnique, Centre de Mathématiques Appliquées.

Emmanuel Halberstadt, a été Maître de conférences à l'Université Paris 6 Pierre et Marie Curie, ancien chargé de cours d'agrégation aux Écoles Normales Supérieures d'Ulm et de Cachan.

François Moulin, ancien élève de l'École Normale Supérieure de la rue d'Ulm, professeur de chaires supérieures au Lycée sainte-Geneviève (spéciales MP*).

Monique Ramis, ancienne élève de l'École Normale Supérieure de Sèvres, a été professeur de chaires supérieures (à Paris, Strasbourg, Toulouse).

Jacques Sauloy, ancien élève de l'École Normale Supérieure de Saint-Cloud, a été maître de conférences à l'Institut de Mathématiques de Toulouse.

Illustration de couverture : © DNYS9 – istock.com

© Dunod, Paris, 2006, 2013, 2018, 2022
ISBN 978-2-10-084670-2
11 rue Paul Bert, 92240 Malakoff
www.dunod.com

e Code de la propriété intellectuelle n'autorisant, aux termes de l'article 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement eservées à l'usage privé du copiste et non destinées à une utilisation collective » , d'autre part, que les analyses et les courtes citations dans un but d'exemple et 'illustration, « toute représentation ou reproduction intégrale ou partielle faite ans le consentement de l'auteur ou de ses ayants droit ou ayants cause est icite » (art. L. 122-4).
ette représentation ou reproduction, par quelque procédé que ce soit, constitue-ait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du ode de la propriété intellectuelle.

Préface

Les mathématiques constituent l'ossature de la science moderne et sont une source intarissable de concepts nouveaux d'une efficacité incroyable pour la compréhension de la réalité matérielle qui nous entoure. Ainsi l'apprentissage des mathématiques est devenu indispensable pour la compréhension du monde par la science. Les nouveaux concepts eux-mêmes sont le résultat d'un long processus de distillation dans l'alambic de la pensée. Essayer de justifier les mathématiques par leurs applications pratiques n'a guère de sens, tant ce processus de création est sous-tendu par la soif de connaître et non l'intérêt immédiat.

Les mathématiques restent l'un des domaines dans lequel la France excelle et ceci malgré la mutilation des programmes dans le secondaire et l'influence néfaste d'un pédagogisme dont l'effet principal est de compliquer les choses simples.

Vues de loin les mathématiques apparaissent comme la réunion de sujets distincts comme la géométrie, qui a pour objet la compréhension du concept d'espace, l'algèbre, art de manipuler les symboles, l'analyse, science de l'infini et du continu, la théorie des nombres etc. Cette division ne rend pas justice à l'un des traits essentiels des mathématiques qui est leur unité profonde de sorte qu'il est impossible d'en isoler une partie sans la priver de son essence. En ce sens les mathématiques ressemblent à un être biologique qui ne peut survivre que comme un tout et serait condamné à périr si on le découpait en morceaux en oubliant son unité fondamentale.

L'une des caractéristiques de l'apprentissage des mathématiques, c'est la possibilité donnée à tout étudiant de devenir son propre maître et en ce sens il n'y a pas d'autorité en mathématiques. Seules la preuve et la rigueur y font la loi. L'étudiant peut atteindre par le travail une maîtrise suffisante pour pouvoir s'il le faut tenir tête au maître. La rigueur, c'est être sûr de soi, et à l'âge où l'on construit sa personnalité, se confronter au monde mathématique est le moyen le plus sûr de construire sur un terrain solide. Il faut, si l'on veut avancer, respecter un équilibre entre les connaissances qui sont indispensables et le « savoir-faire » qui l'est autant. On apprend les maths en faisant des exercices, en apprenant à calculer sans l'aide de l'ordinateur, en se posant des questions et en ne lâchant pas prise facilement devant la difficulté. Seule la confrontation réelle à la difficulté a une valeur formatrice, en rupture avec ce pédagogisme qui complique les choses simples et mélange l'abstraction mathématique avec le jeu qui n'a vraiment rien à voir. Non, les mathématiques ne sont pas un jeu et l'on n'apprend pas les mathématiques en s'amusant.

L'ouvrage qui suit est un cours soigné et complet idéal pour apprendre toutes les Mathématiques qui sont indispensables au niveau de la Licence. Il regorge d'exercices (850) qui

incitent le lecteur à réfléchir et ne sont pas de simples applications de recettes, et respecte parfaitement l'équilibre nécessaire entre connaissances et savoir-faire, permettant à l'étudiant de construire des images mentales allant bien au-delà de simples connaissances mémorisées. Il s'agit d'un ouvrage de référence pour la Licence, non seulement pour les étudiants en mathématiques mais aussi pour tous ceux qui s'orientent vers d'autres disciplines scientifiques. Il insiste sur la rigueur et la précision et va au fond des notions fondamentales les plus importantes sans mollir devant la difficulté et en respectant constamment l'unité des mathématiques qui interdit tout cloisonnement artificiel. Il répond à une demande de tant de nos collègues d'un ouvrage qui les aide à « redresser la barre », mais sera aussi un atout merveilleux pour l'étudiant travaillant seul par la cohérence et la richesse de son contenu.
Il est l'œuvre d'une équipe qui rassemble des mathématiciens de tout premier plan ayant une véritable passion pour l'enseignement. Il était grand temps !

Alain Connes,
Médaille Fields 1982,
Professeur au Collège de France.

Table des matières

Préface iii

Avant-propos xv

I Notations et vocabulaire

I.1 Fondements 3

1 Ensembles . . . 4
1.1 Appartenance, éléments . . . 4
1.2 Définition en compréhension . . . 7
1.3 Constructeurs . . . 8
2 Applications . . . 11
2.1 Applications et graphes . . . 11
2.2 Images et antécédents . . . 13
2.3 L'ensemble $\mathcal{F}(E, F)$ des applications de E dans F . . . 18
3 Suites et familles . . . 19
3.1 Suites d'éléments d'un ensemble . . . 19
3.2 Familles d'éléments d'un ensemble . . . 20
3.3 Familles d'ensembles . . . 21
3.4 Familles de parties d'un ensemble . . . 23
4 Lois de composition . . . 24
4.1 Vocabulaire général . . . 24
4.2 Application au calcul ensembliste . . . 28
5 Relations . . . 28
5.1 Relations binaires sur un ensemble . . . 29
5.2 Relations d'équivalence . . . 31
5.3 Relations d'ordre . . . 33
6 Cardinaux . . . 37
6.1 Induction . . . 37
6.2 Équipotence . . . 39
6.3 Cardinaux finis et cardinaux infinis . . . 42
7 Rudiments de logique . . . 46
7.1 Logique propositionnelle . . . 46
7.2 Prédicats et quantificateurs . . . 50
7.3 Théorèmes et démonstrations . . . 54
EXERCICES . . . 56

II Algèbre

II.1 Arithmétique 61

1 Ensemble des entiers naturels . . . 62
1.1 Relations d'ordre et entiers naturels . . . 62
1.2 Récurrence . . . 63
1.3 Addition et multiplication des entiers naturels . . . 66
2 Dénombrement . . . 70
2.1 Ensembles finis, ensembles dénombrables . . . 70
2.2 Analyse combinatoire . . . 75
3 Divisibilité . . . 80
3.1 Division euclidienne — Numération . . . 80
3.2 Nombres premiers — Factorisation des entiers . . . 83
3.3 Plus grand commun diviseur, algorithme d'Euclide . . . 87
4 Entiers relatifs . . . 91
4.1 Opérations sur les entiers relatifs . . . 91
4.2 Sous-groupes de $\mathbb{Z}$, divisibilité dans $\mathbb{Z}$ . . . 94
5 Nombres rationnels . . . 99
EXERCICES . . . 105

II.2 Groupes, anneaux, corps 111

1 Lois de composition internes . . . 112
2 Groupes . . . 118
2.1 Définitions, règles de calcul . . . 118
2.2 Sous-groupes, morphismes de groupes . . . 120
2.3 Groupe symétrique . . . 125
2.4 Groupe additif des entiers modulo n . . . 130
3 Anneaux . . . 132
3.1 Définitions, règles de calcul . . . 132
3.2 Sous-anneaux, idéaux, morphismes . . . 138
3.3 Divisibilité dans un anneau intègre . . . 143
3.4 Anneau des entiers modulo n . . . 145
4 Corps . . . 151
EXERCICES . . . 154

II.3 Espaces vectoriels et applications linéaires 161

1 Vocabulaire et propriétés élémentaires . . . 162
1.1 La structure d'espace vectoriel . . . 162
1.2 Combinaisons linéaires . . . 165
1.3 Sous-espaces vectoriels . . . 168
2 Applications linéaires . . . 172
2.1 Vocabulaire et exemples . . . 172
2.2 Noyau et image . . . 177
2.3 Quelques applications linéaires particulières . . . 180
2.4 Espaces d'applications linéaires . . . 183
3 Familles de vecteurs . . . 185
3.1 Familles génératrices . . . 185
3.2 Familles libres . . . 187

3.3 Bases 192
3.4 Dimension finie 196
4 Sommes directes et projections 196
4.1 Somme directe de deux sous-espaces vectoriels 196
4.2 Projections 198
EXERCICES 202

II.4 Calcul matriciel élémentaire 207

1 Algèbre matricielle 207
1.1 Définitions et généralités 207
1.2 Matrices carrées 216
1.3 Matrices et applications linéaires 221
2 Opérations élémentaires et algorithmes de Gauß 225
2.1 Opérations élémentaires sur les lignes et sur les colonnes d'une matrice . . . 226
2.2 Algorithmes de Gauß : lignes seules 230
2.3 Algorithmes de Gauß : lignes et colonnes 233
EXERCICES 237

II.5 Le corps des nombres complexes 241

1 Construction et axiomes 241
1.1 Approche axiomatique 242
1.2 Construction effective de $\mathbb{C}$ 243
2 Règles élémentaires de calcul 245
2.1 Représentation cartésienne 245
2.2 Le plan d'Argand-Cauchy 247
2.3 Conjugaison 248
2.4 Module 249
2.5 Racines carrées 253
3 Représentation trigonométrique 256
3.1 Le groupe des nombres complexes de module 1 257
3.2 Racines de l'unité 258
3.3 Arguments d'un nombre complexe 262
3.4 Racines $n^{èmes}$ des nombres complexes 264
3.5 Applications à la trigonométrie 265
4 Quelques applications géométriques 267
4.1 Similitudes planes 267
4.2 Angles de vecteurs et angles de droites 268
4.3 Constructions à la règle et au compas 270
5 Topologie de $\mathbb{C}$ 271
5.1 Rappels sur la convergence dans $\mathbb{C}$ 271
5.2 L'exponentielle complexe 272
5.3 Le théorème de d'Alembert-Gauß 274
EXERCICES 276

II.6 Polynômes et fractions rationnelles 281

1 Polynômes sur un corps quelconque 281
1.1 Construction et axiomes 281
1.2 Règles élémentaires de calcul 283
1.3 Propriétés arithmétiques des polynômes 290

1.4 Fonctions polynomiales et racines d'un polynôme . . . 296
1.5 Polynômes dérivés . . . 301
2 Polynômes sur les corps $\mathbb{R}$ et $\mathbb{C}$ . . . 306
2.1 Applications du théorème de d'Alembert-Gauß . . . 306
2.2 Cyclotomie . . . 307
2.3 Polynômes de Tchebychef . . . 310
2.4 Nombres algébriques . . . 312
3 Fractions et fonctions rationnelles . . . 315
3.1 Le corps des fractions rationnelles . . . 315
3.2 Propriétés arithmétiques de $K(X)$ . . . 318
3.3 Fonctions rationnelles . . . 322
3.4 Développements limités . . . 323
EXERCICES . . . 324

II.7 Espaces vectoriels de dimension finie 331

1 Espaces vectoriels de dimension finie . . . 331
1.1 Définition de la dimension . . . 332
1.2 Applications linéaires en dimension finie . . . 338
2 Applications linéaires et matrices . . . 342
2.1 Écriture matricielle d'une application linéaire . . . 342
2.2 Changements de base . . . 349
3 Déterminants . . . 351
3.1 Déterminant d'une matrice carrée . . . 352
3.2 Mineurs d'une matrice . . . 358
3.3 Déterminant d'un endomorphisme, déterminant d'une famille de n vecteurs . . . 364
3.4 Valeurs propres et vecteurs propres . . . 368
4 Systèmes linéaires . . . 375
4.1 Équations linéaires . . . 375
4.2 Systèmes linéaires . . . 377
EXERCICES . . . 383

II.8 Initiation à l'algorithmique et au calcul formel 391

1 Exemple introductif : l'addition en base b . . . 392
1.1 L'algorithme d'addition . . . 392
1.2 Analyse de l'algorithme d'addition . . . 398
2 Vocabulaire . . . 400
2.1 Langage algorithmique simplifié . . . 400
2.2 Des mathématiques aux algorithmes . . . 404
2.3 Un exemple détaillé : l'algorithme d'Euclide . . . 408
3 Quelques exemples fondamentaux . . . 410
3.1 L'exponentiation dichotomique . . . 410
3.2 Tris et permutations . . . 412
3.3 Polynômes . . . 417
EXERCICES . . . 420

III Géométrie

III.1 Géométrie dans les espaces affines 425

1 Espaces affines . . . 426
1.1 Structure d'espace affine . . . 426
1.2 Barycentres . . . 428
1.3 Sous-espaces affines . . . 429
1.4 Applications affines . . . 432
2 Représentation des sous-espaces affines . . . 436
2.1 Hyperplans . . . 436
2.2 Repère . . . 439
2.3 Systèmes d'équations . . . 440
3 Géométrie affine dans $\mathbb{R}^2$ et dans $\mathbb{R}^3$ . . . 441
3.1 Droites de $\mathbb{R}^2$ . . . 442
3.2 Plans de $\mathbb{R}^3$ . . . 445
3.3 Droites de $\mathbb{R}^3$ . . . 449
3.4 Géométrie euclidienne dans $\mathbb{R}^2$ et $\mathbb{R}^3$ . . . 451
4 Les coniques . . . 456
4.1 Cercles . . . 457
4.2 Coniques . . . 458
4.3 Équations de degré 2 . . . 462
EXERCICES . . . 466

III.2 Courbes paramétrées 469

1 Courbes planes . . . 469
1.1 Notion de courbe paramétrée . . . 469
1.2 Étude locale . . . 471
1.3 Deux exemples . . . 479
2 Courbes en coordonnées polaires . . . 483
2.1 Définition . . . 483
2.2 Tangente . . . 484
2.3 Branches infinies . . . 485
3 Étude métrique d'une courbe plane . . . 486
3.1 Longueur d'une courbe . . . 486
3.2 Paramétrage normal . . . 488
3.3 Courbure . . . 491
3.4 Théorème fondamental . . . 495
4 Courbes de l'espace . . . 497
4.1 Tangente et plan osculateur . . . 498
4.2 Courbure, torsion . . . 500
4.3 Théorème fondamental . . . 502
EXERCICES . . . 502

IV Analyse

IV.1 Nombres réels, suites numériques 509

1 Le corps des nombres réels . . . 510
1.1 Bornes inférieures et supérieures . . . 510
1.2 Le corps des nombres réels . . . 512
1.3 Intervalles de $\mathbb{R}$ . . . 520
2 Suites numériques . . . 522
2.1 Généralités sur les suites . . . 522
2.2 Convergence d'une suite . . . 539
2.3 Cas des suites réelles . . . 555
2.4 Suites bornées . . . 562
2.5 Limites infinies — formes indéterminées . . . 578
3 Un exemple de construction de $\mathbb{R}$ . . . 581
3.1 Écritures décimales . . . 582
3.2 Définition des nombres réels à partir des écritures décimales . . . 588
3.3 Théorème de la borne supérieure . . . 593
3.4 Opérations sur les réels . . . 593
EXERCICES . . . 598

IV.2 Fonctions réelles 619

1 Limites et continuité . . . 619
1.1 Généralités . . . 619
1.2 Limite d'une fonction . . . 621
1.3 Continuité . . . 631
1.4 Théorème des valeurs intermédiaires et image continue d'un segment . . . 634
2 Dérivabilité . . . 640
2.1 Définitions, exemples . . . 640
2.2 Opérations sur les dérivées . . . 642
2.3 Dérivées d'ordre n . . . 644
2.4 Sens de variation et extrema . . . 647
2.5 Théorème de Rolle, accroissements finis . . . 648
2.6 Formules de Taylor . . . 652
2.7 Dérivée de la réciproque . . . 657
2.8 Fonctions convexes . . . 657
3 Étude d'une fonction . . . 668
3.1 Définition et variations . . . 668
3.2 Branches infinies . . . 668
EXERCICES . . . 670

IV.3 Fonctions transcendantes 677

1 Fonctions logarithme et exponentielle . . . 678
1.1 Logarithme népérien . . . 678
1.2 Exponentielle . . . 680
1.3 Représentation graphique des fonctions logarithme népérien et exponentielle . 682
1.4 Logarithmes et exponentielles de base quelconque . . . 682
2 Fonctions racines et puissances . . . 684
2.1 Fonctions racines . . . 684

2.2 Fonctions puissances . . . 684
2.3 Croissances comparées des fonctions puissances, logarithme et exponentielle . 688
3 Fonctions trigonométriques . . . 688
3.1 Fonctions sinus et cosinus . . . 688
3.2 Fonctions tangente et arc-tangente . . . 691
3.3 Fonction arc-tangente . . . 692
3.4 Expressions de $\sin x$ et $\cos x$ en fonction de $\tan \frac{x}{2}$ — Paramétrage du cercle . . 695
3.5 Arc-sinus et arc-cosinus . . . 696
4 Trigonométrie hyperbolique . . . 698
4.1 Sinus, cosinus et tangente hyperboliques . . . 698
4.2 Expressions de $\sinh x$ et $\cosh x$ en fonction de $\tanh \frac{x}{2}$ — Paramétrage de l'hyperbole équilatère . . . 700
4.3 Réciproques des fonctions hyperboliques . . . 701
4.4 Extension au domaine complexe . . . 704
5 Formulaires . . . 705
5.1 Dérivées des fonctions élémentaires . . . 705
5.2 Fonctions trigonométriques inverses . . . 705
5.3 Fonctions trigonométriques hyperboliques inverses . . . 706
EXERCICES . . . 706

IV.4 Séries numériques 713

1 Convergence d'une série . . . 713
1.1 Définitions . . . 713
1.2 Premiers résultats . . . 717
2 Séries à termes réels positifs . . . 720
2.1 Convergence par comparaison . . . 720
2.2 Utilisation d'une intégrale . . . 724
2.3 Application : développement d'un réel positif . . . 727
3 Séries à termes réels ou complexes . . . 730
3.1 Convergence absolue . . . 730
3.2 Séries alternées . . . 732
EXERCICES . . . 735

IV.5 Introduction à l'intégration 739

1 Intégrale des fonctions en escalier . . . 740
1.1 Subdivision d'un segment . . . 740
1.2 Fonctions en escalier . . . 740
1.3 Intégrale d'une fonction en escalier . . . 741
1.4 Propriétés de l'intégrale des fonctions en escalier . . . 742
2 Fonctions continues par morceaux . . . 743
2.1 Définition, exemples . . . 743
2.2 Approximation des fonctions continues par morceaux . . . 745
2.3 Intégrale d'une fonction continue par morceaux . . . 746
3 Propriétés de l'intégrale . . . 748
3.1 Linéarité, relation de Chasles . . . 748
3.2 Inégalités . . . 749
3.3 Cas des fonctions continues . . . 751
3.4 Sommes de Riemann . . . 754

4 Intégration et dérivation, calcul des intégrales . . . 756
4.1 Fonctions continues par morceaux sur un intervalle . . . 756
4.2 Le théorème fondamental de l'analyse . . . 758
5 Méthodes de calcul d'intégrales . . . 761
5.1 Intégration par parties . . . 761
5.2 Changement de variable . . . 764
5.3 Quel changement de variable choisir ? . . . 765
5.4 Intégration des fractions rationnelles . . . 770
EXERCICES . . . 773

IV.6 Introduction aux fonctions vectorielles d'une variable réelle 779

1 Suites vectorielles . . . 779
1.1 Distance entre deux vecteurs . . . 780
1.2 Convergence de suites . . . 782
1.3 Suites vectorielles définies par une récurrence linéaire . . . 785
1.4 Suites réelles définies par une récurrence d'ordre 2 . . . 787
2 Fonctions vectorielles . . . 789
2.1 Continuité . . . 789
2.2 Dérivabilité . . . 791
2.3 Opérations sur les dérivées . . . 792
2.4 Inégalité des accroissements finis . . . 794
2.5 Intégration . . . 795
3 Équations différentielles linéaires . . . 797
3.1 Équations différentielles linéaires scalaires d'ordre 1 . . . 798
3.2 Équations vectorielles d'ordre 1 . . . 802
3.3 Allure des solutions d'une équation homogène en dimension 2 . . . 807
3.4 Équations différentielles linéaires d'ordre 2 à coefficients constants . . . 811
EXERCICES . . . 816

IV.7 Première initiation aux fonctions de plusieurs variables 819

1 Continuité . . . 821
1.1 Ouverts, fermés et compacts . . . 821
1.2 Fonctions continues . . . 823
1.3 Théorème des bornes . . . 824
1.4 Norme d'une application linéaire . . . 825
2 Différentiabilité . . . 826
2.1 Dérivées partielles . . . 826
2.2 Dérivée suivant un vecteur . . . 828
2.3 Différentielle . . . 829
2.4 Matrice jacobienne . . . 830
3 Propriétés fondamentales . . . 832
3.1 Opérations élémentaires . . . 832
3.2 Différentielle d'une application composée . . . 835
3.3 Applications continûment différentiables . . . 838
3.4 Théorème des accroissements finis . . . 839
4 Applications de la notion de différentiabilité . . . 841
4.1 Plan tangent au graphe d'une fonction $f : \mathbb{R}^2 \to \mathbb{R}$ . . . 841
4.2 Dérivation sur $\mathbb{C}$ . . . 846
EXERCICES . . . 850

IV.8 Approximation 855

1 Introduction . . . 855
2 Formules de Taylor . . . 856
2.1 Formule de Taylor avec reste intégral . . . 857
2.2 Formules de Taylor pour les fonctions vectorielles . . . 858
3 Équivalents et notations de Landau . . . 860
3.1 Équivalents . . . 861
3.2 Notations de Landau . . . 867
4 Développements limités. . . . 871
4.1 Définition et premières propriétés — Exemples . . . 871
4.2 Développements limités des fonctions usuelles . . . 875
4.3 Développements limités à droite et à gauche — Développements limités à l'infini . . . 878
4.4 Opérations sur les développements limités . . . 880
4.5 Calculs de limites et d'équivalents . . . 891
5 Méthodes de calcul approché d'intégrales . . . 893
5.1 Méthode des rectangles . . . 894
5.2 Méthode des rectangles médians . . . 895
5.3 Méthode des trapèzes . . . 897
5.4 Méthode de Simpson . . . 899
EXERCICES . . . 903

V Probabilités, statistiques

V.1 Statistique descriptive 915

1 Introduction à la statistique descriptive . . . 915
1.1 Données statistiques . . . 915
1.2 Représentation des données . . . 917
2 Statistique descriptive univariée . . . 922
2.1 Mesures de tendance centrale. . . . 922
2.2 Mesures de dispersion . . . 925
3 Statistique descriptive bivariée . . . 928
3.1 Ajustement linéaire par moindre carrés . . . 929
3.2 Covariance et corrélation . . . 932
3.3 Corrélation et régression . . . 933
3.4 Régression linéaire facile avec des outils logiciels. . . . 934
EXERCICES . . . 935

V.2 Probabilités finies 941

1 Introduction aux probabilités . . . 941
1.1 Expériences aléatoires, événements . . . 941
1.2 Espace de probabilité fini . . . 943
1.3 Probabilités de réunions d'ensembles : règle d'inclusion-exclusion . . . 947
2 Combinatoire . . . 949
2.1 Généralités sur le dénombrement . . . 949
2.2 Dénombrements classiques . . . 951
2.3 Dénombrement appliqué au loto. . . . 953

3 Conditionnement et indépendance 954
3.1 Probabilité conditionnelle 954
3.2 Probabilités composées, formule des probabilités totales 956
3.3 Formule de Bayes 958
3.4 Indépendance de deux événements. 959
3.5 Indépendance d'une famille d'événements 962
EXERCICES 964

V.3 Variables aléatoires 967

1 Lois et variables aléatoires. 967
1.1 Définitions 967
1.2 Histogrammes 968
2 Quelques lois usuelles 969
2.1 Loi de Bernoulli 969
2.2 Loi binomiale et nombre de succès 970
2.3 Echantillonnage et loi hypergéométrique 973
3 Espérance de variables aléatoires réelles 975
3.1 Définition de l'espérance 976
3.2 Propriétés élémentaires de l'espérance 977
3.3 Propriétés de transport 977
3.4 Une application de la linéarité de l'espérance : formule d'inclusion-exclusion. . 978
3.5 Variance 979
3.6 Espérances et variances pour des lois usuelles 981
4 Familles de variables aléatoires 983
4.1 Loi d'un vecteur aléatoire 983
4.2 Covariance et corrélation 986
4.3 Indépendance 987
4.4 Indépendance et covariance 989
4.5 Loi faible des grands nombres 991
EXERCICES 992

Index 999

Avant-propos

Ce livre est le premier d'une série de trois ouvrages de mathématiques pour la licence[1]. Il couvre les programmes de mathématiques des diverses filières scientifiques de première année. Il contient un cours complet, illustré d'exemples et d'applications, et des indications historiques. De plus, il propose au fil du texte de nombreux exercices corrigés qui permettront à l'étudiant de s'entraîner au fur et à mesure de son apprentissage. On trouvera aussi à la fin de chaque « module » des exercices supplémentaires[2] avec des indications de solutions. Une correction détaillée d'une grande partie de ces exercices est accessible sur le site de l'éditeur.

Dans cette nouvelle édition nous avons tenu compte de l'évolution récente des programmes de l'enseignement secondaire (importante pour certains thèmes) et des modifications des enseignements universitaires et, d'autre part, des remarques de nos lecteurs et de nos collègues enseignants.

Nos livres sont conçus comme une aide à l'enseignement oral dispensé par nos collègues dans les cours et travaux dirigés, en particulier par une construction « modulaire ». Les différents sujets apparaissent, pour chaque année d'enseignement, groupés dans un seul volume mais l'ordre de lecture n'est pas imposé, et chaque étudiant peut se concentrer sur tel ou tel aspect en fonction de son programme et de son travail personnel.

Ce livre peut être utilisé par un enseignant comme ouvrage de base pour son cours, dans l'esprit d'une pédagogie encore peu utilisée en France, mais qui a largement fait ses preuves ailleurs. Nous avons aussi pensé à l'étudiant travaillant seul, sans appui d'un corps professoral.

Dans les mathématiques d'aujourd'hui, un certain nombre de théories puissantes sont au premier plan. Leur maniement, au moins à un certain niveau, devra évidemment être acquis par l'étudiant à la fin de ses années de Licence. Mais celui-ci devra aussi avoir appris à calculer, sans s'appuyer exclusivement sur les ordinateurs et les logiciels, à « se débrouiller » devant un problème abstrait ou issu des applications. Nous avons donc, à cette fin, mis en place une approche adaptée. Nous insistons, dès la première année, sur les exigences de rigueur (définitions précises, démonstrations rigoureuses), mais les choses sont mises en place de façon progressive et pragmatique, et nous proposons des exemples riches, dont

[1] Dans ce livre, les deux autres ouvrages seront notés L2 (*Mathématiques Tout-en-un pour la Licence 2*) et L3 (*Mathématiques Tout-en-un pour la Licence 3*).

[2] Certains, plus difficiles, sont marqués d'une ou deux étoiles

l'étude met souvent en œuvre des approches multiples. Nous aidons progressivement le lecteur à acquérir le maniement d'un outillage abstrait puissant, sans jamais nous complaire dans l'abstraction pour elle-même et un formalisme sec et gratuit : le cœur des mathématiques n'est sans doute pas un corpus de théories, si profondes et efficaces soient-elles, mais un certain nombre de problèmes dans toute leur complexité, souvent issus d'une réflexion sur le monde qui nous entoure.

Historiquement les mathématiques se sont développées pendant des siècles en relation avec les autres sciences. Après une phase de « repliement sur elles-mêmes », leurs interactions se développent à nouveau vigoureusement (avec la physique, l'informatique, la mécanique, la chimie, la biologie...). Nous souhaitons, au niveau de l'enseignement des premières années d'université, accompagner ce mouvement et, en pratique, la mise en place ici ou là de filières scientifiques pluridisciplinaires avec une composante mathématique pure ou appliquée. Nous avons, par exemple, introduit de solides initiations aux probabilités et statistiques et à l'algorithmique dès ce volume de première année.

Malgré tout le soin apporté à cet ouvrage il est inévitable que quelques erreurs subsistent. Nous prions le lecteur, qui pourra les signaler à l'éditeur ou à l'un des auteurs pour correction lors d'un nouveau tirage, de nous en excuser.

Une correction détaillée des 530 énoncés supplémentaires est accessible sur le site `http://www.dunod.com` de l'éditeur à la page de ce livre. Ces corrigés sont au format pdf, ce qui permet facilement de faire une recherche sur un mot ou le numéro d'un exercice ou d'imprimer une solution.

Edmond Ramis a enseigné de longues années en classes préparatoires et écrit une série de livres pour ces classes. Ces livres ont fortement contribué à la formation de plusieurs générations d'étudiants, finissant par acquérir un nom commun : « le Ramis ». Plusieurs de ces étudiants sont maintenant enseignants-chercheurs ou chercheurs dans les universités. Au début des années 2000, André Warusfel a eu l'idée de prolonger l'œuvre d'Edmond Ramis par une série d'ouvrages pour l'enseignement universitaire. André nous a hélas quittés. Nous poursuivons l'édition de ces livres au plus près de l'esprit initial.

Jean-Pierre Ramis

Partie I

Notations et vocabulaire

Vers 1870 Cantor, avec Dedekind puis Peano et quelques autres, créa la théorie des ensembles. Elle fut d'abord mal accueillie et, plus tard, conduisit à une crise aussi violente que féconde. Aujourd'hui cette théorie (au moins à son niveau élémentaire), avec son langage et ses principales notations, est l'outil de base du mathématicien.
Après Cantor apparurent un certain nombre de problèmes : discussions entre Lebesgue, Borel et Baire sur l'axiome du choix, graves paradoxes de la théorie des ensembles découverts par Russell (*l'ensemble de tous les ensembles*). Toujours au début du XXe siècle, Hilbert entreprit de donner des bases solides, totalement indiscutables (dans la ligne historique d'Aristote, Leibniz, Boole, Frege, Peano...). Il avait l'ambition de formaliser complètement *tout* le raisonnement mathématique. Mais malgré le génie de Hilbert, ce programme finit par un échec retentissant avec, en 1931, une stupéfiante découverte de Gödel (le théorème d'incomplétude, relié à des paradoxes classiques : *un crétois dit que tous les crétois sont menteurs*).
Toutefois, si la réponse de Hilbert était fausse, sa *question* était bonne ! En effet, elle conduisit, entre autres, à l'invention des ordinateurs et à un extraordinaire développement de la programmation et du calcul... initiés par des travaux de von Neumann et Turing (eux-mêmes fortement influencés par le résultat de Gödel). L'histoire n'est pas finie et tout cela se prolonge dans d'importantes recherches récentes en informatique théorique (comme par exemple les travaux de G.J. Chaïtin, en relation avec le concept physique de l'entropie).

Ce premier module n'est évidemment pas à lire en premier ! Mais il faut y revenir sans cesse, au moins au départ, pour préciser tel ou tel point, lever telle obscurité apparente, bref l'utiliser comme un dictionnaire de scrabble ou un manuel de grammaire. Certains passages (sur les cardinaux, ou les connecteurs logiques) peuvent paraître complexes à première vue. Cela correspond à de vraies difficultés et ne doit surtout pas décourager le lecteur. Il est conseillé de s'y replonger plusieurs fois, en fonction des besoins, et un jour son contenu sera devenu tout à fait abordable. Bonne lecture discursive !

Fondements I.1

Du paradis que Cantor a créé pour nous,
nul ne nous chassera (Hilbert).

La théorie des ensembles a été créée par Georg Cantor à la fin du XIXe siècle pour résoudre des problèmes d'analyse réelle. Elle a suscité des résistances psychologiques (certains mathématiciens y voyaient de la métaphysique) et posé d'importantes difficultés logiques (la « crise des fondements », au début du XIXe siècle). Cependant, elle a rapidement investi les autres parties des mathématiques ; en particulier, l'algèbre moderne, créée en Allemagne entre les deux guerres mondiales, est entièrement formulée dans le langage des ensembles.

Dans leur version moderne, toutes les mathématiques sont fondées sur la théorie des ensembles et *tout objet mathématique est un ensemble*. Ainsi, dans la théorie de Von Neumann, on définit l'entier naturel n comme un ensemble particulier à n éléments ; par exemple, $0 := \varnothing$ (voir la section 6.3).

Nous adopterons un point de vue plus naïf. Nos ensembles et nos applications contiennent (ou mettent en jeu) des éléments, qui sont des objets mathématiques plus ou moins élémentaires et dont on ne précise pas nécessairement la nature. Toutes ces « entités » sont cependant soumises à des *axiomes* que nous rendrons explicites, autant du moins que ce point de vue non formel le permet. Il ne s'agit pas à proprement parler de *fondations* des mathématiques, mais plutôt de la mise en place de l'outillage de base qui interviendra dans *toutes les structures* et *toutes les théories* qui seront exposées dans cet ouvrage. Ainsi, l'exposition n'est pas strictement linéaire : des exemples seront tirés des modules ultérieurs, certaines définitions (lois de composition, nombres entiers, relations d'ordre) apparaîtront à plusieurs reprises si des points de vue différents le justifient. Notons que la section 6 (cardinaux) peut être omise en première lecture

Enfin, nous ne formaliserons pas la logique. Nous admettrons une connaissance intuitive de la notion de vrai et de faux, de l'égalité et même des entiers naturels. Des règles de raisonnement seront décrites *à la fin* de ce module, alors que nous aurons déjà suffisamment de matière pour les illustrer.

Une théorie exposée de manière purement formelle et sans exemples serait bien indigeste ! Et par ailleurs, nous savons bien que dans le secondaire ont été rencontrés de nombreux ensembles, en particulier les ensembles de nombres : $\mathbb{N}$ (entiers naturels), $\mathbb{N}^*$ (entiers naturels non nuls), $\mathbb{Z}$ (entiers relatifs), $\mathbb{Q}$ (nombres rationnels), $\mathbb{R}$ (nombres réels), $\mathbb{R}_+$ (nombres réels positifs ou nuls), $\mathbb{R}_+^*$ (nombres réels strictement positifs), $\mathbb{C}$ (nombres complexes), etc. Tous ces ensembles seront redéfinis formellement[1] dans les modules correspondants de ce cours. Cependant, le lecteur est invité à profiter de la connaissance intuitive qu'il en a déjà pour illustrer les définitions et constructions qui vont suivre.

1 Ensembles

Le lecteur qui ne désire pas s'initier à l'axiomatique des ensembles peut aborder la lecture de ce module à la page 7 ; s'il est familier avec les définitions de base, il peut même commencer par le théorème 1 de la page 14.

1.1 Appartenance, éléments

Les ensembles sont formés d'éléments. Ce que sont ces derniers n'est pas précisé. On introduit donc une relation particulière entre un élément x et un ensemble E, la *relation d'appartenance*. Cette relation s'écrit $x \in E$, ce qui se lit « x appartient à E », « x est élément de E », « E contient x ». Notons cependant que cette dernière formulation est ambiguë, à cause de l'inclusion (définition 2 de la page 7). La négation de la relation d'appartenance s'écrit : $x \notin E$, ce qui signifie que $x \in E$ est faux, ou encore que $\neg(x \in E)$ est vrai (en logique, le symbole $\neg$ signifie « non »). On lit : « x n'appartient pas à E », etc.

Les ensembles les plus célèbres sont $\mathbb{N}$ (dont les éléments sont appelés *entiers naturels*), $\mathbb{Z}$ (dont les éléments sont appelés *entiers relatifs*), $\mathbb{Q}$ (dont les éléments sont appelés *nombres rationnels*), $\mathbb{R}$ (dont les éléments sont appelés *nombres réels*) et $\mathbb{C}$ (dont les éléments sont appelés *nombres complexes*).

Axiome 1. L'axiome fondamental est l'*axiome d'extensionalité* pour les ensembles, qui dit qu'un ensemble est totalement caractérisé par ses éléments :

$$\big(\forall x\ ,\ x \in E \Leftrightarrow x \in F\big) \quad \Longrightarrow \quad E = F. \tag{1}$$

Lire : « deux ensembles qui ont les mêmes éléments sont égaux ».

Dans la phrase entre guillemets, l'usage de l'article défini « les » (mêmes éléments) dit qu'il s'agit de *tous* les éléments : c'est pourquoi la formule (1) comporte un quantificateur $\forall x$, qu'il faut lire « quel que soit x », ou encore « pour tout x ».

Remarquons que l'implication réciproque va de soi, pour des raisons de logique pure. L'usage mathématique veut en effet que, si l'on a une égalité $E = F$, alors toute propriété vérifiée par E est

[1] À noter que l'ensemble $\mathbb{N}$ des entiers naturels sera même construit deux fois : une première fois dans ce module (*entiers de Von Neumann*, voir la section 6.3) et une deuxième fois dans le module suivant. C'est cette dernière construction qui doit ici être regardée comme « officielle ».

vérifiée par F (« substitutivité de l'égalité »). C'est même ainsi que Leibniz définissait l'égalité ! Pour en revenir à nos ensembles, si l'on a $E = F$ et $x \in E$, on en déduit que $x \in F$. Pratiquement, on pourra prouver l'égalité de deux ensembles *par équivalence*.

Exercice 1.
Déterminer l'intersection des droites $D : y = 2x + 1$ et $D' : y = 3x - 2$.
Solution. C'est l'ensemble des points (x, y) qui appartiennent à D et à D', c'est-à-dire qui satisfont le système d'équations $(y = 2x + 1\ ,\ y = 3x - 2)$. Les méthodes de résolution du secondaire (par équivalence, justement) permettent de déduire que (x, y) est solution de ce système si, et seulement si, $x = 3$, $y = 7$. Comme nous allons le voir, il y a un ensemble dont le seul élément est le point $P = (3, 7)$, c'est le singleton $\{P\}$. On a donc : $M \in D \cap D' \Leftrightarrow M \in \{P\}$, d'où $D \cap D' = \{P\}$.

La plupart des *axiomes* qui vont apparaître dans ce module mettront en jeu une certaine propriété $P(x)$ d'un élément indéterminé x (P s'appelle un *prédicat*[2]) ; un axiome dira alors que les éléments x tels que $P(x)$ est vrai forment un ensemble, autrement dit, qu'il existe un ensemble E tel que $\forall x\ ,\ x \in E \Leftrightarrow P(x)$. Cela se lit : « quel que soit x, $x \in E$ si, et seulement si, $P(x)$ ». L'axiome d'extensionalité permettra alors de déduire que E est *unique*. En effet, si F est un (autre) ensemble tel que $\forall x\ ,\ x \in F \Leftrightarrow P(x)$, les règles usuelles concernant l'équivalence logique entraînent : $\forall x\ ,\ x \in E \Leftrightarrow x \in F$, donc $E = F$. Après quelques exemples simples, ce procédé sera systématisé en 1.2.

1.1.1 Construction d'ensembles finis

Nous allons commencer par les ensembles « finis » dans un sens intuitif, terme qui sera précisé à la section 6.

Axiome et définition 2. Il existe un ensemble qui n'admet aucun élément.

$$\exists E\ :\ \forall x\ ,\ x \notin E. \tag{2}$$

On le note $\varnothing$ et on l'appelle *ensemble vide*.

L'existence d'*un* (sous-entendu : « au moins ») ensemble qui n'a aucun élément (« un », article indéfini, exprime l'existence, quantificateur $\exists$) est le contenu de cet axiome. Mais, d'après l'axiome d'extensionalité, l'ensemble qui n'a aucun élément est unique. C'est pourquoi, dans la deuxième partie de la phrase, on dit *le* (note), qui sous-entend l'unicité. La formule (2) peut donc être renforcée ainsi :

$$\exists !E\ :\ \forall x\ ,\ x \notin E.$$

Le symbole $\exists !$ signifie « il existe un unique ».

Axiome et définition 3. Soit a un objet mathématique. L'*axiome du singleton* dit qu'il existe un ensemble dont le seul élément est a. On le note $\{a\}$, lire « singleton a ».

Notant $E = \{a\}$, on a donc : $\forall x\ ,\ x \in E \Leftrightarrow x = a$. D'après l'axiome d'extensionalité, un tel ensemble est unique. Tout objet mathématique est donc élément d'un ensemble, et nous appellerons donc *élément* un tel objet. Naturellement, $\{a\} = \{b\}$ équivaut logiquement à $a = b$.

Axiome et définition 4. Soient a, b deux éléments (non nécessairement distincts). L'*axiome de la paire* dit qu'il existe un ensemble dont les seuls éléments sont a et b. On le note $\{a, b\}$. Si $a \neq b$, on l'appelle « paire formée de a et b ».

---

[2] La notion de prédicat et son usage seront abordés de manière plus détaillée à la section 7.2

D'après l'axiome d'extensionalité, l'ensemble $\{a, b\}$ est unique. L'axiome de la paire implique d'ailleurs celui du singleton : lorsque $a = b$, on trouve $\{a, a\} = \{a\}$. On démontre de même que $\{a, b\} = \{b, a\}$ par de la logique pure ! Pour tout x :

$$x \in \{a, b\} \iff (x = a \text{ ou } x = b) \iff (x = b \text{ ou } x = a) \iff x \in \{b, a\},$$

et l'on applique l'axiome d'extensionalité. On prendra donc garde à ne pas confondre la paire $\{a, b\}$ et le couple (a, b) (ces derniers apparaîtront à la page 10) : on n'a $(a, b) = (b, a)$ que si $a = b$.

Exercice 2.
Montrer que les ensembles $\varnothing$, $\{\varnothing\}$, $\{\{\varnothing\}\}$ et $\{\varnothing, \{\varnothing\}\}$ sont deux à deux distincts.
Solution. Seul le premier n'a aucun élément, il est donc différent de chacun des trois autres ; par ailleurs, puisque $\varnothing \neq \{\varnothing\}$ (on vient de le voir), les singletons correspondants sont différents. Le dernier ensemble est une paire parce que $\varnothing \neq \{\varnothing\}$, il n'est donc égal à aucun des trois premiers.

1.1.2 Définition en extension

Chaque fois que l'on se donne des objets $a_1, \ldots, a_n$, on dispose d'une version plus générale des axiomes précédents.

Axiome et définition 5. Il existe un ensemble dont les seuls éléments sont $a_1, \ldots, a_n$ (d'après l'axiome d'extensionalité, il est unique). Cet ensemble est noté $\{a_1, \ldots, a_n\}$. On dit que l'on a défini cet ensemble « en extension », c'est-à-dire en énumérant ses éléments.

Exercice 3.
Donner une condition nécessaire et suffisante pour que l'on ait l'égalité $\{a_1, \ldots, a_n\} = \{a\}$, puis $\{a_1, \ldots, a_n\} = \varnothing$.
Solution. La première égalité a lieu si, et seulement si, les objets $a_1, \ldots, a_n$ sont tous égaux à a. La seconde est en principe impossible. Il existe cependant une convention (en fait, un abus de langage) selon laquelle, lorsque $n = 0$ (aucun objet !), la notation $\{a_1, \ldots, a_n\}$ désigne l'ensemble vide.

Pour $n = 1, 2$, on retrouve les axiomes précédents. Au delà, on trouve les ensembles $\{a, b, c\}$, $\{a, b, c, d\}$, etc. La notation générale $a_1, \ldots, a_n$ est légèrement abusive car elle sous-entend que l'on sait interpréter les $\ldots$ intermédiaires. Dans la pratique, on s'autorise des « définitions en extension incomplètes » comme $\{0, 2, \ldots, 2n\}$ parce que l'on devine aisément la *loi de formation* des éléments énumérés : ici, les entiers de la forme $2k$, où l'entier k varie de 0 à n. Par abus courant, la même notation s'emploie pour énumérer des ensembles infinis, comme $\mathbb{N} = \{0, 1, 2, \ldots\}$ ou l'ensemble $2\mathbb{N} = \{0, 2, 4, \ldots\}$ des entiers naturels pairs. Cela peut être dangereux (ambiguïté sur la loi de formation) et cela cache en réalité des axiomes plus puissants. En fait, pour anticiper, on a une suite $(u_n)_{n \in \mathbb{N}}$ et l'ensemble $\{u_0, u_1, \ldots\}$ est l'ensemble image de cette suite, c'est-à-dire l'ensemble $\{u_n \mid n \in \mathbb{N}\}$.

Exercice 4.
Reconnaître l'ensemble $\{0, 1, -1, 2, -2, \ldots\}$.
Solution. Il s'agit bien sûr de l'ensemble $\mathbb{Z}$. On peut le décrire à l'aide de la suite $(u_n)_{n \in \mathbb{N}}$ définie par les formules $u_{2p} = -p$ et $u_{2p+1} = p + 1$.

1.2 Définition en compréhension

Dans des situations très répandues, on a un prédicat $P(x)$ (*i.e.* une propriété dépendant de l'élément indéterminé x) et l'on se demande s'il existe un ensemble E tel que $\forall x\ ,\ x \in E \Leftrightarrow P(x)$. Si c'est le cas, E est unique (axiome d'extensionalité).

Définition 1. Si un tel ensemble existe, on le note $\{x \mid P(x)\}$, ce qui se lit « l'ensemble des x tels que $P(x)$ ». On dit alors que la propriété P est *collectivisante*.

Exemples. Comme on l'a vu, les propriétés $x \neq x$, $x = a$, $(x = a$ ou $x = b)$ (a et b étant fixés) sont collectivisantes. Pour un ensemble E donné, la propriété $x \in E$ est évidemment collectivisante, et l'on a l'égalité $E = \{x \mid x \in E\}$.
Toutes ces précautions sont nécessaires parce que l'on ne peut pas déclarer tous les prédicats collectivisants :

Exercice 5.
L'un des plus vieux paradoxes de la théorie des ensembles concerne « l'ensemble des éléments qui ne sont pas éléments d'eux-mêmes », c'est-à-dire l'ensemble $E := \{x \mid x \notin x\}$. Ainsi, $\forall x\ ,\ x \in E \Leftrightarrow x \notin x$. Appliquer cette assertion à $x := E$.
Solution. On tombe sur la célèbre contradiction :

$$E \in E \Longleftrightarrow E \notin E.$$

L'existence d'un tel ensemble E entraînerait une contradiction. Selon les principes de la logique mathématique, cette existence est donc fausse, l'ensemble E n'existe pas et la propriété $P(x) := (x \notin x)$ n'est donc pas collectivisante.

Il nous faudra donc expressément déclarer par des axiomes que certains prédicats sont collectivisants et que certains ensembles existent bel et bien, comme on l'a fait pour les définitions en extension. Cependant, dans l'immense majorité des cas, cela sera facilité par l'axiome suivant, appelé *axiome de séparation* :

Axiome et définition 6. Soient E un ensemble et $P(x)$ une propriété des éléments de E. Alors la propriété « $P(x)$ et $x \in E$ » est collectivisante et l'on note l'ensemble associé $\{x \in E \mid P(x)\}$, ce qui se lit « l'ensemble des x éléments de E tels que $P(x)$ ».

Il n'existe donc pas d'ensemble de tous les ensembles : en effet, si un tel ensemble existait, selon l'axiome de séparation, toute propriété serait collectivisante.

Exercice 6.
Les ensembles $\{x \in \mathbb{R} \mid x^2 + 1 = 0\}$ et $\{x \in \mathbb{N} \mid x + 1 = 0\}$ existent-ils ?
Solution. Oui, d'après l'axiome de séparation ! Mais ils sont évidemment vides (donc égaux). Ce sont leurs éléments qui n'existent pas…

1.2.1 Inclusion, parties

Définition 2. On dit que l'ensemble F est *inclus* dans l'ensemble E, ce que l'on note $F \subset E$, si tous les éléments de F sont éléments de E :

$$F \subset E \quad \Longleftrightarrow \quad \big(\forall x\ ,\ x \in F \Rightarrow x \in E\big).$$

On dit aussi que F est une *partie* ou un *sous-ensemble* de E.

On dit également que E contient F (mais c'est ambigu à cause de la relation $x \in E$ qui se dit aussi « E contient x »). L'ensemble vide est un sous-ensemble de tout ensemble. Tout ensemble est sous-ensemble de lui même. L'ensemble $\{x \in E \mid P(x)\}$ (c'est-à-dire l'ensemble des éléments de E qui ont la propriété P) est un sous-ensemble de E.

Exercice 7.
À quelle condition a-t-on $\{a\} \subset E$? $\{a, b\} \subset E$? $\{a\} \subset \{b\}$?
Solution. L'inclusion $\{a\} \subset E$ (resp. $\{a, b\} \subset E$) est équivalente à l'appartenance $a \in E$ (resp. aux appartenances $a \in E$ et $b \in E$).
L'inclusion $\{a\} \subset \{b\}$ est équivalente à l'égalité $a = b$.

En anticipant un peu sur les définitions ultérieures, on voit que la relation d'inclusion est réflexive : $E \subset E$, transitive : $(E \subset F$ et $F \subset G) \Rightarrow E \subset G$ et antisymétrique : $(E \subset F$ et $F \subset E) \Rightarrow E = F$; les deux premières propriétés sont immédiates, la troisième est une simple traduction de l'axiome d'extensionalité. L'inclusion est donc une relation d'ordre (définition 16 de la page 29).

Axiome et définition 7. Soit E un ensemble. La relation $x \subset E$ est collectivisante et définit *l'ensemble des parties de* E, noté $\mathcal{P}(E)$. On a donc : $\mathcal{P}(E) := \{x \mid x \subset E\}$.

Exemples. Le seul sous-ensemble de $\varnothing$ est $\varnothing$, d'où l'égalité $\mathcal{P}(\varnothing) = \{\varnothing\}$.
On vérifie de même que $\mathcal{P}(\{a\}) = \{\varnothing, \{a\}\}$ et que $\mathcal{P}(\{a, b\}) = \{\varnothing, \{a\}, \{b\}, \{a, b\}\}$.
Enfin, $\mathcal{P}\big(\mathcal{P}(\varnothing)\big) = \{\varnothing, \{\varnothing\}\}$.

1.3 Constructeurs

1.3.1 Constructeurs élémentaires

Pour construire la réunion de deux ensembles quelconques, un nouvel axiome est requis.

Axiome et définition 8. Soient E et F deux ensembles. Il existe alors un ensemble G tel que :

$$\forall x \ , \ x \in G \Longleftrightarrow (x \in E \ ou \ x \in F).$$

Cet ensemble (unique d'après l'axiome d'extensionalité) est appelé *réunion* ou *union* de E et de F et noté $E \cup F$. Ainsi, $E \cup F := \{x \mid x \in E \ ou \ x \in F\}$.

Rappelons que le « ou » des mathématiciens n'est pas exclusif : l'affirmation $x \in E \ ou \ x \in F$, n'exclut nullement que l'on ait $x \in E \ et \ x \in F$. Ainsi, $E \cup F$ contient en particulier les éléments communs à E et à F.

Définition 3. Soient E et F deux ensembles. L'ensemble des éléments de E qui sont dans F existe (axiome de séparation) et est unique (axiome d'extensionalité). On l'appelle *intersection* de E et de F et on le note $E \cap F$:

$$\forall x \ , \ x \in E \cap F \Longleftrightarrow (x \in E \ et \ x \in F).$$

Ainsi, $E \cap F := \{x \mid x \in E \ et \ x \in F\}$.
On dit que E et F sont *disjoints* si leur intersection est vide.

On a bien sûr $E \cap F = F \cap E$.

Définition 4. Soient E et F deux ensembles. L'ensemble des éléments de E qui ne sont pas dans F existe (axiome de séparation) et est unique (axiome d'extensionalité). On l'appelle *différence* de E et de F et on le note $E \setminus F$:

$$\forall x\ ,\ x \in E \setminus F \Longleftrightarrow (x \in E \textit{ et } x \notin F).$$

Ainsi, $E \setminus F := \{x \mid x \in E \textit{ et } x \notin F\}$.

Lorsque $F \subset E$, l'ensemble $E \setminus F$ est appelé *complémentaire de F dans E* et noté $\complement_E F$. Remarquons d'ailleurs que l'on ne peut parler de complémentaire de l'ensemble F « dans l'absolu », mais seulement relativement à un ensemble englobant E.

Les opérations de réunion et d'intersection vérifient de nombreuses règles de nature algébrique, qui sont toutes très faciles à vérifier une fois écrites :

Dans ces égalités, E, F et G désignent trois ensembles quelconques.

$E \cup (F \cup G) = (E \cup F) \cup G$	(associativité de la réunion)
$E \cup F = F \cup E$	(commutativité de la réunion)
$\varnothing \cup E = E \cup \varnothing = E$	(l'ensemble vide est neutre pour la réunion)
$E \cup E = E$	(tout ensemble est idempotent pour la réunion)
$E \cap (F \cap G) = (E \cap F) \cap G$	(associativité de l'intersection)
$E \cap F = F \cap E$	(commutativité de l'intersection)
$\varnothing \cap E = E \cap \varnothing = \varnothing$	(l'ensemble vide est absorbant pour l'intersection)
$E \cap E = E$	(tout ensemble est idempotent pour l'intersection)
$E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$	(distributivité de l'intersection par rapport à la réunion)
$E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$	(distributivité de la réunion par rapport à l'intersection)
$E \subset F \Longleftrightarrow E \cap F = E$	
$E \subset F \Longleftrightarrow E \cup F = F.$	

À titre d'exemple, démontrons les deux dernières règles, en commençant par les implications directes (ce qui signifie, de gauche à droite). Supposons donc l'inclusion $E \subset F$. En ce qui concerne la première égalité à démontrer, on a *a priori* l'inclusion $(E \cap F) \subset E$. Par ailleurs, on a à la fois $E \subset E$ (toujours vrai) et $E \subset F$ (c'est l'hypothèse), donc $E \subset (E \cap F)$ (par définition de l'intersection). De la double inclusion $(E \cap F) \subset E$ et $E \subset (E \cap F)$, on déduit enfin la première égalité désirée : $(E \cap F) = E$. L'autre égalité

se démontre également par double inclusion : l'inclusion $F \subset (E \cup F)$ est toujours vraie, l'inclusion réciproque $(E \cup F) \subset F$ vient (par définition de la réunion) de ce que $E \subset F$ (c'est l'hypothèse) et $F \subset F$ (toujours vrai).

Démontrons maintenant les implications réciproques (ce qui signifie, de droite à gauche). Supposons d'abord $(E \cap F) = E$. Alors tout élément de E est élément de $E \cap F$ (ces ensembles ont mêmes éléments puisqu'ils sont égaux) donc de F : autrement dit, $E \subset F$, comme espéré.

Supposons maintenant que $(E \cup F) = F$. Alors tout élément de E est élément de $E \cup F$ (évidemment) donc de F (qui est égal au précédent), et l'on a encore $E \subset F$.

Exercice 8.

À quelle condition a-t-on $E \cup F = E \cap F$?

Solution. Puisque $(E \cap F) \subset E \subset (E \cup F)$ et $(E \cap F) \subset F \subset (E \cup F)$, l'égalité ci-dessus entraîne $E = F$; la réciproque est immédiate.

L'union, l'intersection et la différence de deux parties d'un ensemble E sont encore des parties de E. Anticipant un peu sur la section 4, on peut donc les considérer comme des lois de composition interne sur $\mathcal{P}(E)$. Outre les règles algébriques énoncées plus haut, l'ensemble $\mathcal{P}(E)$ muni des lois $\cup$ et $\cap$ vérifie les propriétés suivantes :

F et G désignent des parties de l'ensemble E.

$E \cap F = F \cap E = F$	(l'ensemble E est neutre pour l'intersection)
$E \cup F = F \cup E = E$	(l'ensemble E est absorbant pour la réunion)
$\complement_E(\complement_E F) = F$	(involutivité du passage au complémentaire)
$\complement_E(F \cup G) = (\complement_E F) \cap (\complement_E G)$	(le passage au complémentaire est un morphisme $\cup \to \cap$)
$\complement_E(F \cap G) = (\complement_E F) \cup (\complement_E G)$	(le passage au complémentaire est un morphisme $\cap \to \cup$)

Les deux dernières formules sont connues sous le nom de *lois de Morgan*. Les deux premières lois sont immédiates d'après les règles précédentes. La troisième loi découle de l'équivalence : $\forall x \in E\ ,\ x \notin F \Leftrightarrow x \in \complement_E F$. La démonstration des lois de Morgan repose sur des règles logiques : soient P et Q deux assertions ; alors la négation de « P et Q » est « non P ou non Q », et la négation de « P ou Q » est « non P et non Q ». On applique ici ces règles à l'assertion $P := (x \in F)$ et à l'assertion $Q := (x \in G)$.

1.3.2 Produits cartésiens

On suppose que l'on sait former à partir de deux objets a et b un *couple* (a, b) de manière à satisfaire la règle suivante :

$$\forall a\ ,\ \forall b\ ,\ \forall c\ ,\ \forall d\ ,\ (a, b) = (c, d) \iff a = c \text{ et } b = d.$$

L'exercice I.1.1 de la page 57 suggère une *construction* d'un tel objet, due au mathématicien polonais Kuratowski. Les éléments a et b sont respectivement appelés *première* et *seconde composante* (ou encore *coordonnée*) du couple (a, b). Si $x = (a, b)$, on écrit parfois $a = p_1(x)$ et $b = p_2(x)$.

Axiome et définition 9. Soient E et F deux ensembles. Il existe un ensemble G dont les éléments sont les couples (a, b) formés d'un élément $a \in E$ et d'un élément $b \in F$:

$$G := \{x \mid \exists a \in E \text{ et } \exists b \in F \ : \ x = (a, b)\}.$$

D'après l'axiome d'extensionalité, un tel ensemble est unique. On le note $E \times F$ et on l'appelle *produit cartésien* des ensembles E et F. De manière équivalente :

$$E \times F := \{(a, b) \mid a \in E \text{ et } b \in F\}.$$

Plus généralement, à partir de n éléments $a_1, \ldots, a_n$, on peut former le n*-uplet* (ou encore n*-uple*) $x = (a_1, \ldots, a_n)$, dont les n composantes (ou coordonnées) sont $a_1 = p_1(x)$,..., $a_n = p_n(x)$. Pour $n = 3, 4, 5, \ldots$, on parle de *triplets, quadruplets, quintuplets,* etc. On a la règle :

$$\forall a_1 \ , \ \ldots \ , \ \forall a_n \ , \ \forall b_1 \ , \ \ldots \ , \ \forall b_n \ ,$$
$$(a_1, \ldots, a_n) = (b_1, \ldots, b_n) \Longleftrightarrow a_1 = b_1 \text{ et } \ldots \text{ et } a_n = b_n.$$

L'axiome ci-dessus se généralise ainsi : les n-uplets $(a_1, \ldots, a_n)$ formés d'éléments $a_1 \in E_1, \ldots, a_n \in E_n$ forment un ensemble, le *produit cartésien* $E_1 \times \cdots \times E_n$. Si l'un des E_i est vide, on ne peut former aucun n-uplet car il n'y a aucune possibilité pour sa $i^{ème}$ composante. Si tous les E_i sont non vides, en choisissant un élément a_i dans chacun des E_i, on forme un n-uplet $(a_1, \ldots, a_n)$ qui est élément de $E_1 \times \cdots \times E_n$. On en conclut que le produit cartésien $E_1 \times \cdots \times E_n$ est vide si, et seulement si, l'un des ensembles E_i est vide.

L'application stricte des règles entraîne que $\big(x, (y, z)\big) \neq \big((x, y), z\big)$ (ces deux couples n'ont pas pas la même première projection). Cependant, on les identifie fréquemment à (x, y, z). De même, on identifiera les ensembles produits $E \times (F \times G)$ et $(E \times F) \times G$ à $E \times F \times G$. Cette convention se généralise naturellement à des produits plus compliqués. On pose souvent $E^n = E \times \cdots \times E$ (n facteurs). Avec les identifications précédentes, cela revient à poser $E^1 = E$, puis, par récurrence, $E^{n+1} = E \times E^n$ (ou $E^n \times E$, au choix). La *diagonale* de E^n est l'ensemble $\{x \in E^n \mid p_1(x) = \cdots = p_n(x)\}$. On peut également le décrire comme $\{(x, \ldots, x) \mid x \in E\}$, où le uplet $(x, \ldots, x)$ a n composantes.

2 Applications

Nous abordons ici l'étude des applications, qui ne sont autres que les fonctions, mais avec un nom savant et une théorie plus précise. Cependant, comme pour les ensembles, nous ne partirons pas d'une définition formelle.

2.1 Applications et graphes

Une *application* d'un ensemble E dans un ensemble F est (naïvement) un « procédé » (non précisé) qui associe à chaque élément $x \in E$ un élément $f(x) \in F$. Une telle application est notée

$$f : \begin{array}{ccc} E & \longrightarrow & F \\ x & \longmapsto & f(x). \end{array}$$

L'élément $f(x) \in F$ est appelé *image* de l'élément $x \in E$ par l'application f.

L'ensemble E est appelé *ensemble de départ* (ou encore *source*) de l'application f ; l'ensemble F est appelé *ensemble d'arrivée* (ou encore *but*) de f. Il y a un *axiome d'extensionalité* pour les applications :

Axiome 10. Deux applications f et g de E dans F sont égales si, et seulement si, elles attribuent la même image à tout élément de E :

$$f = g \iff \forall x \in E\ ,\ f(x) = g(x).$$

En principe, lorsque l'on connaît l'application f, sa source et son but sont donc déterminés. Dans la pratique, on a tendance à identifier deux applications f et g de même source E dès qu'elles vérifient la relation $\forall x \in E\ :\ f(x) = g(x)$: par exemple, les applications

$$f : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & x^2 \end{array} \qquad \text{et} \qquad g : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}_+ \\ x & \longmapsto & x^2. \end{array}$$

Cela peut poser problème, par exemple lorsque l'on discute de la surjectivité (*cf.* la définition 7 de la page 14) : l'application g est surjective, alors que l'application f ne l'est pas. On fait appel au bon sens du lecteur pour vérifier que ce qu'il écrit a bien un sens.

Exemples.

1. Pour tout ensemble E, l'application $x \mapsto x$ de E dans lui-même est appelée *application identité* de E et notée Id_E.
2. Les ensembles E et F étant quelconques (ce dernier non vide), on peut définir, pour tout $a \in F$, l'application $x \mapsto a$ de E dans F : c'est une *application constante*.
3. Si $E \subset F$, l'application $x \mapsto x$ de E dans F est appelée *application canonique* ou *injection canonique* de E dans F.
4. Si f est une application de E dans F et si $E' \subset E$, on peut définir l'application $x \mapsto f(x)$ de E' dans F. Elle est appelée *restriction* de f à E' et notée $f_{|E'}$.
5. Si $F' \subset F$ est tel que $\forall x \in E\ ,\ f(x) \in F'$, l'application $x \mapsto f(x)$ de E dans F' est appelée *corestriction* de f à F'.
6. Si $E = \varnothing$, on convient qu'il y a une et une seule application de E dans F, appelée *application vide*. Par l'abus mentionné plus haut, c'est la même quel que soit F.
7. Si $F = \varnothing$ et $E \neq \varnothing$, il n'y a aucune application de E dans F.
8. Soient $E_1, \ldots, E_n$ des ensembles. Pour chaque $i = 1, \ldots, n$, l'application $p_i : (x_1, \ldots, x_n) \mapsto x_i$ (sa $i^{ème}$ composante) de $E_1 \times \cdots \times E_n$ dans E_i est appelée $i^{ème}$ *projection* ou $i^{ème}$ *application coordonnée*.
9. Soient E, F_1, F_2 des ensembles et $f_1 : E \to F_1$ et $f_2 : E \to F_2$ des applications. On en déduit une application $x \mapsto \big(f_1(x), f_2(x)\big)$ de E dans $F_1 \times F_2$. Cette application peut être notée (f_1, f_2), avec toutefois un risque de confusion avec le couple formé de f_1 et f_2.
   Plus généralement, on peut définir $(f_1, \ldots, f_n) : E \to F_1 \times \cdots \times F_n$.
10. Soient E_1, E_2, F_1, F_2 des ensembles et $f_1 : E_1 \to F_1$ et $f_2 : E_2 \to F_2$ des applications. On en déduit une application $(x_1, x_2) \mapsto \big(f_1(x_1), f_2(x_2)\big)$ de $E_1 \times E_2$ dans $F_1 \times F_2$. Cette application est notée $f_1 \times f_2$.
    Plus généralement, on peut définir $f_1 \times \cdots \times f_n : E_1 \times \cdots \times E_n \to F_1 \times \cdots \times F_n$.

Définition 5. On appelle *graphe* de l'application $f : E \to F$ l'ensemble :

$$\Gamma_f := \{(x, y) \in E \times F \mid y = f(x)\}.$$

On peut également le décrire comme $\{(x, f(x)) \mid x \in E\}$. En vertu de l'axiome d'extensionalité, deux applications de même source et de même but sont égales si, et seulement si, elles ont le même graphe. Certains ouvrages définissent d'ailleurs une application comme un triplet (E, Γ, F), où $\Gamma \subset (E \times F)$ vérifie la propriété suivante :

$$\forall x \in E \ , \ \exists ! y \in F \ : \ (x, y) \in \Gamma,$$

autrement dit, tout $x \in E$ a une unique image $y \in F$.

Exemples.

1. Le graphe de Id_E est la diagonale de E^2.
2. Le graphe de l'application constante $x \mapsto a$ est $E \times \{a\}$.
3. Le graphe de la restriction $f' = f_{|E'}$ est $\Gamma_{f'} = \Gamma_f \cap (E' \times F)$.
4. Le graphe de l'application vide est l'ensemble vide.

2.2 Images et antécédents

Définition 6. Soit $f : E \to F$ une application. On appelle *antécédent* d'un élément $y \in F$ par l'application f tout élément $x \in E$ tel que $f(x) = y$. L'ensemble des éléments de F qui admettent un antécédent par f est une partie de F appelée *ensemble image* ou *image* de f et notée $\mathrm{Im}\, f$:

$$\mathrm{Im}\, f := \{y \in F \mid \exists x \in E \ : \ y = f(x)\}.$$

On écrit également $\mathrm{Im}\, f := \{f(x) \mid x \in E\}$. Il ne faut en principe pas confondre ensemble image et ensemble d'arrivée, tout particulièrement lorsque la question de la surjectivité est en jeu. Pour tout sous-ensemble A de E, l'image $\mathrm{Im}\, f_{|A}$ (pour cette notation, voir l'exemple 4 de la page précédente) de la restriction de f à A est l'ensemble des éléments de F qui ont un antécédent dans A. Ce sous-ensemble de F est appelé *image de A par f* et noté $f(A)$:

$$f(A) := \{y \in F \mid \exists x \in A \ : \ y = f(x)\}.$$

On écrit également $f(A) = \{f(x) \mid x \in A\}$. Ces deux définitions sont liées : on vérifie sans peine que $\mathrm{Im}\, f = f(E)$ et que $f(A) = \mathrm{Im}\, f_{|A}$.

Exemples.

1. L'application Id_E a pour ensemble image E. Une application constante $x \mapsto a$ a pour ensemble image $\{a\}$. L'image de l'application vide est vide. L'image de l'application $x \mapsto x^2$ de $\mathbb{R}$ dans $\mathbb{R}$ est $\mathbb{R}_+$.
2. Selon le module IV.2, l'image par une application continue $f : \mathbb{R} \to \mathbb{R}$ d'un intervalle $I \subset \mathbb{R}$ est un intervalle (théorème des valeurs intermédiaires) ; si I est un segment, c'est-à-dire un intervalle fermé borné, $f(I)$ est lui-même un segment (théorème des bornes).

Un cas particulier intéressant est celui où $E = F$; si $f : E \to E$, on dit que $A \subset E$ est *stable par* f si $f(A) \subset A$. Dans ce cas, par restriction et corestriction on obtient une *application induite sur* A. Par exemple, $\{x\}$ est stable si, et seulement si, $f(x) = x$: on dit alors que x est un *point fixe de* f.
Voici quelques règles de calcul concernant l'image d'une partie de E par une application $f : E \to F$. Ici, A et B désignent des parties de E :

$$\begin{aligned} f(\varnothing) &= \varnothing, \\ f(A \cup B) &= f(A) \cup f(B), \\ f(A \cap B) &\subset f(A) \cap f(B), \\ f(B) \setminus f(A) &\subset f(B \setminus A), \\ A \subset B &\Longrightarrow f(A) \subset f(B). \end{aligned}$$

À titre d'exemple, discutons la troisième règle. Si $x \in A \cap B$, alors $x \in A$ et $x \in B$, donc $f(x) \in f(A)$ et $f(x) \in f(B)$, donc $f(x) \in f(A) \cap f(B)$. Tentons le raisonnement réciproque. Soit $y \in f(A) \cap f(B)$. Alors $y = f(a)$ pour un $a \in A$ et $y = f(b)$ pour un $b \in B$. Si f est injective (*cf.* définition 8 de la page ci-contre), on a nécessairement $a = b \in A \cap B$ et $y \in f(A \cap B)$, ce qui nous apporte une règle supplémentaire :

$$f \text{ injective } \Longrightarrow f(A \cap B) = f(A) \cap f(B).$$

Si l'on ne suppose pas f injective, cette égalité peut être fausse : prendre :

$$E = F = \mathbb{R}, \qquad f : x \mapsto x^2, \qquad A = \mathbb{R}_-^* \quad \text{et} \quad B = \mathbb{R}_+^*.$$

Alors $A \cap B = \varnothing$, donc $f(A \cap B) = \varnothing$, mais $f(A) = f(B) = \mathbb{R}_+^*$, donc $f(A) \cap f(B) = \mathbb{R}_+^*$.

Définition 7. L'application $f : E \to F$ est dite *surjective* si tout élément de F admet (au moins) un antécédent, autrement dit, si $\operatorname{Im} f = F$:

$$f \text{ surjective } \Longleftrightarrow \big(\forall y \in F\ ,\ \exists x \in E\ :\ y = f(x)\big).$$

On parle alors d'application de E *sur* F, ou encore de *surjection*.

Exemples. L'application Id_E est surjective. Une application constante n'est surjective que si son ensemble d'arrivée est un singleton. Soit $f : E \to F$ une application quelconque et soit $F' := \operatorname{Im} f$. Alors la corestriction $f : E \to F'$ est surjective. Par exemple, l'application $x \mapsto x^2$ de $\mathbb{R}$ dans $\mathbb{R}$ n'est pas surjective, alors que sa corestriction à $\mathbb{R}_+$ est surjective.

Théorème 1 (de Cantor).
Il n'existe aucune application surjective d'un ensemble E sur $\mathcal{P}(E)$.

Démonstration. Soit $f : E \to \mathcal{P}(E)$ une application quelconque. Considérons l'ensemble $F := \{x \in E \mid x \notin f(x)\}$. Puisque, pour $x \in E$, $f(x) \subset E$, la condition $x \notin f(x)$ a

un sens. L'axiome de séparation garantit l'existence de $F \subset E$. Nous allons montrer que l'élément F de $\mathcal{P}(E)$ n'est l'image par f d'aucun élément a de E, ce qui impliquera que f n'est pas surjective. Si l'on avait $f(a) = F$ pour un certain $a \in E$, de l'équivalence $\forall x \in E\ ,\ x \in F \Leftrightarrow x \notin f(x)$ (qui est la définition de F) on déduirait, en remplaçant x par a, l'équivalence : $a \in F \Leftrightarrow a \notin f(a)$, c'est-à-dire $a \in F \Leftrightarrow a \notin F$, ce qui est absurde. On ne peut donc avoir $f(a) = F$. ■

Définition 8. On dit que l'application $f : E \to F$ est *injective* (ou que c'est une *injection*) si tout élément de F admet *au plus* un antécédent :

$$f \text{ injective } \iff \big(\forall x, x' \in E\ ,\ f(x) = f(x') \Longrightarrow x = x'\big).$$

On dit que $f : E \to F$ est *bijective* (ou que c'est une *bijection*) si elle est à la fois surjective et injective, autrement dit si tout élément de F admet *exactement* un antécédent :

$$f \text{ bijective } \iff \big(\forall y \in F\ ,\ \exists !x \in E\ :\ y = f(x)\big).$$

Rappelons que la notation $\exists !x$ signifie « il existe un unique x tel que … »

Exemples.

1. L'application Id_E est toujours bijective.
2. Soit $f : E \to F$ une application injective et soit $F' = \operatorname{Im} f$.
   Alors la corestriction $f : E \to F'$ est bijective.
3. Soit I une partie de $\mathbb{R}$. Si $f : I \to \mathbb{R}$ est strictement croissante (ou strictement décroissante), elle est injective.

Exercice 9.
Soient $E = \{a\}$ et $F = \{b, c\}$. Existe-t-il des surjections (resp. des injections, des bijections) de E dans F ?
Solution. Si $b \neq c$, il y a deux applications de E dans F (respectivement définies par $a \mapsto b$ et $a \mapsto c$), et elles sont toutes les deux injectives et non surjectives. Si $b = c$, il n'y a qu'une application (définie par $a \mapsto b$) et elle est bijective.

Définition 9. Supposons l'application $f : E \to F$ bijective. En associant à tout élément $y \in F$ son unique antécédent par f, on définit une application de F dans E. Cette application est appelée *application réciproque de l'application* f (ou simplement *réciproque de* f) et notée f^{-1}. Elle est caractérisée par la relation :

$$\forall x \in E\ ,\ \forall y \in F\ ,\ y = f(x) \Longleftrightarrow x = f^{-1}(y).$$

Il en résulte évidemment que f^{-1} est également bijective et que sa réciproque est $(f^{-1})^{-1} = f$. Pratiquement, on calcule $f^{-1}(y)$ en résolvant l'équation $f(x) = y$ d'inconnue x ; en principe, cette équation doit avoir une unique solution $x = f^{-1}(y)$.

Exemple. La fonction $x \mapsto \sinh x := \dfrac{e^x - e^{-x}}{2}$ de $\mathbb{R}$ dans $\mathbb{R}$ (« sinus hyperbolique ») est strictement croissante (sa dérivée est strictement positive sur $\mathbb{R}$). Elle est donc injective. Comme ses limites en $\pm\infty$ sont $\pm\infty$, l'intervalle $\operatorname{Im} f$ est $\mathbb{R}$. Elle est donc bijective.

Pour calculer $x = f^{-1}(y)$, on résout l'équation $\dfrac{e^x - e^{-x}}{2} = y$ en prenant pour inconnue auxiliaire $X = e^x$. Celle-ci doit avoir pour valeur l'unique racine strictement positive de l'équation $X^2 - 2yX - 1 = 0$, donc $X = y + \sqrt{y^2+1}$; l'application réciproque est donc $y \mapsto \ln\left(y + \sqrt{y^2+1}\right)$.

Définition 10. Soient $f : E \to F$ et $g : F \to G$ deux applications.
L'application $x \mapsto g\big(f(x)\big)$ de E dans G est appelée *composée* de f et g et notée $g \circ f$.

Si de plus $h : G \to H$, alors $h \circ (g \circ f) = (h \circ g) \circ f$: c'est l'application $x \mapsto h\big(g(f(x))\big)$ de E dans H. On prendra garde que, dans la composition $g \circ f$, c'est f qui « agit » en premier, bien que cela se lise « g rond f ». Il peut être utile de visualiser l'action à l'aide d'un diagramme :

$$\begin{array}{ccccc} E & \xrightarrow{f} & F & \xrightarrow{g} & G \\ x & \xmapsto{f} & f(x) & \xmapsto{g} & g\big(f(x)\big). \end{array}$$

Exemples.

1. Pour toute application $f : E \to F$, on a $\mathrm{Id}_F \circ f = f \circ \mathrm{Id}_E = f$.
2. Si f est bijective, on a $f^{-1} \circ f = \mathrm{Id}_E$ et $f \circ f^{-1} = \mathrm{Id}_F$.
3. Si f ou g est constante, alors $g \circ f$ est constante.
4. On ne peut composer $f \circ f$ que si $E = F$.
5. On ne peut composer $f \circ g$ et $g \circ f$ que si la source de l'un est le but de l'autre et réciproquement. On ne peut se poser la question de l'égalité $f \circ g = g \circ f$ que si, en outre, ces deux ensembles sont égaux.
6. Soit $f : \mathbb{R} \to \mathbb{R}$ et soit $g : x \mapsto x + 2\pi$. Alors $f \circ g = g \circ f$ si, et seulement si, l'application $x \mapsto f(x) - x$ est 2π-périodique.

Comme le montre le dernier exemple, on n'a pas, en général $f \circ g = g \circ f$, même si les deux applications ont pour source et but un même ensemble E.

La composée de deux applications injectives (resp. surjectives, resp. bijectives) est une application injective (resp. surjective, resp. bijective). Dans le cas où f et g sont bijectives, on a :

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

Pour le montrer, on résout $(g \circ f)(x) = z$.
Cela équivaut à $g\big(f(x)\big) = z$, donc à $f(x) = g^{-1}(z)$, donc à $x = f^{-1}\big(g^{-1}(z)\big)$.

En fait, il y a de nombreuses règles de ce genre. Soient $f : E \to F$ et $g : F \to G$ deux applications.

1. Si f est bijective, alors $g \circ f$ est injective (resp. surjective, resp. bijective) si, et seulement si, g l'est.
2. Si g est bijective, alors $g \circ f$ est injective (resp. surjective, resp. bijective) si, et seulement si, f l'est.

3. Si $g \circ f$ est injective, alors f est injective.
4. Si $g \circ f$ est surjective, alors g est surjective.

Nous démontrerons seulement les deux dernières, et nous recommandons chaudement au lecteur de prouver les autres à titre d'exercice. Supposons $g \circ f$ injective. Alors on a les implications :

$$f(x) = f(x') \Rightarrow g\big(f(x)\big) = g\big(f(x')\big) \Rightarrow (g \circ f)(x) = (g \circ f)(x') \Rightarrow x = x',$$

d'où l'injectivité de f. Supposons maintenant $g \circ f$ surjective, et soit $z \in G$. Soit $x \in E$ un antécédent de z par $g \circ f$. Alors $f(x) \in F$ est un antécédent de z par g, qui est donc bien surjective.

Définition 11. On définit l'*image réciproque* $f^{-1}(B)$ de $B \subset F$ par l'application $f : E \to F$ comme l'ensemble des antécédents par f des éléments de B :

$$\forall B \subset F\ ,\ f^{-1}(B) := \{x \in E \mid f(x) \in B\}.$$

Il arrive parfois que l'on note $f^{-1}(y)$ l'ensemble $f^{-1}(\{y\})$ des antécédents d'un élément $y \in F$ par f. Avec cette notation, on voit que f surjective (resp. injective, resp. bijective) si, et seulement si, chaque $f^{-1}(y)$ est non vide (resp. vide ou un singleton, resp. un singleton).

Il y a un risque de confusion avec la notation de l'application réciproque (et certains ouvrages utilisent d'ailleurs des notations distinctes). Si f est bijective, l'image réciproque $f^{-1}(B)$ de $B \subset F$ par $f : E \to F$ est égale à l'image $f^{-1}(B)$ de $B \subset F$ par $f^{-1} : F \to E$, et il n'y a pas de problème. Mais l'emploi de l'écriture f^{-1} ne signifie pas nécessairement que f est bijective. D'autre part, si f est bijective et si $x \in E$ est l'unique antécédent de y, il ne faut pas confondre l'image réciproque $f^{-1}(\{y\}) = \{x\}$, parfois notée $f^{-1}(y)$, et l'image de y par l'application réciproque $f^{-1}(y) = x$.

Voici quelques règles de calcul concernant l'image réciproque d'une partie de F par une application $f : E \to F$:

$$\begin{aligned}
f^{-1}(\varnothing) &= \varnothing,\\
f^{-1}(F) &= E,\\
f^{-1}(A \cup B) &= f^{-1}(A) \cup f^{-1}(B),\\
f^{-1}(A \cap B) &= f^{-1}(A) \cap f^{-1}(B),\\
f^{-1}(B \setminus A) &= f^{-1}(B) \setminus f^{-1}(A),\\
A \subset B &\Longrightarrow f^{-1}(A) \subset f^{-1}(B).
\end{aligned}$$

Démontrons par exemple la cinquième règle. On procède par équivalences :
la relation $x \in f^{-1}(B) \setminus f^{-1}(A)$ équivaut à $x \in f^{-1}(B)$ et $x \notin f^{-1}(A)$, c'est-à-dire à $f(x) \in B$ et $f(x) \notin A$, donc à $f(x) \in B \setminus A$, d'où la conclusion.

2.2.1 Diagrammes

Lorsqu'il y a plusieurs ensembles et plusieurs applications entre ces ensembles, il peut être commode d'utiliser des *diagrammes* comme celui-ci :

$$\begin{tikzcd} E \arrow[r, "f"] \arrow[d, "u"] & F \arrow[d, "v"] \\ E' \arrow[r, "f'"] & F' \end{tikzcd}$$

Il arrive fréquemment que l'on veuille exprimer la propriété suivante : si l'on part d'un élément x de l'un de ces ensembles et que l'on calcule ses images successives jusqu'à arriver dans un autre ensemble du diagramme, *le résultat obtenu ne dépend pas du chemin choisi*. On dit alors que *le diagramme est commutatif*. Dans le diagramme ci-dessus, on partirait de $x \in E$, et l'on irait jusqu'à F'. Si le chemin choisi est horizontal puis vertical, le résultat est $v\big(f(x)\big) \in F'$; si le chemin choisi est vertical puis horizontal, le résultat est $f'\big(u(x)\big) \in F'$. Dire que ce diagramme est commutatif signifie donc que $\forall x \in E\ ,\ f'\big(u(x)\big) = v\big(f(x)\big)$, c'est-à-dire que $f' \circ u = v \circ f$.

Exemple. La périodicité d'une fonction de variable réelle peut s'exprimer par un diagramme commutatif. Notons $T_{2\pi}$ la translation $x \mapsto x + 2\pi$ de $\mathbb{R}$ dans lui-même. Une fonction f de $\mathbb{R}$ dans lui-même est 2π-périodique si, et seulement si, le diagramme suivant est commutatif :

$$\begin{tikzcd} \mathbb{R} \arrow[rr, "T_{2\pi}"] \arrow[dr, "f"'] & & \mathbb{R} \arrow[dl, "f"] \\ & \mathbb{R} & \end{tikzcd}$$

Exemple. Anticipons un peu sur la section 4. Supposons que l'on ait une loi de composition interne notée $*$ sur l'ensemble E et une loi de composition interne notée $\star$ sur l'ensemble F. Soit $f : E \to F$ une application.
Dire que f est un morphisme pour les lois $*$ et $\star$ équivaut à dire que le diagramme ci-contre est commutatif, où l'on a respectivement noté m_E et m_F les applications $(x, y) \mapsto x * y$ de $E \times E$ dans E et $(z, t) \mapsto z \star t$ de $F \times F$ dans F.

$$\begin{tikzcd} E \times E \arrow[r, "m_E"] \arrow[d, "f \times f"] & E \arrow[d, "f"] \\ F \times F \arrow[r, "m_F"] & F \end{tikzcd}$$

Les deux diagrammes ci-dessous illustrent ainsi les propriétés correspondantes du logarithme et de l'exponentielle (et l'addition sur $\mathbb{R}$, resp. la multiplication sur $\mathbb{R}_+^*$ ont simplement été notées $+$, resp. $\times$).

$$\begin{tikzcd} \mathbb{R}_+^* \times \mathbb{R}_+^* \arrow[r, "\times"] \arrow[d, "\ln \times \ln"] & \mathbb{R}_+^* \arrow[d, "\ln"] \\ \mathbb{R} \times \mathbb{R} \arrow[r, "+"] & \mathbb{R} \end{tikzcd} \qquad \begin{tikzcd} \mathbb{R} \times \mathbb{R} \arrow[r, "+"] \arrow[d, "\exp \times \exp"] & \mathbb{R} \arrow[d, "\exp"] \\ \mathbb{R}_+^* \times \mathbb{R}_+^* \arrow[r, "\times"] & \mathbb{R}_+^* \end{tikzcd}$$

2.3 L'ensemble $\mathcal{F}(E, F)$ des applications de E dans F

Axiome 11. Les applications de l'ensemble E dans l'ensemble F forment un ensemble, noté $\mathcal{F}(E, F)$.

Exemples.

1. L'ensemble $\mathcal{F}(\varnothing, F)$ est le singleton dont l'unique élément est l'application vide.
2. L'ensemble $\mathcal{F}(\{a\}, F)$ est en bijection naturelle avec F par l'application $f \mapsto f(a)$ de $\mathcal{F}(\{a\}, F)$ dans F. En effet, connaître une application f de $\{a\}$ dans F revient à connaître l'image $f(a) \in F$.

3. Si $a \neq b$, l'ensemble $\mathcal{F}(\{a,b\},F)$ est en bijection naturelle avec F^2 par l'application $f \mapsto \big(f(a),f(b)\big)$ de $\mathcal{F}(\{a,b\},F)$ dans F^2 (connaître une application f de $\{a,b\}$ dans F revient à connaître $f(a),f(b) \in F$).
4. L'ensemble $\mathcal{F}(E,\varnothing)$ est vide si $E \neq \varnothing$ (et un singleton si $E = \varnothing$).
5. L'ensemble $\mathcal{F}(E,\{a\})$ est le singleton dont l'unique élément est l'application constante de valeur a.
6. L'ensemble $\mathcal{F}(E,\{0,1\})$ est en bijection naturelle avec $\mathcal{P}(E)$; en effet, connaître une application f de E dans $\{0,1\}$ revient à connaître le sous-ensemble $A \subset E$ des antécédents de 1, car alors $B = \complement_E A$ est le sous-ensemble des antécédents de 0. Nous retrouverons ce raisonnement en 3.4, lors de l'étude des fonctions caractéristiques.
7. La composition des applications induit une application $(f,g) \mapsto g \circ f$ de $\mathcal{F}(E,F) \times \mathcal{F}(F,G)$ dans $\mathcal{F}(E,G)$.
8. L'application $(f,g) \mapsto < f,g >$ de $\mathcal{F}(E,F) \times \mathcal{F}(E,G)$ dans $\mathcal{F}(E,F \times G)$ est une bijection. Soit $A \subset E$.
9. L'application $f \mapsto f_{|A}$ va de $\mathcal{F}(E,F)$ dans $\mathcal{F}(A,F)$. Si de plus $B \subset E$, on obtient ainsi une application $f \mapsto (f_{|A}, f_{|B})$ de $\mathcal{F}(E,F)$ dans $\mathcal{F}(A,F) \times \mathcal{F}(B,F)$. Si $A \cup B = E$, cette application est injective ; si $A \cap B = \varnothing$, elle est surjective. La preuve de ces affirmations est laissée au lecteur.

On verra dans le module II.1 (théorème 24 page 76) que si E et F sont finis de cardinaux respectifs $n,p \geqslant 1$, alors $\mathcal{F}(E,F)$ est fini de cardinal p^n. Cela explique que l'ensemble $\mathcal{F}(E,F)$ soit parfois noté F^E. Les propriétés élémentaires des puissances :

$$(pp')^n = p^n p'^n,$$
$$p^{n+n'} = p^n p^{n'},$$
$$(p^n)^m = p^{nm}$$

peuvent alors être *interprétées* en termes de bijections. Par exemple, la première et la seconde de ces égalités correspondent respectivement à l'avant-dernière et à la dernière des bijections ci-dessus ; et la troisième à une bijection de $\mathcal{F}(E,\mathcal{F}(F,G))$ dans $\mathcal{F}(E \times F,G)$, qu'on laissera le lecteur expliciter et démontrer. L'analogue des propriétés élémentaires des puissances (rappelées ci-dessus) pour les cardinaux généraux (finis ou infinis) fait l'objet de l'exercice I.1.27.

3 Suites et familles

3.1 Suites d'éléments d'un ensemble

Une notion fondamentale introduite et étudiée dans le secondaire est celle de *suite de nombres réels* : $u_0, u_1, \ldots \in \mathbb{R}$. Chacun des termes de la suite est affecté d'un indice : l'indice du terme général u_n est n. Avant d'en donner une définition formelle, quelques remarques s'imposent :

- Pour distinguer une suite de ses termes, on notera en général $(u_n)_{n\in\mathbb{N}}$ ou $(u_n)_{n\geqslant 0}$ ou encore, pour alléger la notation, tout simplement (u_n). On pourra donc dire au choix « la suite (u_n) » ou « la suite de terme général u_n ».
- Une suite ne commence pas nécessairement au terme d'indice 0. On peut considérer la suite $u_1, u_2, \ldots$, que l'on notera donc au choix $(u_n)_{n\in\mathbb{N}^*}$ ou $(u_n)_{n\geqslant 1}$. On peut considérer plus généralement une suite $(u_n)_{n\geqslant n_0}$ commençant à l'indice $n_0 \in \mathbb{N}$.
- Les termes de la suite ne sont pas nécessairement des réels : on peut s'intéresser à des suites de complexes, de points du plan, de fonctions ...

Dans tous les cas, le processus consiste à associer à chaque entier $n \in \mathbb{N}$ (voire, dans les cas particuliers, $n \in \mathbb{N}^*$, etc) un élément u_n d'un ensemble : $\mathbb{R}$ ou $\mathbb{C}$ ou le plan, etc. On reconnaît une application ! Seule la notation diffère : on écrit u_n au lieu de $u(n)$.

Définition 12. On appelle *suite d'éléments de l'ensemble* E une application u de $\mathbb{N}$ ou de $\mathbb{N}^*$ dans E. L'image $u(n)$ est notée u_n et appelée *terme* de la suite et n est son *indice* ; on dit encore que le terme u_n est *indexé* par n. L'ensemble $\mathcal{F}(\mathbb{N}, E)$ des suites indexées par $\mathbb{N}$ est alors noté $E^{\mathbb{N}}$. De même, l'ensemble $\mathcal{F}(\mathbb{N}^*, E)$ des suites indexées par $\mathbb{N}^*$ est noté $E^{\mathbb{N}^*}$, etc.

Comme observé plus haut, on admet en fait d'autres ensembles de départ, de la forme $\{n_0, n_0+1, \ldots\} = \mathbb{N} \setminus \{0, \ldots, n_0 - 1\}$. Il arrive au contraire que l'on ne considère que des termes indexés par les indices $0, \ldots, n$, resp. $1, \ldots, n$: on parle alors d'une *suite finie*, notée $(u_0, \ldots, u_n)$, resp. $(u_1, \ldots, u_n)$.
Soit $J \subset \mathbb{N}$ un sous-ensemble infini de $\mathbb{N}$. La restriction à J de l'application $n \mapsto u_n$ peut être considérée comme une *suite extraite*, également appelée *sous-suite*. Par exemple, si $(u_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ est une suite de réels, en se restreignant à l'ensemble J des indices pairs $2\mathbb{N} \subset \mathbb{N}$, on obtient une suite extraite $(u_n)_{n\in 2\mathbb{N}}$, que l'on notera plutôt $(u_{2n})_{n\in\mathbb{N}}$.
Pratiquement, pour définir une sous-suite de la suite $(u_n)_{n\in\mathbb{N}}$, on définit une suite strictement croissante d'indices : $n_0 < n_1 < \cdots$, et l'on considère la suite $(u_{n_k})_{k\in\mathbb{N}}$. Un exemple détaillé est traité dans l'exercice I.1.0 à la fin de ce module.

3.2 Familles d'éléments d'un ensemble

Nous allons dans cette section étendre la notion de suite au cas où les indices parcourent un ensemble encore plus général.

Définition 13. Une *famille d'éléments de l'ensemble* E *indexée par l'ensemble* I est une application de I dans E. On note sous la forme $\underline{x} := (x_i)_{i\in I}$ une telle famille, l'image de l'*indice* $i \in I$ étant l'élément $x_i \in E$. L'ensemble $\mathcal{F}(I, E)$ de ces familles est alors noté E^I.

Ce changement de notation est une opération psychologique : on met en avant les *valeurs* $x_i \in E$, le rôle des *indices* $i \in I$ étant de les repérer.

Exemples.

1. En prenant $I := \mathbb{N}$ ou $I := \mathbb{N}^*$, on retrouve la notion de suite :
$$(x_n)_{n\geqslant 0} := (x_0, x_1, x_2, \ldots) \quad \text{ou} \quad (x_n)_{n\geqslant 1} := (x_1, x_2, \ldots).$$
2. En prenant $I := \{0, 1, 2, \ldots, n\}$ ou $I := \{1, 2, \ldots, n\}$ on retrouve la notion de suite finie $(x_0, x_1, x_2, \ldots, x_n)$ ou $(x_1, x_2, \ldots, x_n)$.
3. Soit $J \subset I$ un sous-ensemble d'indices. Par restriction à J de l'application $i \mapsto x_i$, on obtient une famille $(x_i)_{i\in J}$ indexée par J : on dit que c'est une *famille extraite*. Si $I = \mathbb{N}$ et $J \subset \mathbb{N}$, on reconnaît la notion de suite extraite ou sous-suite.
4. Soit $\varphi : K \to I$ une application entre les ensembles d'indices K et I. Par composition de φ avec l'application $i \mapsto x_i$, on définit une famille $(x_{\varphi(k)})_{k\in K} \in E^K$. Dans le cas où φ est injective, c'est une bijection de K sur un sous-ensemble J de I, et cette construction est équivalente à la précédente. Par exemple, prenant $K = I = \mathbb{N}$ et $\varphi(n) = 2n$, on voit que φ est une bijection de $\mathbb{N}$ sur $2\mathbb{N} \subset \mathbb{N}$. Si $(u_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ est une suite de réels, on retrouve la suite extraite $(u_n)_{n\in 2\mathbb{N}}$, que l'on notera plutôt $(u_{2n})_{n\in\mathbb{N}}$.
5. Si l'ensemble I des indices est un produit cartésien : $I := J \times K$, il est d'usage d'employer la « biindexation » $(x_{j,k})_{j\in J, k\in K} \in E^{J\times K}$. C'est le cas de la famille des coefficients d'une matrice (module II.4).
6. Si I est l'ensemble vide (ce qui peut arriver), on parle de *famille vide*.

Malgré son apparente lourdeur, la notion de famille d'éléments indexée par un ensemble d'indices est d'une extrême commodité. Ainsi, lorsque l'on veut opérer sur tous les éléments d'un ensemble quelconque E, il est fréquent que l'on préfère les considérer comme les termes x_i d'une famille. Il suffit pour cela de prendre $I = E$ et de poser $x_i = i$, autrement dit de prendre la famille qui correspond à l'application Id_E. On dit que c'est *la famille canoniquement associée à l'ensemble* E. L'usage est de la noter $(x_i)_{i\in I}$, comme n'importe quelle famille, mais en gardant à l'esprit que chaque élément de E y apparaît une fois et une seule. Par exemple, en algèbre linéaire, on considérera parfois les bases comme des ensembles de vecteurs, parfois comme des familles de vecteurs : dans ce dernier cas, on aura l'usage de notations comme $\sum\limits_{i\in I} \lambda_i x_i$, etc.

3.3 Familles d'ensembles

Nous appellerons *famille d'ensembles indexée par l'ensemble* I la donnée, pour chaque indice $i \in I$, d'un ensemble E_i. Ce n'est pas réellement une définition car nous n'attribuons aucune signification précise à la manière dont sont « donnés » les ensembles E_i. Une telle famille est notée $(E_i)_{i\in I}$.

Axiome et définition 12. Soit $(E_i)_{i\in I}$ une famille d'ensembles. Il existe un ensemble G tel que :
$$\forall x\ ,\ x \in G \iff \big(\exists i \in I\ :\ x \in E_i\big).$$
Cet ensemble est unique (axiome d'extensionalité).
On l'appelle *réunion* de la famille $(E_i)_{i\in I}$, et on le note $\bigcup\limits_{i\in I} E_i$.

On a donc :

$$\bigcup_{i\in I} E_i := \{x \mid \exists i \in I \ : \ x \in E_i\}.$$

Par exemple, la réunion d'une famille vide d'ensembles est vide (ainsi d'ailleurs que la réunion d'une famille d'ensembles vides).

Nous avions postulé que les éléments x_i appartenaient tous à un même ensemble E donné au départ. En réalité, il est possible de définir des familles $(x_i)_{i\in I}$, chaque x_i étant pris dans un ensemble E_i *qui dépend de* i. En vertu de l'axiome ci-dessus, cette définition n'est pas vraiment plus générale : on peut en effet considérer $(x_i)_{i\in I}$ comme une famille d'éléments de l'ensemble $E := \bigcup_{i\in I} E_i$. Pour la même raison, toute famille d'ensembles peut être considérée comme une famille de parties d'un certain ensemble E, *i.e.* comme une famille d'éléments de $\mathcal{P}(E)$.

Si $(E_i)_{i\in I}$ est une famille d'ensembles et que I n'est pas vide, on déduit de l'axiome de séparation (appliqué à l'intérieur de l'un quelconque des E_i) qu'il existe un ensemble G tel que :

$$\forall x \ , \ x \in G \Longleftrightarrow \big(\forall i \in I \ , \ x \in E_i\big).$$

Cet ensemble est unique (axiome d'extensionalité). On l'appelle *intersection* de la famille $(E_i)_{i\in I}$, et on le note $\bigcap_{i\in I} E_i$:

$$\bigcap_{i\in I} E_i := \{x \mid \forall i \in I \ , \ x \in E_i\}.$$

En revanche, on ne sait pas définir l'intersection d'une famille vide d'ensembles. La formule ci-dessus déterminerait dans ce cas un ensemble qui contient tous les x, sans aucune condition.

Soit $(E_i)_{i\in I}$ une famille d'ensembles de réunion $E = \bigcup_{i\in I} E_i$. Le sous-ensemble de E^I formé des familles $(x_i)_{i\in I}$ telles que chaque x_i appartient à l'ensemble E_i est appelé *produit cartésien* ou simplement *produit* de la famille $(E_i)_{i\in I}$, et noté $\prod_{i\in I} E_i$.

$$\prod_{i\in I} E_i := \{(x_i)_{i\in I} \mid \forall i \in I \ , \ x_i \in E_i\}.$$

Pour chaque indice $j \in I$, on dispose alors d'une application $j^{\text{ème}}$ *projection* de $\prod_{i\in I} E_i$ sur E_j, qui, à l'élément $(x_i)_{i\in I}$, associe sa $j^{\text{ème}}$ *composante* ou $j^{\text{ème}}$ *coordonnée* x_j. Le lecteur vérifiera sans peine que, lorsque I est fini, on retrouve les notions déjà rencontrées. Par exemple, si $I = \{1, \ldots, n\}$:

$$\bigcup_{i\in I} E_i = E_1 \cup \ldots \cup E_n, \quad \bigcap_{i\in I} E_i = E_1 \cap \ldots \cap E_n, \quad \prod_{i\in I} E_i = E_1 \times \ldots \times E_n.$$

Il est bien évident que, si l'un des E_i est vide, alors le produit $\prod_{i\in I} E_i$ l'est également. Pour démontrer la réciproque dans le cas d'un produit fini, on avait supposé les E_i non vides et « choisi » un élément dans chacun d'eux pour former une famille $(x_i)_{i\in I}$ particulière qui soit un élément du produit. Au début du vingtième siècle, la possibilité de « choisir » une infinité de tels éléments simultanément et de manière arbitraire a été mise en doute par de nombreux mathématiciens. La résolution moderne de ce conflit, qui frise la métaphysique, a été d'introduire un axiome *ad hoc*. Malgré son allure

fumeuse, cet axiome comporte pas mal de conséquences intéressantes, comme nous le rappellerons à la section 6.2.

Axiome 13 (Axiome du choix). Si les E_i sont non vides, alors $\prod_{i\in I} E_i$ l'est également.

3.4 Familles de parties d'un ensemble

Beaucoup des règles de calcul que nous avons énoncées dans le cas de deux ensembles (ou d'un nombre fini d'ensembles) s'étendent ici. Il ne serait pas raisonnable d'énoncer toutes ces généralisations. En voici quelques exemples. Pour formuler l'associativité de la réunion et de l'intersection généralisées, on considère une famille $(E_i)_{i\in I}$ d'ensembles et l'on suppose que l'ensemble I de ses indices est lui-même une réunion d'ensembles : $I = \bigcup_{j\in J} I_j$.

On a alors :

$$\bigcup_{i\in I} E_i = \bigcup_{j\in J}\left(\bigcup_{i\in I_j} E_i\right), \qquad \bigcap_{i\in I} E_i = \bigcap_{j\in J}\left(\bigcap_{i\in I_j} E_i\right).$$

Supposons maintenant que les E_i sont des parties d'un ensemble E :

$$\complement_E\left(\bigcup_{i\in I} E_i\right) = \bigcap_{i\in I} \complement_E E_i, \qquad \complement_E\left(\bigcap_{i\in I} E_i\right) = \bigcup_{i\in I} \complement_E E_i.$$

Supposons en outre que les F_j $(j \in J)$ sont des sous-ensembles d'un ensemble F et que f est une application de E dans F :

$$f\left(\bigcup_{i\in I} E_i\right) = \bigcup_{i\in I} f(E_i), \qquad f\left(\bigcap_{i\in I} E_i\right) \subset \bigcap_{i\in I} f(E_i),$$

$$f^{-1}\left(\bigcup_{j\in J} F_j\right) = \bigcup_{j\in J} f^{-1}(F_j), \qquad f^{-1}\left(\bigcap_{j\in J} F_j\right) = \bigcap_{j\in J} f^{-1}(F_j).$$

On prendra garde que la deuxième de ces règles comporte une inclusion et non une égalité. Mentionnons enfin que la bijection $(f,g) \mapsto \langle f,g\rangle$ de $\mathcal{F}(E,F) \times \mathcal{F}(E,G)$ sur $\mathcal{F}(E, F\times G)$ rencontrée antérieurement se généralise en une bijection de $\prod_{j\in J} \mathcal{F}(E,F_j)$ sur $\mathcal{F}(E, \prod_{j\in J} F_j)$. Le lecteur courageux démontrera (sans peine) toutes ces règles et en inventera bien d'autres.

3.4.1 Partitions et recouvrements

On définit un *recouvrement* d'un ensemble E comme une famille $(E_i)_{i\in I}$ de parties de E telle que $\bigcup_{i\in I} E_i = E$. Cette notion interviendra surtout en topologie. On définit une *partition* d'un ensemble E comme un recouvrement de E par des ensembles non vides et deux à deux disjoints. On a donc :

$$\bigcup_{i\in I} E_i = E, \qquad \forall i \in I\ ,\ E_i \neq \varnothing, \qquad \forall i \neq j \in I\ ,\ E_i \cap E_j = \varnothing.$$

3.4.2 Parties et applications caractéristiques

On fixe un ensemble E. On se propose ici de « coder » les parties de E à l'aide d'applications. À toute partie $A \subset E$, on associe sa *fonction caractéristique* χ_A (en tant que partie de E) de la manière suivante. C'est une application $\chi_A : E \to \{0,1\}$ définie par les relations :

$$\forall x \in E \ , \ \chi_A(x) := \begin{cases} 1 \text{ si } x \in A, \\ 0 \text{ si } x \notin A. \end{cases}$$

Théorème 2. L'application $A \mapsto \chi_A$ de $\mathcal{P}(E)$ dans $\mathcal{F}\big(E, \{0,1\}\big)$ est bijective.

Démonstration. Soit $f : E \to \{0,1\}$ un élément de $\mathcal{F}\big(E, \{0,1\}\big)$. Soit $A := f^{-1}(1) \subset E$. Il est alors clair que $\chi_A = f$, et, plus précisément, que A est l'unique antécédent de f par l'application en question ; cette dernière est donc bijective. ■

Ainsi, si $E = \{1, \ldots, n\}$, se donner une partie de E revient à se donner n entiers $x_1, \ldots, x_n \in \{0,1\}$, autrement dit, en langage informatique, un « vecteur de bits ». C'est la base de la représentation des ensembles dans le langage de programmation Pascal. Ce théorème permet de donner une interprétation intéressante du théorème 1 de la page 14. Puisque l'on a une bijection de $\mathcal{F}\big(E, \{0,1\}\big)$ sur $\mathcal{P}(E)$, le théorème de Cantor équivaut à la non existence d'une surjection de E sur $\mathcal{F}\big(E, \{0,1\}\big)$. On peut démontrer cette non existence à l'aide du *procédé diagonal de Cantor*. Soit φ une application quelconque de E dans $\mathcal{F}\big(E, \{0,1\}\big)$. Ainsi, pour tout $a \in E$, $\varphi(a)$ est une application de E dans $\{0,1\}$. On peut alors définir une application f de E dans $\{0,1\}$ comme suit :

$$\forall a \in E \ , \ f(a) := \begin{cases} 1 \text{ si } \varphi(a)(a) = 0 \\ 0 \text{ si } \varphi(a)(a) = 1 \end{cases}$$

Par construction, $f(a) \neq \varphi(a)(a)$. *A fortiori*, $f \neq \varphi(a)$. Cela étant vrai quel que soit $a \in E$, f n'est pas dans l'image de φ qui n'est donc pas surjective. Le lecteur est encouragé à réfléchir à cette démonstration, afin d'y reconnaître une simple traduction de la démonstration du théorème 1 de la page 14 en termes de fonctions caractéristiques. Nous verrons à la section suivante que l'usage des fonctions caractéristiques permet de ramener le calcul ensembliste au calcul dans $\{0,1\}$.

4 Lois de composition

Les notions qui apparaissent ici seront étudiées en plus grand détail au module II.2, où apparaîtront également de nombreux exemples.

4.1 Vocabulaire général

On appelle *loi de composition interne* sur un ensemble E une application de $E \times E$ dans E. L'usage est de les noter en « forme infixe ». Cela signifie que l'on introduit un symbole distinct de ceux qui servent à dénoter les éléments de E, par exemple $+$, $\times$, $*$, $\star$ ou $\top$ et que l'on écrit l'image de $(a,b) \in E \times E$ dans E sous la forme $a + b$, $a \times b$, $a * b$, $a \star b$ ou $a \top b$ selon le cas. Dans certains cas, qui s'apparentent à la multiplication des nombres, on peut même omettre tout symbole d'opération et écrire $a.b$ ou même ab au lieu de $a \times b$. Lorsque la loi est notée $+$, on parle de notation *additive* et les composés d'éléments sont

appelés *sommes*. Dans les autres cas, on parle de notation *multiplicative* et les composés sont plutôt appelés *produits*.

Les exemples les plus célèbres de lois de composition interne sont l'addition et la multiplication sur les ensembles de nombres $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ et $\mathbb{C}$; mais l'on peut également citer la réunion et l'intersection sur $\mathcal{P}(E)$.

Soit E un ensemble muni d'une loi de composition interne notée $\star$. On dit qu'une partie A de E est *stable pour la loi* $\star$ si l'image de $A \times A$ est incluse dans A :

$$\forall (a,b) \in A \times A \ , \ a \star b \in A.$$

Dans ce cas, l'application induite de $A \times A$ dans A est une loi de composition interne sur A, appelée *loi induite* sur A ; en général, on la note de la même manière que la loi sur E, donc ici $\star$. Par exemple, chacune des inclusions $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$ définit une partie stable pour l'addition et pour la multiplication. De même, si $E' \subset E$, l'ensemble $\mathcal{P}(E')$ est une partie de $\mathcal{P}(E)$ qui est stable pour la réunion et l'intersection.

Avec les mêmes notations, soient A et B deux parties de E.
On définit alors l'ensemble $A \star B \subset E$ comme l'image de $A \times B \subset E \times E$ par la loi $\star$:

$$A \star B := \{a \star b \mid a \in A \ , \ b \in B\}.$$

Cela permet d'*étendre* la loi $\star$ à $\mathcal{P}(E)$.

Soient E et F deux ensembles respectivement munis des lois de composition $*$ et $\star$. On peut munir le produit $E \times F$ d'une loi de composition $\top$ en posant : $(a_1, b_1) \top (a_2, b_2) = (a_1 * a_2, b_1 \star b_2)$. La loi $\top$ est alors appelée *loi produit des lois* $*$ *et* $\star$. Cette construction se généralise au produit cartésien d'une famille $(E_i)_{i \in I}$ d'ensembles. Supposons chacun des E_i muni d'une loi de composition, et que toutes ces lois sont notées $\star$ (cela se produit fréquemment et ne cause en général pas de confusion). Le produit de deux éléments $(x_i)_{i \in I}, (y_i)_{i \in I} \in \prod_{i \in I} E_i$ est alors l'élément $(x_i \star y_i)_{i \in I} \in \prod_{i \in I} E_i$, dans lequel chaque composante $x_i \star y_i$ a bien sûr été calculée avec la loi $\star$ de E_i.

Par exemple, si E est muni de la loi $\star$ (maintenant unique !), le produit E^I est muni de la loi produit que nous noterons encore $\star$ et qui est définie par la formule : $(x_i)_{i \in I} \star (y_i)_{i \in I} = (x_i \star y_i)_{i \in I}$. Rappelons que nous avons, à la section 3, identifié les ensembles E^I et $\mathcal{F}(I, E)$. À partir de la loi $\star$ sur E nous avons donc également défini une loi sur $\mathcal{F}(I, E)$. Celle-ci est ainsi définie : soient f et g deux applications de I dans E ; alors $f \star g$ est l'application $i \mapsto f(i) \star g(i)$ de I dans E.

4.1.1 Propriétés spéciales des lois de composition

On dit que la loi $\star$ sur l'ensemble E est *associative* si l'on a :

$$\forall a, b, c \in E \ , \ a \star (b \star c) = (a \star b) \star c.$$

On peut alors écrire plus simplement $a \star b \star c$. Cette convention se généralise au produit de n éléments quelconques. Par exemple, le produit de n éléments tous égaux à a est souvent noté a^n (qui se lit « a puissance n », et l'on dit alors que n est l'*exposant*). Toutes les lois que nous avons mentionnées (addition et multiplication de nombres, réunion et intersection

de parties) sont associatives. Les puissances d'un élément et, plus généralement, les produits de familles d'éléments obéissent à de nombreuses lois algébriques qui seront étudiées au module II.2.

Exemple. Lorsque la loi n'est pas supposée associative, notons c_n le nombre de produits *a priori* distincts que l'on peut former en utilisant $n+1$ fois le seul élément a. Pour $n=0$, il n'y a que a, donc $c_0 = 1$. Pour $n=1$, il n'y a que $a \star a$, donc $c_1 = 1$. Pour $n=2$, il y a $a \star (a \star a)$ et $(a \star a) \star a$ donc $c_2 = 2$. Les c_n sont les *nombres de Catalan* (exercice I.1.13 de la page 58).

On dit que la loi $\star$ est *commutative* si l'on a :

$$\forall a, b \in E \ , \ a \star b = b \star a.$$

Toutes les lois que nous avons mentionnées (addition et multiplication de nombres, réunion et intersection de parties) sont commutatives.

On dit que $e \in E$ est *neutre à gauche* ou *un élément neutre à gauche* pour la loi $\star$ si l'on a $e \star a = a$ quel que soit $a \in E$. On définit de même *neutre à droite* par $a \star e = a$. On appelle *neutre* ou *élément neutre* tout élément neutre à gauche et à droite. Toutes les lois que nous avons mentionnées (addition et multiplication de nombres, réunion et intersection de parties) admettent un élément neutre ; nous laissons au lecteur le soin de préciser dans chaque cas lequel.

Exemple. S'il y a un élément neutre à gauche e et un élément neutre à droite e', ils sont égaux : on a $e \star e' = e'$ et $e \star e' = e$, donc $e = e'$. En particulier, *s'il y a un élément neutre, il est unique*.

Supposons que la loi $\star$ sur E admette un élément neutre. On dit alors que $a \in E$ est *inversible à gauche* s'il admet un *inverse à gauche*, c'est-à-dire un élément a' tel que $a' \star a = e$. On dit que a est *inversible à droite* s'il s'il admet un *inverse à droite* (ou symétrique), c'est-à-dire un élément a'' tel que $a \star a'' = e$. On dit que a est inversible s'il admet un *inverse*, c'est-à-dire un élément qui est son inverse à gauche et à droite.

Il est fréquent de noter *additivement*, autrement dit par le symbole $+$ une loi de composition commutative. Dans ce cas, le symétrique d'un élément a (s'il existe) est presque toujours appelé *opposé* (et non inverse) de a et généralement noté $-a$. L'appelation *inverse* est en principe réservée au cas d'une loi notée *multiplicativement*, autrement dit, par le symbole $\times$, le symbole $.$ ou avec symbole omis.

Exemple. Supposons la loi associative. Si a admet un inverse a' à gauche et un inverse a'' à droite, ils sont égaux : les deux manières de calculer $a' \star a \star a''$ donnent $a' = a''$. Il en découle que *si la loi est associative, l'inverse d'un élément, s'il existe, est unique*.

Une notion moins fondamentale est celle d'élément *simplifiable à gauche* : c'est un élément a tel que $a \star x = a \star y \Rightarrow x = y$.
De même, a est *simplifiable à droite* si $x \star a = y \star a \Rightarrow x = y$.
On dit que a est *simplifiable* (ou régulier) s'il est simplifiable à gauche et à droite.
Exemple. Si la loi est associative et munie d'un élément neutre, tout élément inversible à gauche (resp. à droite) est simplifiable à gauche (resp. à droite). En effet, avec les notations antérieures, si $a \star x = a \star y$, en composant à gauche par a', on trouve (grâce à l'associati-

vité) $(a' \star a) \star x = (a' \star a) \star y$, c'est-à-dire (puisque $a' \star a = e$) $x = y$. Le calcul du côté droit est similaire.

Remarquons cependant que tous les éléments de $\mathbb{N}$ (resp. les éléments non nuls de $\mathbb{Z}$) sont simplifiables pour l'addition (resp. pour la multiplication), mais qu'aucun n'est inversible, à l'exception de 0 (resp. de ± 1). Pour l'addition dans $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ (resp. pour la multiplication dans $\mathbb{Q}, \mathbb{R}, \mathbb{C}$), tous les éléments sont inversibles (resp. tous les éléments non nuls).

Voici encore deux notions d'usage moins fréquent. On dit que a est *idempotent* si $a \star a = a$. Par exemple, tout neutre est idempotent (et, dans $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ munis de l'addition, ce sont les seuls, alors que pour la multiplication il y a en plus 0). On dit que a est *absorbant à gauche* (resp. *absorbant à droite*) si l'on a $a \star b = a$ quel que soit b (resp. $b \star a = a$). Par exemple, 0 est absorbant pour la multiplication dans $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$.

4.1.2 Morphismes

Nous allons considérer simultanément deux ensembles munis de lois de composition interne, afin d'exprimer les liens possibles entre les « structures » ainsi définies.
Nous noterons $(E, \star)$ et $(F, \top)$ ces deux structures.

Définition 14. Un *morphisme* de $(E, \star)$ dans $(F, \top)$ est une application f de E dans F telle que :

$$\forall a, b \in E \ : \ f(a \star b) = f(a) \top f(b).$$

Un morphisme bijectif est appelé *isomorphisme*. S'il existe un isomorphisme de $(E, \star)$ dans $(F, \top)$, on dit que $(E, \star)$ et $(F, \top)$ sont *isomorphes*. Un morphisme de $(E, \star)$ dans lui-même (avec la même loi !) est appelé *endomorphisme*. Un isomorphisme de $(E, \star)$ dans lui-même (avec la même loi !) est appelé *automorphisme*.

Les vérifications (faciles) des faits suivants sont laissées en exercice au lecteur. Le composé de deux morphismes est un morphisme. L'identité, le composé de deux isomorphismes, l'application réciproque d'un isomorphisme sont des isomorphismes. La relation « être isomorphe » est donc une relation d'équivalence (au sens de la section 5).

Exemples.

1. L'application exponentielle est un morphisme de $\mathbb{R}$ muni de l'addition dans $\mathbb{R}_+^*$ muni de la multiplication.
2. L'application logarithme népérien est un morphisme de $\mathbb{R}_+^*$ muni de la multiplication dans $\mathbb{R}$ muni de l'addition.
3. Fixons le réel strictement positif a. L'application $n \mapsto a^n$ est un morphisme de $\mathbb{Z}$ muni de l'addition dans $\mathbb{R}_+^*$ muni de la multiplication.
4. Fixons l'entier naturel n. L'application $a \mapsto a^n$ est un automorphisme de $\mathbb{R}_+^*$ muni de la multiplication.
5. L'application $A \mapsto \complement_E A$ de $\mathcal{P}(E)$ dans lui-même est un isomorphisme de $\big(\mathcal{P}(E), \cap\big)$ dans $\big(\mathcal{P}(E), \cup\big)$ et également un isomorphisme de $\big(\mathcal{P}(E), \cup\big)$ dans $\big(\mathcal{P}(E), \cap\big)$ (lois de Morgan). Ce n'est cependant pas un automorphisme car la loi n'est pas la même au départ et à l'arrivée.

Si $(E, \star)$ et $(F, \top)$ sont isomorphes, alors l'une de ces deux lois est associative (resp. commutative) si, et seulement si, l'autre l'est. De même, les isomorphismes transforment neutres (à gauche ou à droite) en neutres (à gauche ou à droite), inverses (à gauche ou à droite) en inverses (à gauche ou à droite), simplifiables (à gauche ou à droite) en simplifiables (à gauche ou à droite), absorbants en absorbants, idempotents en idempotents.

4.2 Application au calcul ensembliste

Nous avons rencontré en 3.4 la bijection $A \mapsto \chi_A$ de $\mathcal{P}(E)$ dans $\mathcal{F}(E, \{0,1\})$. Nous allons voir que, quitte à munir $\mathcal{F}(E, \{0,1\})$ de quelques lois de composition très naturelles, cette bijection est, en bien des sens, un isomorphisme. On commence pour cela par munir $\{0,1\}$ de diverses opérations, notées $\cup$, $\cap$ et $\oplus$. Par les constructions générales qui précèdent, ces opérations s'étendent automatiquement à $\mathcal{F}(E, \{0,1\})$. Pour définir des lois sur $\{0,1\}$, nous écrirons leurs tables :

$\cup$	0	1
0	0	1
1	1	1

$\cap$	0	1
0	0	0
1	0	1

$\oplus$	0	1
0	0	1
1	1	0

(3)

On vérifie alors que la bijection mentionnée plus haut est un isomorphisme de $\big(\mathcal{P}(E), \cup\big)$ sur $(\mathcal{F}(E, \{0,1\}), \cup)$, de $\big(\mathcal{P}(E), \cap\big)$ sur $(\mathcal{F}(E, \{0,1\}), \cap)$, de $\big(\mathcal{P}(E), \oplus\big)$ sur $(\mathcal{F}(E, \{0,1\}), \oplus)$. La loi $\oplus$ sur $\mathcal{P}(E)$ est la *différence symétrique* définie par : $A \oplus B := (A \backslash B) \cup (B \backslash A)$. On vérifiera au module II.2 que $\big(\mathcal{P}(E), \oplus\big)$ est un groupe (et même un anneau, avec la loi $\cap$) et au module II.3 que c'est un espace vectoriel sur le corps à deux éléments. Mentionnons une autre application du calcul dans $\{0,1\}$: la démonstration de la formule du crible au module II.2.

5 Relations

Nous avons manipulé, depuis le début de ce module, des propriétés $P(x)$ mettant en jeu un seul objet inconnu x sur lequel portait la discussion (des « prédicats ») ; par exemple $x \in E$, $(x = a$ ou $x = b)$, etc. On peut tout aussi bien considérer des propriétés $R(x, y)$ mettant en jeu *deux* objets inconnus x et y, par exemple $x < y$, $x \in y$, $x \subset y$, $x \equiv y \pmod 9$, etc. voire des propriétés $R(x_1, \ldots, x_n)$ mettant en jeu *plusieurs* objets inconnus $x_1, \ldots, x_n$, par exemple « les entiers $x_1, \ldots, x_n$ sont premiers entre eux dans leur ensemble » ou encore « les fonctions $f_1, \ldots, f_n$ n'ont pas de zéro commun ».

Nous nous intéresserons principalement aux propriétés de la forme $R(x, y)$, que nous appellerons *relations binaires*. Des exemples de telles relations sont $x = y$, $x \in y$ et $x \subset y$. Cependant, de même que l'axiome de séparation concerne un prédicat $P(x)$ défini sur un ensemble, nous distinguerons les relations $R(x, y)$ qui n'ont de sens que lorsque x est élément d'un certain ensemble E et y est élément d'un certain ensemble F. Une relation entre éléments de E et éléments de F est appelée *correspondance entre E et F*. Comme exemple, nous avons la relation $y = f(x)$ lorsque f est une application de E dans F, mais aussi la relation $x = g(y)$ lorsque g est une application de F dans E et même la relation $\varphi(x, y) = 0$ lorsque φ est une application de $E \times F$ dans $\mathbb{R}$ (par exemple).

Définition 15. Le *graphe* de la correspondance $R(x,y)$ entre E et F est l'ensemble :

$$\Gamma_R \quad := \quad \{(x,y) \in E \times F \mid R(x,y)\}.$$

Par exemple, le graphe de la correspondance $y = f(x)$ est le graphe Γ_f de l'application f. Réciproquement, pour toute partie $G \subset E \times F$, on peut définir une correspondance entre E et F dont le graphe est G : il suffit de prendre pour $R(x,y)$ la relation $(x,y) \in G$.

5.1 Relations binaires sur un ensemble

Les relations binaires les plus utiles en mathématiques portent sur des éléments x, y d'un même ensemble E. On parle alors de *relation binaire sur* E. Si F est une partie de E, la *relation induite par* R *sur* F est simplement la relation $R(x,y)$ entre éléments de F. On peut d'ailleurs partir d'une relation binaire quelconque entre objets de même nature, par exemple $x \subset y$, et considérer la relation induite sur un ensemble, par exemple ici $\mathcal{P}(E)$.
Pour une relation binaire $R(x,y)$ sur un ensemble E, l'usage veut que l'on emploie très souvent une notation « infixe » $x \mathcal{R} y$ au lieu de $R(x,y)$, ce que nous ferons désormais. À titre d'exemple, nous connaissons les relations $x \leqslant y$ et $x < y$ sur $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$, la relation $x \mid y$ (« x divise y ») sur $\mathbb{N}, \mathbb{Z}$ ou même sur l'anneau $K[X]$ des polynômes sur le corps K (module II.6), la relation $x \subset y$ sur $\mathcal{P}(E)$, etc.

Définition 16. On dit que la relation $\mathcal{R}$ est *réflexive* si :

$$\forall x \in E \ , \ x \mathcal{R} x.$$

On dit que la relation $\mathcal{R}$ est *transitive* si :

$$\forall x, y, z \in E \ , \ (x \mathcal{R} y \quad \text{et} \quad y \mathcal{R} z) \Rightarrow x \mathcal{R} z.$$

On dit que la relation $\mathcal{R}$ est *symétrique* si :

$$\forall x, y \in E \ , \ x \mathcal{R} y \Rightarrow y \mathcal{R} x.$$

On dit que la relation $\mathcal{R}$ est *antisymétrique* si :

$$\forall x, y \in E \ , \ (x \mathcal{R} y \quad \text{et} \quad y \mathcal{R} x) \Rightarrow x = y.$$

Une *relation d'équivalence* est une relation réflexive, transitive et symétrique.
Une *relation d'ordre* est une relation réflexive, transitive et antisymétrique.

Ces notions ont été définies pour des relations binaires sur un ensemble, mais elles s'étendent sans problème à des relations $R(x,y)$ quelconques. Il est par exemple évident que la relation $x = y$ est une relation d'équivalence et que la relation $x \subset y$ est une relation d'ordre, sans qu'il soit nécessaire de préciser un ensemble E de référence.

Exemples.

1. Les relations d'égalité (sur n'importe quel ensemble) et de congruence modulo a (sur $\mathbb{Z}$) sont des relations d'équivalence.
2. Les relations $\leqslant$ sur $\mathbb{R}$, $\subset$ sur $\mathcal{P}(E)$, et $\mid$ (divisibilité) sur $\mathbb{N}$ sont des relations d'ordre.
3. La seule relation à la fois d'ordre et d'équivalence est l'égalité.

On vérifie immédiatement que la relation induite sur un ensemble E par une relation réflexive (resp. transitive, resp. symétrique, resp. antisymétrique, resp. d'équivalence, resp. d'ordre) est elle-même réflexive (resp. transitive, resp. symétrique, resp. antisymétrique, resp. d'équivalence, resp. d'ordre). Nous étudierons les relations d'équivalence et les relations d'ordre dans la suite de cette section.

5.1.1 Relations engendrées, clôtures

Puisqu'une relation binaire sur un ensemble E est essentiellement déterminée par son graphe, qui est une partie quelconque de $E \times E$, on peut appliquer à ces relations la plupart des notions définies sur $\mathcal{P}(E \times E)$. Soient $\mathcal{R}$ et $\mathcal{R}'$ deux relations binaires sur un même ensemble E. Il peut arriver que l'une entraîne l'autre : $\forall x \in E \ , \ x \ \mathcal{R} \ y \Rightarrow x \ \mathcal{R}' \ y$. On dit alors que $\mathcal{R}'$ est *plus faible* que $\mathcal{R}$, car elle contient moins d'information ; on dit également que $\mathcal{R}$ est *plus forte* que $\mathcal{R}'$. Par exemple la congruence modulo 2 dans $\mathbb{Z}$ est plus faible que la congruence modulo 4. Dire que $\mathcal{R}'$ est plus faible que $\mathcal{R}$ est équivalent à dire que $\Gamma_{\mathcal{R}} \subset \Gamma_{\mathcal{R}'}$. On pourra donc également dire que $\mathcal{R}'$ *contient* $\mathcal{R}$, ou que $\mathcal{R}$ *est contenue dans* $\mathcal{R}'$. La plus faible de toutes les relations binaires est celle dont le graphe est le plus grand, donc celle telle que $\Gamma_{\mathcal{R}} = E \times E$, autrement dit, $x \ \mathcal{R} \ y$ est vérifié quels que soient x et y ; ainsi, affirmer que $x \ \mathcal{R} \ y$ n'apporte aucune information sur x et y. La plus forte est celle dont le graphe est vide, autrement dit, $x \ \mathcal{R} \ y$ n'est jamais vérifié.

Si l'on se donne une famille $(\mathcal{R}_i)_{i \in I}$ de relations binaires sur un ensemble E, on peut définir une relation en prenant pour son graphe $\bigcap\limits_{i \in I} \Gamma_{\mathcal{R}_i}$. Cette relation, appelée *intersection des* $\mathcal{R}_i$ et notée $\bigcap\limits_{i \in I} \mathcal{R}_i$ est la plus faible des relations plus fortes que toutes les $\mathcal{R}_i$. Elle est définie en termes logiques par l'équivalence :

$$x \ \mathcal{R} \ y \Longleftrightarrow \forall i \in I \ , \ x \ \mathcal{R}_i \ y.$$

Par exemple, dans $\mathbb{Z}$, l'intersection de toutes les relations $x \equiv y \pmod{n}$ (pour $n \in \mathbb{N}^*$) est l'égalité : en effet, le seul entier $x - y$ divisible par tous les entiers naturels non nuls n est 0. Une intersection de relations réflexives (resp. transitives, resp. symétriques, resp. antisymétriques, resp. d'équivalence, resp. d'ordre) est elle-même réflexive (resp. transitive, resp. symétrique, resp. antisymétrique, resp. d'équivalence, resp. d'ordre). Par ailleurs, comme on l'a vu, toute relation est contenue dans une relation qui est à la fois réflexive, transitive et symétrique (c'est la relation dont le graphe est $E \times E$). On peut de même définir une relation de graphe $\bigcup\limits_{i \in I} \Gamma_{\mathcal{R}_i}$. Cette relation, appelée *réunion des* $\mathcal{R}_i$ et notée $\bigcup\limits_{i \in I} \mathcal{R}_i$ est moins utile.

Définition 17. L'intersection de toutes les relations transitives (resp. réflexives transitives) qui contiennent une relation $\mathcal{R}$ donnée est appelée *clôture transitive de* $\mathcal{R}$ (resp. *clôture réflexive transitive de* $\mathcal{R}$). L'intersection de toutes les relations d'équivalence qui contiennent une relation $\mathcal{R}$ donnée est appelée *relation d'équivalence engendrée par la relation* $\mathcal{R}$.

Exercice 10.
Soit $E = \{a, b, c, d\}$. Soit $\mathcal{R}$ la relation dont le graphe est $G = \{(a, b), (b, c)\}$.
Déterminer sa clôture transitive et sa clôture réflexive transitive.
Solution. Elles ont pour graphes respectifs :

$$\{(a, b), (b, c), (a, c)\} \quad \text{et} \quad \{(a, b), (b, c), (a, c), (a, a), (b, b), (c, c), (d, d)\}.$$

La relation d'équivalence engendrée par $\mathcal{R}$ a pour graphe :

$$\{(a, b), (b, a), (b, c), (c, b), (a, c), (c, a), (a, a), (b, b), (c, c), (d, d)\}.$$

On ne peut parler de relation d'ordre engendrée par une relation $\mathcal{R}$, car $\mathcal{R}$ n'est peut-être contenue dans aucune relation d'ordre. Considérons par exemple, sur l'ensemble à trois éléments distincts $E = \{a, b, c\}$, la relation de graphe $\{(a, b), (b, c), (c, a)\}$. Toute relation d'ordre $\mathcal{R}'$ qui la contient vérifierait $a \mathcal{R}' b$ et $b \mathcal{R}' c$, donc $a \mathcal{R}' c$ (transitivité) ; et $c \mathcal{R}' a$, donc $a = c$ (antisymétrie), ce qui contredit l'hypothèse.

La notion de clôture transitive est importante en mathématiques discrètes et en informatique théorique à cause de son interprétation en termes de « graphe », ici entendu dans le sens de « représentation par sommets et arêtes ». Reprenons l'exemple de $E = \{a, b, c, d\}$ et de la relation $\mathcal{R}$ dont le graphe est $G = \{(a, b), (b, c)\}$. Nous interprétons a, b, c, d comme des points, les sommets du graphe ; et $(a, b), (b, c)$ comme des arêtes reliant a à b et b à c (mais pas dans le sens contraire : il s'agit d'arêtes *orientées*). Le dessin de gauche montre le graphe correspondant. Si l'on suit le chemin formé des arêtes (a, b) et (b, c) on constate que l'on peut aller de a à c, on dit que c est accessible à partir de a. Le dessin de droite est le graphe qui correspond à la clôture transitive de $\mathcal{R}$.

$$\begin{tikzcd} a \arrow[d] & d \\ b \arrow[r] & c \end{tikzcd} \qquad\qquad \begin{tikzcd} a \arrow[d] \arrow[dr] & d \\ b \arrow[r] & c \end{tikzcd}$$

5.2 Relations d'équivalence

Rappelons qu'une relation d'équivalence sur un ensemble E est une relation binaire réflexive, symétrique et transitive.

Exemples.

1. Sur tout ensemble E, l'égalité est une relation d'équivalence, et c'est la plus forte (elle est contenue dans toutes les autres). La relation d'équivalence la plus faible est celle dont le graphe est $E \times E$, c'est-à-dire celle qui est satisfaite par tous les couples $(x, y) \in E \times E$.
2. La relation $x \equiv y \pmod{a}$ est une relation d'équivalence dans $\mathbb{Z}$ ($a \in \mathbb{Z}$).
   La relation $x \equiv y \pmod{2\pi}$ est une relation d'équivalence dans $\mathbb{R}$.
3. Plus généralement, soit G un groupe commutatif (module II.2) et soit H un sous-groupe de G. La relation de congruence modulo H est une relation d'équivalence. Cela s'applique en particulier au cas où G est un anneau et H un idéal (module II.2) et au cas où G est un espace vectoriel et H un sous–espace vectoriel (module II.3).
4. Dans l'ensemble $\mathcal{F}(I, \mathbb{R})$ des fonctions numériques sur un intervalle ouvert I de $\mathbb{R}$, tel que $0 \in I$, on définit la relation $f \mathcal{R} g$ (pour $n \in \mathbb{N}^*$ fixé) par la formule : $f = g + \mathrm{o}(x^n)$ (module IV.8), c'est une relation d'équivalence. De même, si l'on écrit $f \sim g$ pour $f = g + \mathrm{o}(g)$, on obtient une autre relation d'équivalence.
5. Dans l'ensemble $\mathbb{N}\times\mathbb{N}$, on pose $(a, b) \sim (c, d)$ si, et seulement si, $a+d = b+c$. C'est une relation d'équivalence, qui servira, au module II.1, à *construire* l'ensemble $\mathbb{Z}$. De même, la relation d'équivalence définie par $(a, b) \sim (c, d)$ si, et seulement si, $ad = bc$ dans l'ensemble $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$, servira, au module II.2, à *construire* l'ensemble $\mathbb{Q}$.
6. En géométrie euclidienne plane, on dit improprement que deux triangles ABC et $A'B'C'$ sont « égaux » si l'on peut déplacer l'un pour le faire coïncider avec l'autre. On obtient ainsi une relation d'équivalence sur l'ensemble des triangles.

Définition 18. Soit $\mathcal{R}$ une relation d'équivalence sur l'ensemble E. On appelle *classe d'équivalence de* $x \in E$ et l'on note $\mathrm{Cl}(x)$ l'ensemble des éléments de E équivalents à x :

$$\mathrm{Cl}(x) := \{y \in E \mid x \mathcal{R} y\}.$$

La relation $\mathcal{R}$ étant réflexive, $x \in \mathrm{Cl}(x)$.

Proposition 3.
(i) Deux éléments de E sont équivalents si, et seulement s'ils ont la même classe :

$$\forall x, y \in E \ , \ x \mathcal{R} y \Longleftrightarrow \mathrm{Cl}(x) = \mathrm{Cl}(y).$$

(ii) Les classes d'équivalence forment une partition de E. Autrement dit : elles sont non vides, leur réunion est E et elles sont deux à deux disjointes.

Démonstration. Supposons $x \mathcal{R} y$. Soit $z \in \mathrm{Cl}(x)$. Alors $x \mathcal{R} z$ (par définition de $\mathrm{Cl}(x)$), donc $y \mathcal{R} z$ (par symétrie et transitivité de $\mathcal{R}$). On a ainsi démontré que $\mathrm{Cl}(x) \subset \mathrm{Cl}(y)$. L'inclusion réciproque se démontre de la même manière, et l'on a bien $\mathrm{Cl}(x) = \mathrm{Cl}(y)$. Supposons réciproquement que $\mathrm{Cl}(x) = \mathrm{Cl}(y)$. Comme on l'a remarqué, $y \in \mathrm{Cl}(y)$, donc $y \in \mathrm{Cl}(x)$ (par hypothèse), donc $x \mathcal{R} y$ par définition de $\mathrm{Cl}(x)$. Cela achève la preuve de l'équivalence (i).
Pour tout x de E, on a $x \in \mathrm{Cl}(x)$ (qui est donc non vide). La réunion des classes d'équivalence contient donc tous les éléments de E, c'est donc bien E. Soient $\mathrm{Cl}(x)$ et $\mathrm{Cl}(y)$ deux classes d'équivalence. Si elles ne sont pas disjointes, il existe $z \in \mathrm{Cl}(x) \cap \mathrm{Cl}(y)$, donc tel que $x \mathcal{R} z$ et $y \mathcal{R} z$ $\big($par définition de $\mathrm{Cl}(x)$ et $\mathrm{Cl}(y)\big)$. Par symétrie et transitivité de $\mathcal{R}$, on a $x \mathcal{R} y$, donc $\mathrm{Cl}(x) = \mathrm{Cl}(y)$, ce qui achève la preuve de (ii). ■

Réciproquement, soit $(E_i)_{i \in I}$ une partition de E. En posant que $x \mathcal{R} y$ lorsque x et y appartiennent au même sous-ensemble E_i, on définit une relation d'équivalence pour laquelle les classes d'équivalence sont les E_i.

Exemple. Si $f : E \to F$ est une application, la relation définie par : $x \mathcal{R} y$ si, et seulement si, $f(x) = f(y)$ est une relation d'équivalence. La classe de $x \in E$ est $\mathrm{Cl}(x) = f^{-1}\big(f(x)\big)$. Réciproquement, toute relation d'équivalence $\mathcal{R}$ peut s'obtenir de cette manière : on définit $f : E \to \mathcal{P}(E)$ en posant $f(x) = \mathrm{Cl}(x)$, et l'on a bien $x \mathcal{R} y$ si, et seulement si, $f(x) = f(y)$.

Définition 19. Soit $\mathcal{R}$ une relation d'équivalence sur l'ensemble E. On appelle *ensemble de représentants* pour la relation $\mathcal{R}$ une partie A de E dont l'intersection avec chaque classe d'équivalente contient un unique élément. De manière équivalente, chaque élément de E est équivalent à un unique élément de A :

$$\forall x \in E \ , \ \exists! y \in A \ : \ x \mathcal{R} y.$$

Un ensemble de représentants s'obtient donc en choisissant dans chaque classe un élément particulier, que l'on considérera comme le *représentant* de la classe. Les exemples suivants montrent que cette notion d'aspect barbare est naturelle et utile.

Exemples.

1. L'ensemble $\{0, 1, 2, 3, 4, 5, 6, 7, 8\}$ est un ensemble de représentants pour la relation de congruence modulo 9 dans $\mathbb{Z}$ (module II.1).
2. Les intervalles $[0, 2\pi[$ et $]-\pi, \pi]$ sont des ensembles de représentants pour la relation de congruence modulo 2π dans $\mathbb{R}$.
3. Sur l'ensemble $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$, la relation d'équivalence $(a, b) \sim (c, d)$ si, et seulement si, $ad = bc$ (qui sert à construire $\mathbb{Q}$, module II.2) admet pour ensemble de représentants l'ensemble formé de $(0, 1)$ et des couples $(p, q) \in (\mathbb{Z} \setminus \{0\}) \times \mathbb{N}^*$ tels que p et q sont premiers entre eux (« forme irréductible d'une fraction »).

5.3 Relations d'ordre

Rappelons qu'une relation d'ordre sur un ensemble E est une relation réflexive, transitive et antisymétrique. On dit aussi que l'on a un *ordre* sur E. La plupart des relations d'ordre sont notées $\leqslant$ ou $\preceq$ (exceptions notables : l'inclusion et la divisibilité). On dit alors que $(E, \leqslant)$ (ou, par abus, E lui-même) est un *ensemble ordonné*. Lorsque $x \leqslant y$, on dit que y *majore* x, ou que x *minore* y. Les éléments x et y sont alors *comparables*. L'ordre est dit *total* si deux éléments $x, y \in E$ quelconques sont comparables.

Il est parfois commode de raisonner sur l'*ordre strict* associé à $\leqslant$, que nous noterons alors $<$. On le définit en posant $x < y$ si, et seulement si, $(x \leqslant y$ et $x \neq y)$. On dit que y *majore strictement* x, ou que x *minore strictement* y. Cette relation est transitive et antiréflexive (c'est-à-dire que l'on n'a jamais $x < x$). Réciproquement, si la relation $<$ est transitive et antiréflexive, on définit une relation d'ordre en posant $x \leqslant y$ si, et seulement si, $(x < y$ ou $x = y)$ (c'est un exercice amusant, laissé au lecteur).

Exemples.

1. La relation $\leqslant$ sur l'ensemble $\mathbb{R}$ est une relation d'ordre total. La relation induite sur $\mathbb{N}$, $\mathbb{Z}$ et $\mathbb{Q}$ l'est donc également. L'ordre strict associé est la relation d'inégalité stricte $<$.
2. La relation $\leqslant$ sur l'ensemble $\mathcal{F}(\mathbb{R}, \mathbb{R})$ des fonctions numériques sur $\mathbb{R}$ est une relation d'ordre. Ce n'est pas un ordre total : par exemple les fonctions $x \mapsto x$ et $x \mapsto 0$ ne sont pas comparables.

   Ici, $f < g$ signifie : $\forall x \in \mathbb{R}\ ,\ f(x) \leqslant g(x)$ et $\exists x \in \mathbb{R}\ :\ f(x) < g(x)$. Il ne faut donc pas commettre l'erreur d'en déduire que $\forall x \in \mathbb{R}\ ,\ f(x) < g(x)$.
3. La relation $\subset$ sur l'ensemble $\mathcal{P}(\mathbb{N})$ est une relation d'ordre. Ce n'est pas un ordre total car $\{0\}$ et $\{1\}$, par exemple, ne sont pas comparables (aucun des deux n'est inclus dans l'autre).
4. La relation $|$ (divisibilité) sur l'ensemble $\mathbb{N}$ est une relation d'ordre. Ce n'est pas un ordre total car 2 et 3, par exemple, ne sont pas comparables (aucun des deux ne divise l'autre).
5. La relation $|$ (divisibilité) sur l'ensemble $\mathbb{Z}$ n'est pas une relation d'ordre car elle n'est pas antisymétrique : les éléments a et $-a$ se divisent mutuellement.

6. Soient $(E_1, \leqslant), \ldots, (E_n, \leqslant)$ des ensembles ordonnés.
   L'*ordre produit* sur $E := E_1 \times \cdots \times E_n$ est défini par : $(x_1, \ldots, x_n) \preceq (y_1, \ldots, y_n)$ si, et seulement si, $x_1 \leqslant y_1, \ldots, x_n \leqslant y_n$. Ce n'est en général pas un ordre total. Par exemple, sur $\mathbb{R} \times \mathbb{R}$, $(0,1)$ et $(1,0)$ ne sont pas comparables pour l'ordre produit.
   L'*ordre lexicographique* sur E est défini par : $(x_1, \ldots, x_n) \preceq (y_1, \ldots, y_n)$ si, et seulement si, $(x_1, \ldots, x_n) = (y_1, \ldots, y_n)$ ou bien, i étant le plus petit indice tel que $x_i \neq y_i$, on a $x_i < y_i$. On démontre (*cf.* l'exercice I.1.23 de la page 59) que, si les E_i sont totalement ordonnés, c'est un ordre total. Par exemple, sur $\mathbb{R} \times \mathbb{R}$ muni de l'ordre lexicographique, $(0,1) \preceq (1,0)$. Si $A = \{a, b, c\}$ est muni de l'ordre alphabétique (tel que $a \leqslant b \leqslant c$), l'ordre lexicographique sur $A \times A$ est donné par les relations :
   $$(a,a) \leqslant (a,b) \leqslant (a,c) \leqslant (b,a) \leqslant (b,b) \leqslant (b,c) \leqslant (c,a) \leqslant (c,b) \leqslant (c,c).$$

Soit $(E, \leqslant)$ un ensemble ordonné. On dit qu'un élément de E est *maximal* s'il n'est strictement majoré par aucun élément de E, et qu'il est *minimal* s'il n'est strictement minoré par aucun élément de E ; autrement dit :

$$x \text{ maximal} \iff \big(\forall y \in E\ ,\ x \leqslant y \Rightarrow x = y\big),$$
$$x \text{ minimal} \iff \big(\forall y \in E\ ,\ y \leqslant x \Rightarrow x = y\big).$$

On dit qu'un élément de E est *maximum*, ou que c'est le *plus grand élément*, s'il majore tous les éléments de E, et qu'il est *minimum*, ou que c'est le *plus petit élément*, s'il minore tous les éléments de E ; autrement dit :

$$x \text{ maximum} \iff \big(\forall y \in E\ ,\ y \leqslant x\big),$$
$$x \text{ minimum} \iff \big(\forall y \in E\ ,\ x \leqslant y\big).$$

S'il y a un maximum, il est unique et c'est l'unique élément maximal ; on le note alors $\max E$. De même, s'il y a un minimum, il est unique et c'est l'unique élément minimal ; on le note alors $\min E$. De plus, dans un ensemble totalement ordonné, maximal entraîne maximum (et minimal entraîne minimum). Cependant, dans un ensemble non totalement ordonné, ce n'est pas le cas. Par exemple, dans l'ensemble des parties non vides de $\{0,1\}$, ordonné par l'inclusion, les parties $\{0\}$ et $\{1\}$ sont toutes deux minimales et il n'y a pas de minimum.

Soit maintenant $A \subset E$ une partie de E. On dit qu'un élément $M \in E$ est un *majorant* de A (ou qu'il *majore* A) s'il majore tous les éléments de A : $\forall a \in A\ ,\ a \leqslant M$. On dit de même qu'un élément $m \in E$ est un *minorant* de A (ou qu'il *minore* A) s'il *minore* tous les éléments de A : $\forall a \in A\ ,\ m \leqslant a$. On dit encore que la partie A est *majorée* ou *minorée*. Une partie majorée et minorée est dite bornée. Ces définitions s'étendent directement à des familles d'éléments de E : par exemple, une suite $(u_n)_{n \in \mathbb{N}}$ de réels est majorée s'il existe un réel M qui majore tous les u_n.

Définition 20. Soient E un ensemble ordonné et $A \subset E$. Si l'ensemble des majorants de A admet un plus petit élément, cet élément est appelé *borne supérieure de* A et noté $\sup A$. Si l'ensemble des minorants de A admet un plus grand élément, cet élément est appelé *borne inférieure de* A et noté $\inf A$.

Exemples.

1. Tout ensemble ordonné fini non vide admet au moins un élément maximal et un élément minimal.
2. Les ensembles $\mathbb{Z}$, $\mathbb{Q}$ et $\mathbb{R}$ n'admettent ni élément maximal ni élément minimal. L'ensemble $\mathbb{N}$ n'admet pas de maximal, mais il admet un minimum : 0.
3. Dans l'ensemble totalement ordonné $\mathbb{Q}$, le sous-ensemble des x tels que $x^2 < 2$ n'admet ni maximum ni borne supérieure. Dans l'ensemble totalement ordonné $\mathbb{R}$, le sous-ensemble des x tels que $x^2 < 2$ n'admet pas de maximum, mais il admet $\sqrt{2}$ pour borne supérieure.
4. L'ensemble ordonné $(\mathcal{P}(E), \subset)$ admet $\varnothing$ pour minimum et E pour maximum. Si $X, Y \in \mathcal{P}(E)$, les majorants de $\{X, Y\}$ sont les sous-ensembles qui contiennent à la fois X et Y et $\{X, Y\}$ admet pour borne supérieure $X \cup Y$. De même, les minorants de $\{X, Y\}$ sont les sous-ensembles qui sont contenus à la fois dans X et Y et $\{X, Y\}$ admet pour borne inférieure $X \cap Y$.
5. L'ensemble ordonné $(\mathbb{N}, |)$ admet 0 pour maximum et 1 pour minimum. Si $a, b \in \mathbb{N}$, les majorants de $\{a, b\}$ sont les multiples communs à a et b, et $\{a, b\}$ admet pour borne supérieure le ppcm de a et b. De même, les minorants de $\{a, b\}$ sont les diviseurs communs à a et b, et $\{a, b\}$ admet pour borne inférieure le pgcd de a et b (module II.1).

Soient E et F deux ensembles ordonnés. L'application $f : E \to F$ est dite *croissante* si $x \leqslant y$ entraîne $f(x) \leqslant f(y)$ et *décroissante* si $x \leqslant y$ entraîne $f(y) \leqslant f(x)$. Une application est dite *monotone* si elle est soit croissante soit décroissante. On dit que f est *strictement croissante* si $x < y$ entraîne $f(x) < f(y)$; on définit de même une application *strictement décroissante* ou *strictement monotone*. Par exemple, toute application monotone injective est strictement monotone.

Si l'ensemble de départ est $E = \mathbb{N}$, l'application f est une suite et l'on parle de *suite croissante, décroissante, monotone, strictement croissante, strictement décroissante, strictement monotone*. Dans ce cas, la propriété de monotonie se vérifie plus simplement en comparant des termes consécutifs de la suite. Par exemple, une suite $(u_n)_{n\in\mathbb{N}}$ est strictement croissante si, et seulement si, pour tout entier n, $u_n < u_{n+1}$.

On vérifie sans peine que la composée de deux applications croissantes est croissante ; que la composée (dans n'importe quel ordre) d'une application croissante et d'une application décroissante est décroissante ; enfin, que la composée de deux applications décroissantes est croissante.

Cependant, il n'est pas toujours vrai que la réciproque d'une application croissante bijective est croissante. Par exemple, l'application identité de $(\mathbb{N}^*, |)$ dans $(\mathbb{N}^*, \leqslant)$ est croissante, parce que, pour des entiers naturels non nuls, $a \mid b \Rightarrow a \leqslant b$. Mais l'application réciproque, qui est l'application identité de $(\mathbb{N}^*, \leqslant)$ dans $(\mathbb{N}^*, |)$ n'est pas croissante, parce que la réciproque de l'implication ci-dessus est, bien sûr, fausse. Cette situation tient à ce que l'un des deux ordres est plus faible que l'autre.

On appellera *isomorphisme* entre deux ensembles ordonnés une application bijective qui est croissante et dont la réciproque est croissante. Dans ce cas, $x \leqslant y$ est logiquement équivalent à $f(x) \leqslant f(y)$. C'est cette relation qui exprime au mieux la similarité de deux relations d'ordre.

Soient E et F deux ensembles ordonnés et $f : E \to F$ une bijection croissante. Si E est totalement ordonné, f est un isomorphisme (vous êtes évidemment encouragés à le démontrer). Sinon, ce n'est pas nécessairement le cas. Par exemple, si l'on prend pour E l'ensemble $\mathbb{N}$ muni de la relation d'ordre $|$ (divise) et pour F l'ensemble $\mathbb{N}$ muni de la relation d'ordre usuelle $\leqslant$, l'application identité de $\mathbb{N}$ est une bijection croissante de E dans F mais pas de F dans E, ce n'est donc pas un isomorphisme.

5.3.1 Propriétés algébriques

Supposons E muni d'une relation d'ordre $\leqslant$ et d'une loi de composition interne $\star$. On peut se poser la question de la compatibilité entre les deux structures. Nous n'en donnerons pas une définition totalement générale, à cause de l'exemple des lois d'addition et de multiplication de $\mathbb{R}$. En effet, si l'on a l'égalité $a \leqslant b$ dans $\mathbb{R}$, on a, quel que soit c : $a+c \leqslant b+c$, mais pas nécessairement $ac \leqslant bc$. La compatibilité ne s'exprime pas de la même manière dans les deux cas. Anticipant les définitions du module II.2, nous distinguerons le cas d'un groupe commutatif, celui d'un anneau et celui d'un corps. Supposons d'abord que E est un groupe commutatif noté additivement. Nous dirons que *la relation d'ordre $\leqslant$ et la loi $+$ sont compatibles*, ou encore que $(E, +, \leqslant)$ est un *groupe commutatif ordonné*, si :

$$\forall a, b, c \in E \ , \ a \leqslant b \Longrightarrow a + c \leqslant b + c.$$

Avec les raisonnements habituels dans $(\mathbb{R}, +)$, on en déduit, entre autres, les règles de calcul suivantes, dans lesquelles $a, b, c, \ldots \in E$ sont quelconques :

$$a \leqslant b \Longleftrightarrow a + c \leqslant b + c,$$

$$a \leqslant b \Longleftrightarrow c + a \leqslant c + b,$$

$$a \leqslant b \Longleftrightarrow 0 \leqslant b - a,$$

$$a \leqslant b \Longleftrightarrow a - b \leqslant 0,$$

$$a \leqslant b \Longleftrightarrow -b \leqslant -a,$$

$$a \leqslant b \text{ et } c \leqslant d \Longrightarrow a + c \leqslant b + d.$$

Si l'on note E_+ l'ensemble des éléments positifs ou nuls de E :

$$E_+ := \{x \in E \mid 0 \leqslant x\},$$

on voit que E_+ vérifie les propriétés suivantes :

$$E_+ + E_+ \subset E_+,$$

$$E_+ \cap (-E_+) = \{0\}.$$

Réciproquement, si, dans le groupe commutatif $(E, +)$, il y a un sous-ensemble E_+ vérifiant ces propriétés, en posant $x \leqslant y$ si, et seulement si, $y - x \in E_+$, on fait de E un groupe commutatif ordonné. Une condition nécessaire et suffisante pour que l'ordre soit total est la suivante :

$$E_+ \cup (-E_+) = E.$$

Considérons maintenant le cas d'un anneau $(E, +, .)$. Supposons que l'on a déjà fait de $(E, +)$ un groupe ordonné par la relation $\leqslant$. Nous dirons que *la relation d'ordre $\leqslant$ et la structure d'anneau*

sont compatibles, ou encore que $(E, +, ., \leqslant)$ est un *anneau ordonné*, si :

$$\forall a, b \in E \ , \ \big(0 \leqslant a \ \text{ et } \ 0 \leqslant b\big) \Longrightarrow 0 \leqslant a.b.$$

Si la relation d'ordre a été définie à l'aide de E_+, cela équivaut à la condition :

$$E_+.E_+ \quad \subset \quad E_+.$$

On a alors les règles usuelles de calcul :

$$\begin{aligned} (a \leqslant b \ \text{ et } \ 0 \leqslant c) \quad &\Longrightarrow \quad (a.c \leqslant b.c \ \text{ et } \ c.a \leqslant c.b), \\ (a \leqslant b \ \text{ et } \ c \leqslant 0) \quad &\Longrightarrow \quad (b.c \leqslant a.c \ \text{ et } \ c.b \leqslant c.a). \end{aligned}$$

Ainsi, $\mathbb{Z}$ est un anneau totalement ordonné. Nous réserverons enfin le nom de *corps ordonné* au cas d'un corps qui est un anneau totalement ordonné. Ainsi, $\mathbb{Q}$ et $\mathbb{R}$ sont des corps ordonnés, mais on verra au module II.5 qu'il est impossible de faire de $\mathbb{C}$ un corps ordonné.

6 Cardinaux

6.1 Induction

6.1.1 Induction noetherienne

On dit qu'une suite *stationne en* $n_0 \in \mathbb{N}$, ou encore qu'elle est *stationnaire* si elle est constante à partir du rang n_0, c'est-à-dire si : $\forall n \geqslant n_0 \ , \ u_n = u_{n_0}$.

Proposition et définition 4. Soit E un ensemble ordonné. Les propriétés suivantes sont équivalentes :

(i) Toute suite croissante de E est stationnaire.

(ii) Il n'existe pas dans E de suite strictement croissante.

(iii) Toute partie non vide de E admet un élément maximal.

On dit alors que l'ensemble ordonné E est *noetherien*.

Démonstration. Supposons vraie la propriété (i). S'il existait dans E une suite strictement croissante, elle serait croissante et non stationnaire, contredisant (i). Cela montre que (i) implique (ii). Supposons vraie la propriété (ii). Nous allons démontrer (iii) par l'absurde. Soit $A \subset E$ une partie non vide n'admettant pas d'élément maximal. Nous allons construire à partir de ces hypothèses une suite strictement croissante dans A, donc dans E. Puisque A est non vide, on peut choisir $u_0 \in A$. Puisque A n'admet pas d'élément maximal, u_0 est strictement majoré dans A, et l'on peut choisir $u_1 \in A$ tel que $u_0 < u_1$. En itérant ce procédé, on construit une suite strictement croissante, contredisant (ii). Cela montre que (ii) implique (iii). Supposons vraie la propriété (iii). Soit $(u_n)_{n\in\mathbb{N}}$ une suite croissante dans E. L'image $\{u_n \mid n \in \mathbb{N}\}$ de cette suite est une partie non vide de E, qui admet donc, par hypothèse, un élément maximal u_{n_0}. Comme la suite est croissante, elle stationne en n_0. ∎

Il existe un « principe d'induction noetherien » qui généralise le raisonnement par récurrence dans $\mathbb{N}$, mais il repose plutôt sur l'existence d'éléments minimaux. Nous appellerons donc *artinien* un ensemble ordonné dans lequel toute partie non vide admet un élément minimal ; de manière équivalente, toute suite décroissante est stationnaire ; ou encore, il n'existe pas de suite strictement décroissante. Par exemple, tout ensemble ordonné fini est à la fois artinien et noetherien.

Théorème 5 (Principe d'induction noetherienne). Soit E un ensemble artinien. Soit $P(x)$ une propriété définie sur E et vérifiant la condition d'*hérédité* :

$$\forall x \in E\ ,\ \big(\forall y < x\ ,\ P(y)\big) \Longrightarrow P(x).$$

Alors P est vraie sur E tout entier : $\forall x \in E\ ,\ P(x)$.

Démonstration. Soit A l'ensemble des $x \in E$ qui ne vérifient pas $P(x)$. Supposons A non vide. Il admet alors un élément minimal x. Ainsi, pour tout $y < x$, $y \notin A$, autrement dit, $P(y)$. D'après la condition d'hérédité, on a donc $P(x)$, ce qui n'est pas possible puisque $x \in A$. L'ensemble A est donc vide, ce qui est la conclusion voulue. ∎

Exemple. L'ensemble des polynômes unitaires sur un corps K, ordonné par la relation de divisibilité, est artinien. En effet, si $P \mid Q$, alors $\deg P \leqslant \deg Q$ et l'on sait qu'il n'existe pas de suite strictement décroissante d'entiers naturels. Démontrons par induction noetherienne que tout polynôme unitaire est produit de polynômes irréductibles ; on sous-entend ici que le produit vide vaut 1. Il suffit de démontrer l'hérédité de la propriété. Soit donc P un polynôme unitaire tel que tout diviseur strict de P est produit d'irréductibles. Si P est irréductible, on n'a rien à démontrer. Sinon, $P = QR$ où Q et R sont unitaires et divisent strictement P. Par hypothèse d'induction noetherienne, ils sont produits d'irréductibles, donc $P = QR$ l'est aussi. La propriété est donc héréditaire, donc vraie pour tout polynôme unitaire.

6.1.2 Ensembles bien ordonnés

Les ensembles noetheriens ou artiniens généraux (non totalement ordonnés) interviendront surtout en algèbre commutative.

Définition 21. On dit que l'ensemble E est *bien ordonné* si toute partie non vide de E admet un plus petit élément. On dit aussi que son ordre est un *bon ordre*.

L'exemple le plus célèbre d'ensemble bien ordonné est $(\mathbb{N}, \leqslant)$; il sera étudié en détail au module II.1. Comme un plus petit élément est minimal, tout ensemble bien ordonné est artinien. En prenant une paire $\{a, b\}$, on voit que l'ordre est de plus total (si a est le plus petit, $a \leqslant b$, si b est le plus petit, $b \leqslant a$). Réciproquement, tout ordre total artinien est un bon ordre : en effet, toute partie non vide admet un élément minimal, qui est nécessairement minimum (puisque l'ordre est total). Un théorème dû à Zermelo dit que *tout ensemble peut être muni d'un bon ordre*. C'est immédiat pour un ensemble fini : il suffit de le munir de n'importe quel ordre total, ce qui ne présente aucune difficulté ! Pour un ensemble dénombrable (voir plus loin), c'est encore possible : on utilise n'importe quelle bijection $f : E \to \mathbb{N}$ et l'on pose $x \leqslant y$ si, et seulement si, $f(x) \leqslant f(y)$. Cependant, il faut prendre garde que les ensembles totalement ordonnés dénombrables $\mathbb{Z}$ et $\mathbb{Q}$ ne sont pas bien ordonnés. On trouvera à l'exercice I.1.24 de la page 60 un exemple d'ensemble dénombrable artinien mais pas bien ordonné. Pour apprécier à sa juste valeur le théorème de Zermelo, le lecteur est invité à chercher un bon ordre sur $\mathbb{R}$ ou sur $\mathcal{P}(\mathbb{N})$.

6.1.3 Le lemme de Zorn

Pour motiver la définition suivante, considérons un ensemble M muni d'une loi $\star$, et soit $(X_i)_{i \in I}$ une famille de parties stables de M. La réunion X des X_i n'a aucune raison d'être stable. En effet, si $x, y \in X$ alors il existe des indices i et j tels que $x \in X_i$ et $y \in X_j$, mais, à moins d'avoir $i = j$, on ne peut rien en déduire. Cependant, si l'on suppose que deux quelconques des parties X_i sont comparables pour l'inclusion, on aura soit $X_i \subset X_j$ soit $X_j \subset X_i$; dans le premier cas, $x, y \in X_j \Rightarrow x \star y \in X_j \Rightarrow x \star y \in X$, et similairement dans le second cas.

Définition 22.
(i) On appelle *chaîne* d'un ensemble ordonné E une famille non vide d'éléments de E qui sont deux à deux comparables.
(ii) On dit que E est *inductif* si toute chaîne de E est majorée.

Ainsi, la réunion d'une chaîne de parties stables de E est stable. L'ensemble des parties stables, ordonné par inclusion, est donc inductif.

Théorème 6 (Lemme de Zorn). Tout ensemble inductif admet un élément maximal. Plus précisément, dans un ensemble inductif, tout élément est majoré par un élément maximal.

Nous ne démontrerons pas ce théorème. Il est essentiellement équivalent au théorème de Zermelo, et à l'axiome du choix. Nous appliquerons le lemme de Zorn dès la section suivante, puis au module II.3. Pour illustrer le théorème sur l'exemple des parties stables de M, il faut raffiner celui-ci : en effet, il est trivial que M lui-même est stable, et le lemme de Zorn ne nous apporte rien. Mais on peut également fixer un élément arbitraire $a \in M$, et considérer l'ensemble des parties stables qui ne contiennent pas a. On vérifie que cet ensemble est inductif (si les X_i ne contiennent pas a, leur réunion non plus). On en conclut qu'il existe une partie stable maximale ne contenant pas a, ce qui n'est nullement évident !

6.2 Équipotence

Nous dirons que les ensembles E et F sont *équipotents* s'il existe une bijection de E sur F. Tout ensemble est équipotent à lui-même (en prenant pour bijection l'identité) ; si E est équipotent à F et F à G, alors E est équipotent à G (en prenant la composée des deux bijections) ; enfin, si E est équipotent à F, alors F est équipotent à E (en prenant la réciproque de la bijection). L'équipotence est donc une relation d'équivalence. Pour des ensembles finis, l'équipotence est la relation « avoir le même nombre d'éléments ». Ce qu'il y a de commun entre tous les ensembles équipotents à un ensemble fini donné est donc un nombre entier (voir le module II.1). Pour des ensembles quelconques, la notion générale est celle de cardinal, mais sa définition requiert un axiome.

Axiome et définition 14. On peut associer à chaque ensemble E un objet appelé *cardinal de* E et noté $\operatorname{card} E$, qui satisfait à la règle suivante : deux ensembles E et F ont même cardinal si, et seulement s'ils sont équipotents.

On note $0 := \operatorname{card} \varnothing$. Les ensembles équipotents à un singleton donné (resp. à une paire donnée) sont les singletons (resp. les paires), et l'on note $1 := \operatorname{card}\{a\}$ quel que soit a et $2 := \operatorname{card}\{a, b\}$ quels que soient a et b distincts, etc. Dans certaines présentations de la théorie des ensembles, on appelle « puissance » ce que nous appelons « cardinal » et l'on appelle « cardinal de E » un ensemble choisi dans la classe de E. Voici les exemples les plus simples. Le seul ensemble équipotent à $\varnothing$ est $\varnothing$, donc $0 := \varnothing$. Ensuite $1 := \{\varnothing\}$, $2 := \{\varnothing, \{\varnothing\}\}$, puis, en général, $n + 1 := n \cup \{n\}$: c'est la construction de Von Neumann des entiers (section 6.3).

Définition 23. On dit que $\operatorname{card} E \leqslant \operatorname{card} F$ (lire : « inférieur ou égal ») s'il existe une application injective de E dans F.

Pour vérifier que cette définition a bien un sens (c'est-à-dire qu'elle ne dépend que de $\operatorname{card} E$ et $\operatorname{card} F$ et non de E et F), il faut toutefois vérifier que si E, E' sont équipotents et si F, F' le sont également, l'existence d'une injection $E \to F$ équivaut à celle d'une injection $E' \to F'$. On procède comme suit : soient $f : E \to E'$ et $g : F \to F'$ des bijections. Si $u : E \to F$ est une

injection, alors $u' := g \circ u \circ f^{-1} : E' \to F'$ est une injection ; réciproquement, si $u' : E' \to F'$ est une injection, alors $u := g^{-1} \circ u' \circ f : E \to F$ est une injection. Pour la plupart des constructions sur les cardinaux, de telles précautions seront nécessaires, mais nous en laisserons en général les détails au lecteur.

Proposition 7. Si $E \neq \varnothing$, la relation $\operatorname{card} E \leqslant \operatorname{card} F$ équivaut à l'existence d'une application surjective de F sur E.

Démonstration. Cela découle de la proposition suivante. ■

Proposition et définition 8. Soient E et F deux ensembles non vides.
(i) Soit $f : E \to F$ une application injective. Il existe alors une application $r : F \to E$ telle que $r \circ f = \mathrm{Id}_E$. L'application r est surjective. On dit que c'est une *rétraction* de f.
(ii) Soit $g : F \to E$ une application surjective. Il existe alors une application $s : E \to F$ telle que $g \circ s = \mathrm{Id}_F$. L'application s est injective. On dit que c'est une *section* de g.

Démonstration. (i) Pour tout élément $y \in \operatorname{Im} f$, on définit $r(y)$ comme l'unique antécédent de y par f (cet antécédent existe parce que $y \in \operatorname{Im} f$, il est unique parce que f est supposée injective). Pour tout $y \in F \setminus \operatorname{Im} f$, on définit $r(y)$ comme un élément arbitraire de E (c'est possible parce que E est non vide). On vérifie sans peine que l'on a bien $r\big(f(x)\big) = x$ pour tout $x \in E$. Cela implique en particulier que tout $x \in E$ admet au moins un antécédent par r, à savoir $f(x)$. L'application r est donc bien surjective.
(ii) Pour tout $x \in E$, on définit $s(x)$ comme l'un quelconque des antécédents de x par g (c'est possible puisque celle-ci est surjective). On vérifie sans peine que l'on a bien $g\big(s(x)\big) = x$ pour tout $x \in E$. Si l'on a $s(x) = s(x')$, alors $x = g\big(s(x)\big) = g\big(s(x')\big) = x'$, et l'application s est bien injective. ■

Théorème 9 (Cantor, Schröder, Bernstein). S'il existe une injection de E dans F et une injection de F dans E, alors E et F sont équipotents.

La démonstration de ce théorème, bien qu'élémentaire, est assez difficile : en réserver la lecture pour un jour de forme !

Démonstration. Notons $f : E \to F$ et $g : F \to E$ ces injections. L'application $h := g \circ f$ réalise une bijection de E sur $h(E)$ et l'ensemble $E' := g(F)$ (qui est donc en bijection avec F) est tel que $h(E) \subset E' \subset E$. Il suffit de trouver une bijection de E sur E'.
Les images itérées $h^n(E)$ (définies par $h^0(E) := E$ et $h^{n+1}(E) := h\big(h^n(E)\big)$) forment une suite décroissante de parties de E. Soit R leur intersection. On vérifie que h induit une bijection de R sur lui-même. Nous poserons $A_0 := E \setminus h(E)$, puis, par récurrence, $A_{n+1} := h(A_n) = h^n(E) \setminus h^{n+1}(E)$. Ainsi, h induit une bijection de A_n sur A_{n+1} et les A_n sont deux à deux disjoints. On peut écrire :

$$E = (A_0 \cup \cdots \cup A_n \cup \cdots) \cup R,$$

la réunion étant disjointe. L'effet de h est donc d'envoyer A_n bijectivement sur A_{n+1} et R bijectivement sur lui-même.
Écrivons maintenant $A_0 = E \backslash h(E)$ comme union disjointe de $B_0 := E \backslash E'$ et de $C_0 := E' \backslash h(E)$. On introduit de même les images itérées $B_{n+1} := h(B_n)$ et $C_{n+1} := h(C_n)$, de sorte que

chaque A_n est union disjointe de B_n et de C_n. On a maintenant deux écritures en réunion disjointe :

$$\begin{aligned} E &= (B_0 \cup C_0 \cup B_1 \cup C_1 \cup \cdots \cup B_n \cup C_n \cup \cdots) \cup R, \\ E' &= (C_0 \cup B_1 \cup C_1 \cup \cdots \cup B_n \cup C_n \cup \cdots) \cup R. \end{aligned}$$

Il est maintenant facile de construire la bijection de E sur E' : sur les C_i et sur R, c'est l'identité ; sur chaque B_n, c'est la bijection $h : B_n \to B_{n+1}$. ■

Corollaire 10. La relation $\leqslant$ entre cardinaux est une relation d'ordre.

Démonstration. Puisque l'identité est injective et que la composée de deux applications injectives est injective, cette relation est bien réflexive et transitive. L'antisymétrie est conséquence du théorème qui précède. ■

Théorème 11. C'est un ordre total.

La démonstration de ce théorème est non seulement assez difficile, mais non élémentaire en ce qu'elle fait appel au lemme de Zorn (que nous avons admis) et donc à l'axiome du choix.

Démonstration. Soient E et F deux ensembles quelconques. Il s'agit de démontrer qu'il existe une injection de l'un dans l'autre. Nous allons démontrer qu'il existe une bijection de l'un sur une partie de l'autre, ce qui revient au même. Nous introduisons pour cela l'ensemble $\mathcal{E}$ dont les éléments sont les triplets (A, f, B) formés d'une partie A de E, d'une partie B de F et d'une bijection f de A sur B. Cet ensemble est non vide, car il contient $(\varnothing, f, \varnothing)$ (où f est l'unique bijection de $\varnothing$ sur lui-même). Nous poserons $(A, f, B) \preceq (A', f', B')$ si, et seulement si, $A \subset A'$, $B \subset B'$ et $f'_{|A} = f$ (autrement dit : la restriction[3] de f' à A est f). On vérifie aisément que c'est une relation d'ordre sur $\mathcal{E}$. On va montrer que celui-ci est inductif. Soit $\mathcal{C} = (A_i, f_i, B_i)_{i \in I}$ une chaîne de $\mathcal{E}$. Notons $A = \bigcup_{i \in I} A_i$ et $B = \bigcup_{i \in I} B_i$. Soit $x \in A$; quel que soit $i \in I$ tel que $x \in A_i$, $f_i(x) \in B$ a la même valeur. En effet, si $x \in A_j$, on a par exemple $(A_i, f_i, B_i) \preceq (A_j, f_j, B_j)$ (car $\mathcal{C}$ est une chaîne), donc f_i est la restriction de f_j, donc $f_j(x) = f_i(x)$. On peut donc poser $f(x) = f_i(x)$ pour n'importe quel $i \in I$ tel que $x \in A_i$. On voit facilement que $f : A \to B$ est une bijection, et que (A, f, B) est un élément de $\mathcal{E}$ qui majore $\mathcal{C}$: l'ensemble $\mathcal{E}$ est bien inductif. D'après le lemme de Zorn, il admet un élément maximal (A, f, B). Montrons que $A = E$ ou $B = F$. Si ce n'était pas le cas, il y aurait des éléments $a \in E \setminus A$ et $b \in F \setminus B$. En posant $A' = A \cup \{a\}$, $B' = B \cup \{b\}$ et en définissant $f' : A' \to B'$ par $f'_{|A} = f$ et $f'(a) = b$, on construirait un élément (A', f', B') de $\mathcal{E}$ qui majore strictement (A, f, B), contredisant ainsi la maximalité de ce dernier. On a donc bien $A = E$ ou $B = F$.

Dans le premier cas, $\operatorname{card} E \leqslant \operatorname{card} F$; dans le second cas, $\operatorname{card} F \leqslant \operatorname{card} E$. ■

Exemple. L'application $x \mapsto \{x\}$ est injective de E dans $\mathcal{P}(E)$. D'après le théorème 1 de la page 14 de Cantor, il n'existe pas de surjection de E dans $\mathcal{P}(E)$, donc pas d'injection de $\mathcal{P}(E)$ dans E (proposition 8 de la page précédente). On a donc $\operatorname{card} E < \operatorname{card} \mathcal{P}(E)$.

[3] C'est aussi une corestriction puisque f' va de A' dans B' et f de A dans B ; mais nous avons convenu de ne pas dénoter explicitement la corestriction lorsque ce n"était pas absolument indispensable.

Bien que les cardinaux ne forment pas un ensemble, on va définir des opérations sur eux, en s'autorisant à employer le vocabulaire de la partie 4. Soient E et F deux ensembles. Pour définir la somme des cardinaux E et F, on choisit arbitrairement un ensemble E' équipotent à E et un ensemble F' équipotent à F, de telle manière que E' et F' soient disjoints ; par exemple $E' := E \times \{0\}$ et $F' := F \times \{1\}$ sont disjoints et l'on a des bijections $\begin{array}{ccc} E & \longrightarrow & E' \\ x & \longmapsto & (x,0) \end{array}$ et $\begin{array}{ccc} F & \longrightarrow & F' \\ y & \longmapsto & (y,1). \end{array}$

Ce procédé (dû au mathématicien Von Neumann) est analogue au « renommage » en informatique. On pose alors $\operatorname{card} E + \operatorname{card} F = \operatorname{card}(E' \cup F')$; il est en effet immédiat qu'avec un autre choix E'', F'', on aurait une bijection de $E' \cup F'$ sur $E'' \cup F''$, d'où l'égalité $\operatorname{card}(E' \cup F') = \operatorname{card}(E'' \cup F'')$. On démontre alors facilement que l'addition des cardinaux est une loi associative, commutative, d'élément neutre 0. En revanche, aucun cardinal n'est simplifiable (exercice I.1.32 de la page 60).

Pour définir le produit de deux cardinaux, on vérifie comme ci-dessus que $\operatorname{card}(E \times F)$ ne dépend que de $\operatorname{card} E$ et de $\operatorname{card} F$. On le note $\operatorname{card} E \times \operatorname{card} F$. C'est encore une loi associative, commutative, d'élément neutre 1. Elle est de plus distributive par rapport à l'addition. On en déduit notamment que, si k est un entier naturel non nul (un cardinal fini), et $\aleph$ (cette lettre hébraïque, d'usage traditionnel dans la théorie depuis Cantor, se lit « aleph ») un cardinal quelconque, le produit $k\aleph$ est égal à la somme de k termes $\aleph + \cdots + \aleph$.

On remarque encore que $\operatorname{card}\big(\mathcal{F}(E,F)\big)$ ne dépend que de $\operatorname{card} E$ et de $\operatorname{card} F$. On le note $(\operatorname{card} F)^{\operatorname{card} E}$, ce qui se lit « $\operatorname{card} F$ à la puissance $\operatorname{card} E$ ». Cette notation paraîtra plus naturelle après lecture du module II.1, où il est démontré que si E et F sont finis et ont respectivement n et p éléments, alors $\mathcal{F}(E,F)$ est fini et a p^n éléments. De nombreuses règles sur les puissances de nombres restent valables (exercice I.1.27 de la page 60). Par exemple, si k est un entier naturel non nul (un cardinal fini), et $\aleph$ un cardinal quelconque, la puissance $\aleph^k$ est égale au produit de k facteurs $\aleph \cdots \aleph$.

Théorème 12. (i) Pour tout ensemble E, on a $\operatorname{card} \mathcal{P}(E) = 2^{\operatorname{card} E}$.
(ii) Pour tout cardinal $\aleph$, on a $\aleph < 2^{\aleph}$.

Démonstration. On a démontré (théorème 2 de la page 24) que les ensembles $\mathcal{P}(E)$ et $\mathcal{F}(E,\{0,1\})$ étaient équipotents, d'où la première assertion. La seconde découle de l'exemple de la page précédente. ∎

Toutes ces opérations sont des fonctions croissantes de chacun de leurs arguments. Par exemple, si $\aleph_1 \leqslant \aleph_2$, alors $\aleph_1\aleph_3 \leqslant \aleph_2\aleph_3$ (ici, $\aleph_1$, $\aleph_2$ et $\aleph_3$ désignent des cardinaux quelconques). Cette règle se prouve en remarquant que si $f : E_1 \to E_2$ est injective, alors l'application $f \times \mathrm{Id}_{E_3} : E_1 \times E_3 \to E_2 \times E_3$ l'est aussi.

6.3 Cardinaux finis et cardinaux infinis

Nous proposons dans cette section une construction de $\mathbb{N}$ et une définition de la finitude et de la dénombrabilité plus proches de la théorie des ensembles « pure » que de la pratique courante des mathématiques. Ces notions seront reprises d'un point de vue plus pratique dans le module II.1.

Définition 24. Un ensemble E est dit *fini* s'il n'existe pas de bijection de E sur une partie de E autre que lui-même ; il est dit *infini* dans le cas contraire.

De manière équivalente, pour qu'un ensemble E soit infini, il faut, et il suffit, qu'il existe une

application injective non bijective de E dans lui-même. Il y a un critère similaire pour les surjections (que vous devriez formuler, avant de prouver les deux critères).
On démontre par récurrence que chacun des ensembles de la suite définie par $E_0 = \varnothing$, $E_{k+1} = E_k \cup \{E_k\}$ est fini (exercice I.1.29 de la page 60), puis que tout ensemble fini est équipotent à l'un de ces ensembles et à un seul. À ce stade, on ne peut cependant pas démontrer l'existence d'ensembles infinis : c'est le rôle de l'*axiome de l'infini*. Nous l'énonçons sous une forme un peu spéciale.

Axiome et définition 15.
Les « entiers de Von Neumann » E_k forment un ensemble, que l'on note $\mathbb{N}$.

Ayant postulé son existence, on peut en effet démontrer que cet ensemble est infini. Réciproquement, on peut démontrer que, s'il existe un ensemble infini E , les E_k forment un ensemble $\mathbb{N}$ et il existe une application injective de $\mathbb{N}$ dans E (exercice I.1.29 de la page 60).

Pour que l'on puisse considérer ce qui précède comme une véritable construction de $\mathbb{N}$, il faudrait que l'on puisse formuler et démontrer un principe de récurrence.
C'est en effet possible de la manière suivante. On définit, pour tout ensemble E, son « successeur » :

$$\mathrm{succ}(E) := E \cup \{E\}.$$

Le principe de récurrence dit simplement que toute partie de $\mathbb{N}$ qui contient 0 (c'est-à-dire, dans ce modèle, $\varnothing$) et qui est stable par l'application succ, est égale à $\mathbb{N}$.
Une construction purement ensembliste de $\mathbb{N}$ par les des entiers de Von Neumann, qui ne fait pas appel à « l'arithmétique intuitive », implicite dans la numérotation $E_0, E_1, \ldots$, est présentée à la section 6.3.2.
Des propriétés qui dérivent de la présentation ci-dessus font l'objet des exercices I.1.22 et I.1.29.

Les ensembles finis et leurs cardinaux (qui sont les entiers naturels, c'est-à-dire les éléments de $\mathbb{N}$) seront abordés par une voie différente et plus intuitive au module II.1. Nous ne retiendrons que les faits suivants :

1. L'ensemble $\mathbb{N}$ des entiers naturels est infini. Son cardinal est noté $\aleph_0$ (lire Aleph zéro).
2. Tout ensemble fini se plonge injectivement dans $\mathbb{N}$. On a donc, pour tout entier naturel n, considéré comme un cardinal : $n < \aleph_0$.
3. Pour tout ensemble infini E, il existe une injection de $\mathbb{N}$ dans E. Donc, pour tout cardinal infini $\aleph$, on a $\aleph_0 \leqslant \aleph$.

6.3.1 Cardinaux infinis

Ainsi, $\aleph_0$ est le plus petit des cardinaux infinis. Nous dirons qu'un ensemble E est *dénombrable* s'il existe une *énumération* de E, c'est-à-dire une bijection de $\mathbb{N}$ sur E. Cela équivaut à dire que E est équipotent à $\mathbb{N}$, *i.e.* que $\mathrm{card}\ E = \aleph_0$. Nous allons prouver quelques formules concernant $\aleph_0$ en nous basant en partie sur des bijections explicites (et nous laisserons alors au lecteur le soin d'expliciter les bijections réciproques), en partie sur l'arithmétique des cardinaux.

Tout d'abord, on a donné dans l'exercice 4 de la page 6 une énumération de $\mathbb{Z}$. On en déduit que $\mathrm{card}\ \mathbb{Z} = \aleph_0$. Mais $\mathbb{Z}$ est l'union disjointe de $\mathbb{N}$ et de $-\mathbb{N}^*$, et ce dernier ensemble est énuméré par $n \mapsto -n-1$. On a donc, par définition de la somme des cardinaux, la formule : $\aleph_0 + \aleph_0 = \aleph_0$.

On peut énumérer $\mathbb{N} \times \mathbb{N}$ en parcourant successivement les ensembles suivants : $\{(0,0)\}$, puis $\{(1,0),(0,1)\}$, puis $\{(2,0),(1,1),(0,2)\}$, etc. Le lecteur pourra démontrer (c'est *très* amusant) que l'image de n s'obtient comme suit : on encadre n par $\dfrac{k(k+1)}{2} \leqslant n < \dfrac{(k+1)(k+2)}{2}$,

ce qui revient à écrire $n = \dfrac{k(k+1)}{2} + \ell$, avec $0 \leqslant \ell \leqslant k$. Cette écriture est unique. On associe alors à n le couple $(k-\ell, \ell) \in \mathbb{N} \times \mathbb{N}$. De cette énumération, on déduit que $\operatorname{card}\big(\mathbb{N} \times \mathbb{N}\big) = \aleph_0$. On a donc, par définition du produit des cardinaux, la formule : $\aleph_0 \aleph_0 = \aleph_0$.

Comme application, calculons $\operatorname{card} \mathbb{Q}$. Comme $\mathbb{N} \subset \mathbb{Q}$, on a $\aleph_0 \leqslant \operatorname{card} \mathbb{Q}$. Par ailleurs, on définit une application injective en associant à tout $x \in \mathbb{Q}$ l'unique couple $(p, q) \in \mathbb{Z} \times \mathbb{N}^*$ tel que $\dfrac{p}{q} = x$ et p, q sont premiers entre eux. On a donc $\operatorname{card} \mathbb{Q} \leqslant \aleph_0 \aleph_0 = \aleph_0$. Comme $\leqslant$ est une relation d'ordre, on a $\operatorname{card} \mathbb{Q} = \aleph_0$.

On peut enfin démontrer que l'ensemble $\mathcal{P}_f(\mathbb{N})$ des parties finies de $\mathbb{N}$ est dénombrable (exercice I.1.19 de la page 59).

Pour établir le prochain théorème, nous allons énoncer quelques règles :

1. Soient $\aleph_1$ et $\aleph_2$ des cardinaux. Alors : $\aleph_1 \leqslant \aleph_2 \iff \exists \aleph_3 \ : \ \aleph_2 = \aleph_1 + \aleph_3$. Dire que $\aleph_1 \leqslant \aleph_2$ équivaut en effet à dire qu'il existe des ensembles E_1, E_2 de cardinaux $\aleph_1, \aleph_2$ et tels que $E_1 \subset E_2$, c'est-à-dire tels que E_2 est l'union disjointe de E_1 et d'un ensemble E_3.
2. Soit $\aleph$ un cardinal. Alors : $\aleph_0 \leqslant \aleph \iff \aleph = \aleph_0 + \aleph$. En effet, si $\aleph_0 \leqslant \aleph$, on peut écrire $\aleph = \aleph_0 + \aleph_1$, d'où $\aleph + \aleph_0 = \aleph_0 + \aleph_0 + \aleph_1 = \aleph_0 + \aleph_1$. La réciproque est immédiate.
3. Soit A un ensemble dénombrable qui est inclus dans un ensemble B non dénombrable. Alors $B \setminus A$ est équipotent à B. En effet, l'ensemble $B \setminus A$ est infini, sinon B serait dénombrable. Soit $\aleph$ son cardinal. Alors $\aleph = \aleph_0 + \aleph$ (règle ci-dessus), mais $\aleph_0 + \aleph$ est le cardinal de B.

Théorème et définition 13.
Le cardinal de $\mathbb{R}$ est $2^{\aleph_0}$. On dit qu'un ensemble équipotent à $\mathbb{R}$ *a la puissance du continu*.

Démonstration. La numération en base 2 des nombres réels assure que tout $x \in \,]0,1[$ admet une unique écriture sous la forme $\sum\limits_{n \geqslant 0} \varepsilon_n 2^{-n}$, où la suite $(\varepsilon_n)_{n \in \mathbb{N}}$ est un élément de $\{0,1\}^{\mathbb{N}}$ tel que : les ε_n ne sont pas tous nuls ; ils ne sont pas égaux à 1 à partir d'un certain rang. On obtient ainsi une bijection de $]0,1[$ sur $\{0,1\}^{\mathbb{N}} \setminus A$, où A est une partie dénombrable de $\{0,1\}^{\mathbb{N}}$. D'après la dernière règle, $\operatorname{card} \,]0,1[\, = 2^{\aleph_0}$. On conclut à l'aide de n'importe quelle bijection de $\mathbb{R}$ sur $]0,1[$, par exemple l'application $x \mapsto \dfrac{1}{2}\left(1 + \dfrac{x}{\sqrt{1+x^2}}\right)$. ∎

L'un des plus célèbres problèmes de la théorie des ensembles fut longtemps de savoir s'il existait un cardinal entre la puissance du dénombrable et celle du continu, autrement dit, tel que $\aleph_0 < \aleph < 2^{\aleph_0}$. L'*hypothèse du continu*, formulée par Cantor, affirmait que c'était impossible. Des théorèmes dûs aux logiciens Kurt Gödel en 1939 et Paul Cohen en 1963 entraînent que *ni l'hypothèse du continu ni sa négation ne sont démontrables dans le cadre de la théorie usuelle des ensembles.*

6.3.2 Construction purement ensembliste de $\mathbb{N}$

On reprend la notation introduite plus haut pour tout ensemble E :

$$\operatorname{succ}(E) := E \cup \{E\}.$$

Ainsi, les ensembles $E_0, E_1, \ldots$ sont obtenus à partir de $E_0 := \varnothing$ par *itération* de la fonction succ et c'est cette itération qui sous-entend le recours à l'arithmétique intuitive. Ce qui suit vise à l'éviter. Nous commençons par une variante de l'axiome de l'infini (c'est d'ailleurs sa forme authentique) :

Axiome 16 (Axiome de l'infini purement ensembliste).
Il existe un ensemble $\mathcal{E}$ tel que $\varnothing \in \mathcal{E}$ et que $\forall x \in \mathcal{E}\ ,\ \operatorname{succ}(x) \in \mathcal{E}$.

Nous allons fonder la construction de $\mathbb{N}$ sur cet ensemble $\mathcal{E}$, qui n'est certainement pas unique ; mais nous verrons plus loin que notre construction ne dépend pas d'un choix particulier, elle est intrinsèque.

Il résulte de la définition que l'on a une application $x \mapsto \mathrm{succ}(x)$ de $\mathcal{E}$ dans lui-même. Nous la noterons S. L'ensemble des parties de $\mathcal{E}$ qui sont stables par S est clos par intersection arbitraire, et l'objet suivant est donc bien défini :

Définition 25. On appelle $\mathbb{N}$ l'intersection de toutes les parties de $\mathcal{E}$ dont $\varnothing$ est élément et qui sont stables par S.

C'est donc le plus petit sous-ensemble de $\mathcal{E}$ dont $\varnothing$ soit élément et qui soit stable par S. Le principe de récurrence en découle immédiatement. Pour le formuler, on introduit la notation $0 := \varnothing$.

Proposition 14. Soit $P(x)$ une propriété des éléments de $\mathbb{N}$ telle que : $P(0)$ est vraie ; et, pour tout $x \in \mathbb{N}$, si $P(x)$ est vraie alors $P(S(x))$ est vraie. Alors, pour tout $x \in \mathbb{N}$, $P(x)$ est vraie.

Démonstration. L'ensemble des $x \in \mathbb{N}$ tels que $P(x)$ soit vraie est une partie de $\mathbb{N}$ dont $\varnothing$ est élément et qui est stable par S, c'est donc $\mathbb{N}$. ∎

Corollaire 15.
Soit $\mathcal{E}'$ un ensemble vérifiant les propriétés de l'axiome de l'infini. Alors $\mathbb{N} \subset \mathcal{E}'$.

Démonstration.
On prouve immédiatement par récurrence que tous les éléments de $\mathbb{N}$ appartiennent à $\mathcal{E}'$. ∎

Il en découle que $\mathbb{N}$ est l'intersection de toutes les parties de $\mathcal{E}'$ dont $\varnothing$ est élément et qui sont stables par S', application induite par succ sur $\mathcal{E}'$: autrement dit, notre construction de $\mathbb{N}$ est bien intrinsèque. Pour aller plus loin dans l'étude, nous aurons besoin de deux nouvelles notions.

Définition 26. Un ensemble E est dit *héréditaire* si :

$$\forall x \in E \ , \ x \subset E.$$

Il est dit *strictement héréditaire* si de plus $E \notin E$.

Noter que, dans certaines axiomatiques de la théorie des ensemble, la condition $E \notin E$ irait de soi et on n'aurait pas besoin de faire la différence entre les deux notions ci-dessus (voir par exemple l'exercice I.1.4).

Lemme 16. Si E est héréditaire, resp. strictement héréditaire, il en est de même de $\mathrm{succ}(E)$.

Démonstration. Notons d'abord que, par définition, $E \subset \mathrm{succ}(E)$. Soit $x \in \mathrm{succ}(E)$, *i.e.* $x \in E$ ou $x = E$. Si $x = E$, puisque $E \subset \mathrm{succ}(E)$, on conclut bien que $x \subset \mathrm{succ}(E)$. Si $x \in E$, comme E est héréditaire, on a $x \subset E$; comme $E \subset \mathrm{succ}(E)$, on conclut encore que $x \subset \mathrm{succ}(E)$. Si maintenant E est strictement héréditaire, du fait que $E \in \mathrm{succ}(E)$ alors que $E \notin E$, on déduit que l'inclusion $E \subset \mathrm{succ}(E)$ est stricte. On ne peut donc avoir $\mathrm{succ}(E) \subset E$, donc $\mathrm{succ}(E) \in E$ est impossible (puisque $\mathrm{succ}(E)$ est héréditaire) et $\mathrm{succ}(E) = E$ aussi (inclusion stricte) donc $\mathrm{succ}(E) \in \mathrm{succ}(E)$ est également impossible et $\mathrm{succ}(E)$ est bien strictement héréditaire. ∎

Puisque $\varnothing$ est, de manière évidente, strictement héréditaire, il vient par récurrence que tous les éléments de $\mathbb{N}$ le sont.

Lemme 17. Si E et F sont héréditaires, et si $\mathrm{succ}(E) \subset \mathrm{succ}(F)$, alors $E \subset F$.

Démonstration. Puisque $E \in \mathrm{succ}(E)$, on a $E \in \mathrm{succ}(F)$ donc soit $E \in F$ qui entraîne (F étant héréditaire) la conclusion, soit $E = F$ qui l'entraîne aussi. ■

L'application $S : \mathbb{N} \to \mathbb{N}$ est donc injective. Comme $x \in S(x)$, on ne peut avoir $S(x) = \varnothing$ et elle n'est donc pas bijective. D'après la caractérisation qui suit la définition 24, on en déduit :

Théorème 18. L'ensemble $\mathbb{N}$ est infini.

Nous laissons au lecteur le soin de poursuivre la théorie et d'établir les propriétés naturelles attendues de $\mathbb{N}$. Voici un guide, constitué de faits qui s'enchaînent sans difficulté :

1. La relation $\in$ est une relation d'ordre strict (*i.e.* transitive et antiréflexive) sur $\mathbb{N}$.
   Notons-la $<$ et notons $\leqslant$ la relation d'ordre associée.
2. On a $x < \mathrm{succ}(x)$ pour tout $x \in \mathbb{N}$.
3. Soit $D(x)$ (ensemble des « descendants » de x) le plus petit sous-ensemble de N dont x est élément et stable par S. Il se construit comme on a construit $\mathbb{N}$ à partir de $\varnothing$ et il admet de même un principe de récurrence (avec initialisation à x) que l'on va exploiter ci-dessous.
4. Tout $y \in D(x)$ majore x ; de plus, $D(x) = \{x\} \cup D(S(x))$ (union disjointe).
5. Soit $A(x)$ (ensemble des « ascendants » de x) l'ensemble des $y \in \mathbb{N}$ tels que $x \in D(y)$. Alors $\mathbb{N} = A(x) \cup D(x)$ et $A(x) \cap D(x) = \{x\}$.

Une fois ces faits établis, la conclusion suivante est immédiate :

Théorème 19. L'ensemble $(\mathbb{N}, \leqslant)$ est totalement ordonné.

7 Rudiments de logique

En mathématiques, on s'intéresse à des *objets* : éléments (par exemple nombres), ensembles, applications… On formule sur ces objets des *assertions*, qui obéissent (indépendamment de leurs contenus concrets) à des règles logiques. Si, partant des *axiomes* ou des *définitions* on peut, en respectant ces règles logiques, *démontrer* une assertion, elle acquiert le statut de *théorème*. La démonstration de théorèmes est l'un des principaux buts des mathématiques. Le mot « théorème » a de nombreux synonymes : proposition, lemme, corollaire… selon l'importance et l'usage. Disons, en première approximation, qu'une proposition est un peu moins importante qu'un théorème ; qu'un lemme est avant tout un outil ou une étape commode dans la démonstration d'une proposition ou d'un théorème ; enfin, qu'un corollaire est une conséquence plus ou moins immédiate d'une proposition ou d'un théorème.

7.1 Logique propositionnelle

Les règles logiques qui structurent une démonstration disent comment l'on détermine si des affirmations sont vraies ou fausses en sachant que d'autres affirmations sont vraies ou fausses. L'usage est d'appeler *propositions* les affirmations sur lesquelles on raisonne. Dans toute cette section, le mot « proposition » sera exclusivement employé dans ce sens.

Il y a d'abord des propositions élémentaires, au contenu concret : $\mathrm{i}^2 = -1$, $\pi \in \mathbb{Q}$, $5 < 7$, 5 est pair ... Elles sont souvent des *relations*, ici : égalité $=$, appartenance $\in$, comparaison $\leqslant$, divisibilité $\mid$ ... Du point de vue de la logique élémentaire, on n'attache aucune importance au sens de ces propositions (du moment qu'elles sont correctement construites) : on demande seulement qu'elles aient une *valeur de vérité*, *i.e.* qu'elles soient *vraies* ou *fausses*. Ainsi, les quatre propositions prises en exemple plus haut sont respectivement : vraie, fausse, vraie, fausse.

On peut bâtir de nouvelles propositions à l'aide de *connecteurs logiques*. Les principaux sont *non* (symbole $\neg$), *ou* (symbole $\vee$), *et* (symbole $\wedge$), *implique* (symbole $\Rightarrow$) et *équivaut* (symbole $\Leftrightarrow$). Bien qu'ils évoquent des processus intellectuels familiers, on n'y attache pas vraiment de signification : on veut simplement qu'ils expriment des opérations sur les valeurs de vérité. Par exemple, on demande que $\neg P$ (lire « non P ») ait pour valeur « vrai » si P a pour valeur « faux », et pour valeur « faux » si P a pour valeur « vrai ». Chacun de ces quatre connecteurs *binaires* (c'est-à-dire à deux arguments P et Q) admet une description du même genre :

1. On veut que $P \vee Q$ (lire « P ou Q ») ait pour valeur « vrai » si, et seulement si, P ou Q a pour valeur « vrai », et pour valeur « faux » dans le cas contraire. Le $\vee$, encore appelé « ou logique », est appelé *disjonction* et l'on doit remarquer que *la disjonction mathématique n'est pas exclusive* : dire que « P ou Q » est vrai n'exclut pas que « P et Q » soit vrai.
2. On veut que $P \wedge Q$ (lire « P et Q ») ait pour valeur « vrai » si, et seulement si, P et Q ont tous deux pour valeur « vrai », et que, dans tout autre cas, $P \wedge Q$ ait pour valeur « faux ». Le $\wedge$, encore appelé « et logique », est appelé *conjonction*.
3. On veut que $P \Rightarrow Q$ (lire « P implique Q ») ait pour valeur « vrai » si, ou bien P a pour valeur « vrai » et Q aussi, ou bien P a pour valeur « faux » et Q est quelconque.
4. On veut que $P \Leftrightarrow Q$ (lire « P équivaut à Q ») ait pour valeur « vrai » si P et Q ont même valeur de vérité, et pour valeur « faux » si P et Q n'ont pas même valeur de vérité.

Au lieu de dire que P a pour valeur « vrai » ou pour valeur « faux », on peut procéder comme en algèbre et considérer P comme une *variable propositionnelle* à laquelle on peut affecter les valeurs V et F (pour « vrai » et « faux »). Les règles ci-dessus se représentent par des *tables de vérité* :

P	$\neg P$
V	F
F	V

P	Q	$P \vee Q$	$P \wedge Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
V	V	V	V	V	V
V	F	V	F	F	F
F	V	V	F	V	F
F	F	F	F	V	V

Le « ou » des mathématiciens des logiciens est dit *inclusif* à cause de la règle $\mathrm{V} \vee \mathrm{V} = \mathrm{V}$. Intuitivement, elle dit par exemple que la condition « il fait beau ou la vie est belle » est réalisée si par chance il fait beau *et* que la vie est belle : ce « ou » n'exprime donc pas un dilemme, contrairement peut-être à l'usage courant.

Naturellement, ces connecteurs ne sont pas indépendants. Par exemple, il est évident que $P \Leftrightarrow Q$ a toujours même valeur que $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$. On peut le vérifier intuitivement en dressant des tables de vérité :

P	Q	$P \Rightarrow Q$	$Q \Rightarrow P$	$(P \Rightarrow Q) \wedge (Q \Rightarrow P)$
V	V	V	V	V
V	F	F	V	F
F	V	V	F	F
F	F	V	V	V

L'expression $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$ a donc même table de vérité que $P \Leftrightarrow Q$. On les considère comme *logiquement équivalentes*. On pourrait d'ailleurs *définir* la seconde à partir de la première. Nous exprimerons cette équivalence comme une égalité :

$$(P \Leftrightarrow Q) = (P \Rightarrow Q) \wedge (Q \Rightarrow P).$$

De même, on vérifie facilement l'égalité suivante :

$$(P \Rightarrow Q) = \big((\neg P) \vee Q\big).$$

L'usage mathématique du connecteur logique « implique » s'éloigne parfois de l'intuition. Par exemple, on peut dire que

- Si la terre tourne autour du soleil, 2 est pair (les deux sont vrais).
- Si le soleil tourne autour de la terre, 2 est impair (les deux sont faux).
- Si $1 = 2$, alors $1 \neq 2$ (le premier est faux et le second vrai).

Cette façon un peu mécanique de manier l'implication logique est parfois appelée « implication matérielle »

À cause des égalités énoncées ci-dessus, dans certaines présentations de la logique propositionnelle, on part des seuls connecteurs $\neg$, $\vee$ et $\wedge$ et l'on en dérive les autres. Voici maintenant les règles de calcul fondamentales concernant ces connecteurs. Le lecteur est invité à les vérifier à l'aide de tables de vérité. Il est également invité à réfléchir sur l'analogie

de ces règles avec celles qui portaient sur les opérations ensemblistes en 1.3.

$$
\begin{aligned}
P \vee (Q \vee R) &= (P \vee Q) \vee R \\
P \vee Q &= Q \vee P \\
\mathrm{V} \vee P = P \vee \mathrm{V} &= \mathrm{V} \\
\mathrm{F} \vee P = P \vee \mathrm{F} &= P \\
P \vee P &= P \\
P \wedge (Q \wedge R) &= (P \wedge Q) \wedge R \\
P \wedge Q &= Q \wedge P \\
\mathrm{V} \wedge P = P \wedge \mathrm{V} &= P \\
\mathrm{F} \wedge P = P \wedge \mathrm{F} &= \mathrm{F} \\
P \wedge P &= P \\
P \wedge (Q \vee R) &= (P \wedge Q) \vee (P \wedge R) \\
P \vee (Q \wedge R) &= (P \vee Q) \wedge (P \vee R) \\
\neg(\neg P) &= P \\
\neg(P \vee Q) &= (\neg P) \wedge (\neg Q) \\
\neg(P \wedge Q) &= (\neg P) \vee (\neg Q)
\end{aligned}
$$

Remarquons qu'en appliquant les trois dernières lois, on obtient l'égalité :

$$P \wedge Q = \neg\big((\neg P) \vee (\neg Q)\big).$$

Par conséquent, on pourrait définir tous les connecteurs logiques à partir des deux seuls connecteurs $\neg$ et $\vee$. En réalité, on peut faire mieux (ou pire ?), voir l'exercice I.1.34 de la page 60.

Sans en avoir formellement introduit la définition, nous manipulons depuis un moment des *formules propositionnelles* obtenues en combinant des arguments P, Q, R ... à l'aide de connecteurs imbriqués.

Définition 27. On appelle *tautologie* une formule propositionnelle qui prend la valeur « vrai » quelles que soient les valeurs données à ses arguments. On appelle *contradiction* une formule qui prend la valeur « faux » quelles que soient les valeurs données à ses arguments. Une formule qui n'est pas une contradiction est dite *satisfiable* .

On vérifie immédiatement que les formules suivantes sont des tautologies : $P \vee (\neg P)$, $P \Rightarrow P$, $(P \wedge Q) \Rightarrow P$, $P \Rightarrow (P \vee Q)$. On y reconnaît des règles fondamentales de la logique (par exemple, la première est le *principe du tiers exclu*). De même, la formule $P \wedge (\neg P)$ est une contradiction. Sa négation (qui est donc une tautologie) est le *principe de non contradiction*. En fait, ce ne sont que des cas particuliers de principes plus généraux : si l'on remplace dans ces formules les arguments P et Q par des formules

quelconques (mettant en jeu autant d'arguments que l'on voudra), on obtient encore des tautologies.

Remarque. Il peut sembler facile de déterminer si une formule logique est satisfiable : on dresse sa table de vérité et l'on voit si la valeur V y apparaît. Cependant, si la formule comporte n variables propositionnelles, la table de vérité en comporte 2^n, ce qui peut être très grand. Même pour des formules de type assez particulier (« formes clausales »), on ne connaît actuellement aucun algorithme efficace pour déterminer leur satisfiabilité : c'est le problème dit SAT , et il est lié à l'un des sept « problèmes du millénaire » de la fondation Clay, dont la résolution serait récompensée de un million de dollars, le problème $P = NP$?.

7.2 Prédicats et quantificateurs

Un *prédicat* est une proposition $P(x)$ qui dépend d'une variable x. Définissons par exemple $A(x)$ par « x est pair », $B(x)$ par $x > x - 1$ et $C(x)$ par $x \neq x$. Le premier a un sens si x est entier, le second si x est réel, le troisième pour un x arbitraire. Nous nous cantonnerons, autant que possible, à des prédicats définis sur un ensemble (par exemple $\mathbb{Z}$ et $\mathbb{R}$ en ce qui concerne A et B). La valeur de vérité de $P(x)$ dépend maintenant du choix de x. Autrement dit, si l'on remplace la variable x par une valeur déterminée, on obtient une proposition qui pourra être vraie ou fausse. Ainsi, la proposition $A(2)$ est vraie alors que $A(1)$ est fausse, le prédicat $B(x)$ est vrai (c'est-à-dire, prend la valeur vrai) pour tout réel x et $C(x)$ n'est vrai pour aucun x.

On peut également considérer des prédicats à plusieurs variables $P(x, y)$, etc. Définissons par exemple $R(x, y)$ par $x \mid y$, $S(x, y)$ par $x \in \{x, y\}$ et $T(x, y)$ par $(x < y) \wedge (y < x)$. Le premier est défini, par exemple, pour $x, y \in \mathbb{N}$ (et vrai ou faux selon leurs valeurs), le second pour x, y quelconques (et toujours vrai) et le troisième pour x, y réels (et jamais vrai). Il n'est d'ailleurs pas nécessaire que les arguments $x, y, \ldots$ prennent leurs valeurs dans le même domaine, comme le montre la relation « x majore A », où $x \in \mathbb{R}$ et $A \in \mathcal{P}(\mathbb{R})$.

On peut opérer logiquement sur les prédicats, avec les mêmes connecteurs que précédemment. Par exemple, si $P(x)$ et $Q(x)$ sont des prédicats, $P(x) \vee Q(x)$ est un prédicat que l'on pourra noter $P \vee Q$. Cela peut d'ailleurs être source d'ambiguïté. Ainsi, lorsque l'on écrit $x > 1 \Rightarrow x^2 > 1$, on devrait en principe y voir un prédicat $P(x)$ qui dépend de l'argument $x \in \mathbb{R}$. Mais un abus courant veut que l'on entende par là $\forall x \in \mathbb{R}\ ,\ x > 1 \Rightarrow x^2 > 1$, qui est une proposition dans laquelle il n'y a plus de variable. Nous allons étudier de telles constructions en évitant soigneusement ce genre d'abus.

7.2.1 Quantificateurs

Le prédicat $x = x$ est vrai quel que soit x. On exprime cela en disant que la proposition $\forall x, x = x$ est vraie. Le point crucial, dans cette affirmation, est que c'est bien une proposition. Elle ne dépend pas de x. On dit que la lettre x n'y figure plus, ou encore qu'elle est *muette*. On peut la remplacer par y une lettre y quelconque : on considère que la proposition $\forall x, x = x$ et la proposition $\forall y, y = y$ sont identiques.

Lorsque $P(x)$ est un prédicat défini sur un ensemble E, on définit de même $\forall x \in E\ ,\ P(x)$, ce qui se lit « quel que soit $x \in E$, $P(x)$ » (ou encore, « pour tout x ... »). Autant que possible, nous n'utiliserons que des quantificateurs portant sur un ensemble E. Il y a d'ailleurs d'autres écritures possibles : $\forall x > 0$ signifie implicitement $\forall x \in \mathbb{R}_+^*$, etc. Lorsque $E = \{a_1, \ldots, a_n\}$, la proposition $\forall x \in E\ ,\ P(x)$ est équivalente à la suivante : $P(a_1) \wedge \cdots \wedge P(a_n)$, et l'on voit bien pourquoi x n'y figure pas. Le quantificateur $\forall$ est donc une sorte de généralisation du connecteur $\wedge$.

Plus généralement, si $P(x, y, z, \ldots)$ est un prédicat qui dépend de plusieurs arguments $x, y, z, \ldots$ variant dans un ensemble E, les formules quantifiées $\forall x \in E\ ,\ P(x, y, z, \ldots)$ et $\exists x \in E\ :\ P(x, y, z, \ldots)$ définissent des prédicats qui ne dépendent plus que des arguments $y, z, \ldots$. La lettre x n'y figure plus, elle est devenue muette. Mais les lettres $y, z, \ldots$ y figurent encore, elles sont « parlantes ». Si par exemple, dans le prédicat sur $\mathbb{R}^2$ défini par $P(x, y) := (y = x^2)$, on quantifie la variable x par $\exists$, on obtient le prédicat sur $\mathbb{R}$ défini par $Q(y) := (\exists x \in \mathbb{R}\ :\ y = x^2)$, dans lequel la lettre x est muette et la lettre y ne l'est pas.

Reprenons le prédicat $A(x)$ défini par « x est pair ». La proposition $\forall x \in \mathbb{Z}\ ,\ A(x)$ est fausse, puisque $A(1)$ l'est. Mais $A(2)$ est vraie, et l'on en conclut que la proposition $\exists x \in \mathbb{Z}\ :\ A(x)$ (qui se lit « il existe $x \in \mathbb{Z}$ tel que $A(x)$ ») est vraie. On précise parfois la formulation en disant « il existe *au moins un* $x \in \mathbb{Z}$ tel que $A(x)$ ». En effet, on affirme ici l'existence d'un tel x, sans rien dire de l'unicité. Lorsque l'on veut affirmer l'unicité, on écrit $\exists !$, lire « il existe un unique ... ». Par exemple, la proposition $\exists ! x \in \{1, 2, 3\}\ :\ x$ est pair est vraie. Nous n'étudierons pas en détail cette variante.

Si $P(x)$ est un prédicat, on peut former la proposition $\exists x\ :\ P(x)$. Si le prédicat $P(x)$ est défini sur l'ensemble E, on formera de même $\exists x \in E\ :\ P(x)$ (et nous préférerons cet usage). Lorsque $E = \{a_1, \ldots, a_n\}$, la proposition $\exists x \in E\ :\ P(x)$ est équivalente à la suivante : $P(a_1) \vee \cdots \vee P(a_n)$, et l'on voit bien pourquoi x n'y figure pas. Le quantificateur $\exists$ est donc une sorte de généralisation du connecteur $\vee$.

Le *quantificateur existentiel* $\exists$ obéit aux mêmes règles que le *quantificateur universel* $\forall$. Il transforme un prédicat $P(x)$ sur l'ensemble E en une proposition $\exists x \in E\ :\ P(x)$ dans laquelle la lettre x est muette. De même, un prédicat $P(x, y, z, \ldots)$ serait transformé en un prédicat $\exists x \in E\ :\ P(x, y, z, \ldots)$ portant sur les seules variables $y, z, \ldots$.

7.2.2 Quantificateurs et connecteurs logiques

Voyons l'effet des quantificateurs sur les prédicats construits à l'aide de connecteurs logiques. La première paire de règles concerne la négation :

$$\big(\forall x \in E\ ,\ \neg P(x)\big) \Longleftrightarrow \neg\big(\exists x \in E\ :\ P(x)\big),$$
$$\big(\exists x \in E\ :\ \neg P(x)\big) \Longleftrightarrow \neg\big(\forall x \in E\ ,\ P(x)\big).$$

Nier qu'il existe un élément ayant une propriété, c'est affirmer que tout élément a la propriété contraire. Nier que tout élément ait une propriété, c'est affirmer l'existence d'un élément ayant la propriété contraire. Cette règle a une conséquence curieuse au premier abord, mais que l'on finit par trouver raisonnable. Si E est l'ensemble vide, alors $\forall x \in E\ ,\ P(x)$ est vrai, indépendamment du prédicat P. En effet, le nier reviendrait à affirmer que $\exists x \in E\ :\ \neg P(x)$, ce qui n'est pas possible puisqu'il n'existe aucun $x \in E$.

Voici les règles concernant la conjonction :

$$\big(\forall x \in E\ ,\ (P(x) \wedge Q(x))\big) \Longleftrightarrow \big((\forall x \in E\ ,\ P(x)) \wedge (\forall x \in E\ ,\ Q(x))\big), \tag{4}$$
$$\big(\exists x \in E\ :\ (P(x) \wedge Q(x))\big) \Longrightarrow \big((\exists x \in E\ :\ P(x)) \wedge (\exists x \in E\ :\ Q(x))\big). \tag{5}$$

Remarquons simplement que la réciproque de cette dernière règle n'est pas vraie : prendre $E = \mathbb{Z}$, et les prédicats « x est pair » et « x est impair ».

Voici les règles concernant la disjonction :

$$\big((\forall x \in E\ ,\ P(x)) \vee (\forall x \in E\ ,\ Q(x))\big) \Longrightarrow \big(\forall x \in E\ ,\ (P(x) \vee Q(x))\big)$$
$$\big(\exists x \in E\ :\ (P(x) \vee Q(x))\big) \Longleftrightarrow \big((\exists x \in E\ :\ P(x)) \vee (\exists x \in E\ :\ Q(x))\big).$$

Cette fois, c'est la réciproque de la première règle qui est fausse : prendre le même exemple que précédemment !

On peut combiner les quantificateurs portant sur plusieurs variables successivement. Pour décrire les règles logiques correspondantes, nous prendrons le cas d'un prédicat à deux variables $P(x, y)$ défini pour $x \in E$ et $y \in F$. S'il y a plus de variables, des règles analogues s'appliquent, que le bon sens permet de deviner dans tous les cas. Voici les règles fondamentales.

$$\big(\forall x \in E\ ,\ \forall y \in F\ ,\ P(x, y)\big) \Longleftrightarrow \big(\forall y \in F\ ,\ \forall x \in E\ ,\ P(x, y)\big),$$
$$\big(\exists x \in E\ :\ \exists y \in F\ :\ P(x, y)\big) \Longleftrightarrow \big(\exists y \in F\ :\ \exists x \in E\ :\ P(x, y)\big),$$
$$\big(\exists x \in E\ :\ \forall y \in F\ ,\ P(x, y)\big) \Longrightarrow \big(\forall y \in F\ ,\ \exists x \in E\ :\ P(x, y)\big).$$

Les deux premières règles sont évidentes et les formules indiquées peuvent d'ailleurs respectivement se condenser en $\forall (x, y) \in E \times F\ ,\ P(x, y)$ et en $\exists (x, y) \in E \times F\ :\ P(x, y)$. La troisième règle mérite peut-être un moment de réflexion. Nous allons surtout évoquer l'implication réciproque, qui est source de nombreuses erreurs, parce qu'elle est fausse. Prenons le cas du prédicat $P(x, y)$ défini par $x \leqslant y$, avec $x \in E$ et $y \in F$, où E et F sont deux sous-ensembles d'un même ensemble ordonné G. Le membre gauche de l'implication se lit ainsi : « il existe un élément de E qui minore tous les éléments de F ». Le membre droit de l'implication se lit ainsi : « tout élément de F est minoré par un élément de E ». Il est clair que l'existence d'un $a \in E$ qui minore tous les éléments de F

entraîne que tous les éléments de F sont minorés par (au moins) un élément de E, à savoir l'élément a mentionné. Mais la réciproque est fausse : par exemple, si $E = F = \mathbb{R}$, tout élément de $\mathbb{R}$ est bien minoré par un élément de $\mathbb{R}$ (lui-même), mais aucun élément de $\mathbb{R}$ ne minore tous les éléments de $\mathbb{R}$.

7.2.3 Variables libres et variables muettes

Nous avons vu que, dans une formule telle que :

$$\forall x \in E \ , \ P(x, y, z, \ldots) \qquad \text{ou} \qquad \exists x \in E \ : \ P(x, y, z, \ldots),$$

la variable x est muette. Selon une autre terminologie, en usage chez les logiciens, on dit plutôt que que la variable x est *liée* par le quantificateur $\forall x$ ou $\exists x$, alors que les variables $y, z, \ldots$ sont *libres* : elles sont par exemple libres de « prendre des valeurs ». Comme on l'a dit, dans la proposition $\forall x \in E \ , \ P(x)$, la variable x peut être remplacée par une autre variable sans que le sens change, mais on ne peut pas lui affecter une valeur. Dans le cas où plusieurs variables sont présentes, les règles de remplacement d'une variable muette sont un peu plus compliquées.

Considérons par exemple le prédicat $P(x, y)$ défini pour $x, y \in \mathbb{R}$ par la formule $y \leqslant x^2$: les variables x et y sont libres.
Dans la formule $\forall x \in \mathbb{R} \ , \ P(x, y)$, c'est-à-dire $\forall x \in \mathbb{R} \ , \ y \leqslant x^2$, la variable y est libre, la variable x est liée. C'est donc un prédicat $Q(y)$. Il est équivalent à $y \leqslant 0$ (c'est immédiat). Puisque la lettre x est muette, on peut, par exemple, la remplacer par la lettre u, ce qui donne : $\forall u \in \mathbb{R} \ , \ y \leqslant u^2$. Il est clair que c'est exactement le même prédicat $Q(y)$. Cependant, si l'on remplaçait la lettre x par la lettre y, on obtiendrait la formule $\forall y \in \mathbb{R} \ , \ y \leqslant y^2$, dans laquelle il n'y a aucune lettre libre : c'est donc une proposition (d'ailleurs fausse).
Autre exemple : partons du prédicat $P(x, y)$ défini pour $x, y \in \mathbb{R}$ par la formule $y < x$. Le prédicat $\exists x \ : \ P(x, y)$ ne dépend que de y (et il prend la valeur « vrai » pour tout choix de y). Si l'on remplace x par u, le prédicat ne change pas. Mais si l'on remplace x par y, on obtient une proposition, d'ailleurs fausse. Ainsi, *lorsque l'on veut remplacer une lettre muette par une autre, il convient de choisir une lettre qui ne figure nulle part dans la formule*.

Outre les quantificateurs $\forall$ et $\exists$, il existe un certain nombre de symboles mathématiques qui ont la propriété de lier des variables. On les appelle *symboles mutificateurs*. Voici la liste des plus courants, avec, le cas échéant, une illustration des règles de remplacement.

1. Définition d'un ensemble en compréhension : l'ensemble $\{x \in \mathbb{R} \mid x > y\}$ (qui n'est défini que si y est connu par ailleurs) est le même que l'ensemble $\{u \in \mathbb{R} \mid u > y\}$, mais pas le même que l'ensemble $\{y \in \mathbb{R} \mid y > y\}$.
2. Définition d'une application $x \mapsto f(x)$: la fonction $x \mapsto x^2 + y$ (qui n'est définie que si y est connu par ailleurs) est la même que la fonction $u \mapsto u^2 + y$ mais pas la même que la fonction $y \mapsto y^2 + y$.
3. Le cas d'une famille indexée s'apparente aux deux précédents : on a par exemple, en général $(x_i)_{i \in I} = (x_j)_{j \in I}$; mais pas si la lettre j figure dans la définition de x_i : la suite géométrique $(p^n)_{n \in \mathbb{N}}$ est la même que la suite géométrique $(p^m)_{m \in \mathbb{N}}$, mais pas la même que la suite $(p^p)_{p \in \mathbb{N}}$.

4. Opérations portant sur des familles, par exemple intersection $\bigcap_{i\in I} E_i$, réunion $\bigcup_{i\in I} E_i$, ou produit cartésien $\prod_{i\in I} E_i$ d'une famille d'ensembles ; ou bien somme $\sum_{i\in I} x_i$ ou produit $\prod_{i\in I} x_i$ d'une famille de réels. La prudence s'impose particulièrement dans le cas d'expressions imbriquées, comme $\sum_{n\geqslant 1}\sum_{p\geqslant 1} n^{-p}$. L'apprentissage des règles de manipulation d'indices se fera par la pratique, en algèbre (sommes finies) et en analyse (sommes infinies).
5. Passages à la limite : $\lim_{x\to a} f(x) = \ell$, ou équivalents : $f(x) \underset{x\to a}{\sim} g(x)$.
6. Intégrales : $\displaystyle\int_a^x f(t)\,\mathrm{d}t = \int_a^x f(u)\,\mathrm{d}u$, le statut de l'expression $\displaystyle\int_a^x f(x)\,\mathrm{d}x$ étant ambigu !

7.3 Théorèmes et démonstrations

Ce qui délimite l'activité mathématique, ce n'est pas qu'elle aborde des objets mathématiques (nombres, figures ...), c'est le fait qu'elle repose sur des démonstrations de théorèmes. Le mathématicien part d'axiomes et de définitions. S'il est avancé dans l'étude, il dispose également de théorèmes déjà démontrés. Il obtient de nouveaux théorèmes par des démonstrations. Celles-ci sont des enchaînements de déductions qui obéissent à des règles logiques. Cette activité peut être formalisée au point de rendre possible la vérification mécanique des démonstrations (mais pas leur invention) - mais c'est très difficile au delà des théories les plus superficielles. Cas spécial : le calcul, qui est une forme de démonstration, est plus facilement mécanisable. C'est bien connu pour le calcul numérique, mais cela vaut également pour le calcul symbolique, comme on le verra au module II.8. Nous ne tenterons pas de formaliser l'activité mathématique, et les mathématiciens ne le font pas en général. En fait, il se produit des erreurs même chez les plus aguerris. Pour les éviter autant que possible, il faut du bon sens, de la méthode, de la rigueur, et le jugement des pairs (c'est-à-dire la relecture). Voici quelques exemples de méthodes.

1. Démonstration par l'absurde. Pour démontrer l'énoncé P, on fait l'hypothèse qu'il est faux et l'on en déduit une contradiction.

Exercice 11.

Démontrer par l'absurde l'énoncé : $\sqrt{2}$ est irrationnel.

Solution. Si $\sqrt{2}$ était rationnel, on pourrait l'écrire p/q avec $p, q \in \mathbb{N}^*$ premiers entre eux. On aurait donc $p^2 = 2q^2$, donc p pair : $p = 2n$, donc $q^2 = 2n^2$, donc q pair : $q = 2m$, mézalor p et q ne seraient pas premiers entre eux, ce qui contredirait l'hypothèse. Ainsi, on ne peut écrire $n = p/q$ et $\sqrt{2}$ est bien irrationnel.

---

2. Méthode de l'hypothèse auxiliaire. Cette méthode s'applique lorsque l'on veut démontrer une implication $P \Rightarrow Q$. On suppose temporairement P vraie (c'est « l'hypothèse auxiliaire »), et l'on procède à des déductions jusqu'à prouver que Q est vraie. Pour l'illustrer, nous la combinerons avec les deux méthodes suivantes.

3. Démonstration par disjonction de cas. Si l'on est certain que $P \vee Q$ est vrai, pour démontrer R, il suffit de démontrer $P \Rightarrow R$, puis de démontrer $Q \Rightarrow R$. Démontrons ainsi que le carré de tout réel est positif ou nul. Soit $x \in \mathbb{R}$. Il est lui-même positif ou nul, ou bien négatif ou nul. Dans le premier cas, $x^2 = xx$ est le produit de deux réels positifs ou nuls, donc il est positif ou nul. Dans le deuxième cas, $x^2 = (-x)(-x)$ est encore le produit de deux réels positifs ou nuls, donc positif ou nul.
4. Démonstration par contraposée. Cette méthode s'applique lorsque l'on veut démontrer une implication $P \Rightarrow Q$. On démontre à la place l'implication *contraposée* $(\neg Q) \Rightarrow (\neg P)$, qui lui est équivalente.
   Démontrons ainsi l'implication : $f \in \mathcal{C}(\mathbb{R}, \mathbb{R})$ est inversible $\Rightarrow$ f a un signe constant ; ici, $\mathcal{C}(\mathbb{R}, \mathbb{R})$ désigne l'ensemble des fonctions continues de $\mathbb{R}$ dans $\mathbb{R}$. L'hypothèse P signifie : $\exists g \in \mathcal{C}(\mathbb{R}, \mathbb{R}) \ : \ fg = 1$. La conclusion Q signifie :
   $$\big(\forall x \in \mathbb{R}\ ,\ f(x) > 0\big) \vee \big(\forall x \in \mathbb{R}\ ,\ f(x) < 0\big).$$
   Nous allons démontrer que $(\neg Q) \Rightarrow (\neg P)$. Pour cela, nous introduisons l'hypothèse auxiliaire $\neg Q$, autrement dit, le temps de la démonstration, nous supposons $\neg Q$ vraie. Expliciter la négation de Q est un petit exercice sur les lois de Morgan et les quantificateurs. On trouve : $\big(\exists x \in \mathbb{R} \ : \ f(x) \leqslant 0\big) \wedge \big(\exists x \in \mathbb{R} \ : \ f(x) \geqslant 0\big)$. Dire que « A et B » est vrai signifie que A est vrai et que B est vrai. Puisque $\exists x \in \mathbb{R} \ : \ f(x) \leqslant 0$ est vrai, on peut choisir $a \in \mathbb{R}$ tel que $f(a) \leqslant 0$. Puisque $\exists x \in \mathbb{R} \ : \ f(x) \geqslant 0$ est vrai, on peut choisir $b \in \mathbb{R}$ tel que $f(b) \geqslant 0$. Noter qu'en vertu du lien entre $\exists$ et $\wedge$ (voir la règle (5) de la page 52), on ne peut pas exiger que a et b soient les mêmes ! On a $a \leqslant b$ ou $b \leqslant a$ (encore une disjonction de cas !) Dans le premier cas, f s'annule en un point de $[a, b]$, dans le deuxième cas, en un point de $[b, a]$ (théorème des valeurs intermédiaires, module IV.2). Dans tous les cas, il existe c tel que $f(c) = 0$. Montrons alors (par l'absurde !) que f n'est pas inversible. Si elle avait un inverse g, on aurait $fg = 1$, donc $0 = 0g(c) = f(c)g(c) = 1$, ce qui est impossible. On a donc démontré $\neg P$, en ayant pour cela supposé $\neg Q$. On a donc démontré $\neg Q \Rightarrow \neg P$. On a donc démontré $P \Rightarrow Q$.
5. Démonstration par récurrence simple. Pour démontrer $\forall n \in \mathbb{N}\ ,\ P(n)$, on démontre $P(0)$ et, pour tout n, $P(n) \Rightarrow P(n+1)$.

Exercice 12.

Démontrer par récurrence $\forall n \in \mathbb{N}\ ,\ n < 2^n$.

Solution. C'est vrai pour $n := 0$ car $0 < 1$. Pour démontrer $P(n) \Rightarrow P(n+1)$, on fait l'*hypothèse de récurrence* $P(n) : n < 2^n$, « pour un certain n » : c'est donc encore un cas d'hypothèse auxiliaire. Supposons $n < 2^n$, on a $n + 1 < 2^n + 1$, et comme $1 \leqslant 2^n$, on en déduit $n + 1 < 2^n + 2^n = 2^{n+1}$, soit $P(n+1)$.
On a donc bien démontré l'implication $P(n) \Rightarrow P(n+1)$ (on dit parfois l'*hérédité*).
Le principe de récurrence (module II.1) permet de conclure.

---

6. Démonstration par récurrence à deux pas. Pour démontrer $\forall n \in \mathbb{N}\ ,\ P(n)$, on démontre $P(0)$ et $P(1)$, et, pour tout n, $\big(P(n) \wedge P(n+1)\big) \Rightarrow P(n+2)$. Il existe évidemment une version à k pas ! Démontrons ainsi que, pour tout $n \in \mathbb{N}$, $2^n - (-1)^n$

est multiple de 3. C'est vrai pour $n := 0$ et 1 (vérification immédiate). Supposons que $a := 2^n - (-1)^n$ et $b := 2^{n+1} - (-1)^{n+1}$ sont multiples de 3. On constate alors (petit calcul) que $2^{n+2} - (-1)^{n+2} = 2a + b$, qui est multiple de 3. En fait, ce mode de démonstration s'applique souvent à des suites définies par des récurrences à deux pas : ici, $u_{n+2} = u_{n+1} + 2u_n$. La précaution à ne pas perdre de vue est que *l'initialisation de la récurrence doit se faire en* $n := 0$ *et* $n := 1$*, ou bien* $n := 1$ *et* $n := 2$. Par exemple, la propriété « $2^n + (-1)^n$ est multiple de 3 » vérifie la même relation d'hérédité, et elle est vraie pour $n := 0$ mais fausse pour $n := 1$.

7. Démonstration par récurrence forte. On veut encore démontrer $\forall n \in \mathbb{N}\ ,\ P(n)$. Ici, l'hérédité prend la forme : $\big(\forall m < n\ ,\ P(m)\big) \Rightarrow P(n)$. En principe, il n'y a même pas besoin d'initialisation. Le lecteur se reportera à 6.1 pour des exemples ; voir aussi l'exercice I.1.38 de la page 60.

Remarque. Nous avons mentionné la démonstration par disjonction de cas. Il est possible de même de formuler des *définitions par disjonction de cas*. En voici un exemple typique, qui introduit d'ailleurs une notation d'intérêt très courant.

Exemple. La *notation de Kronecker*, sous sa forme la plus générale, est la suivante :

$$\delta_{a,b} := \begin{cases} 1 & \text{si } a = b, \\ 0 & \text{sinon.} \end{cases}$$

Les lettres a, b peuvent désigner des nombres (entiers ou non), mais également n'importe quel genre d'objets mathématiques.

Exercice type corrigé

I.1.0

1) Soit E un ensemble totalement ordonné et soit $(u_n)_{n\geqslant 0}$ une suite d'éléments de E. On suppose que la suite $(u_n)_{n\geqslant 0}$ n'admet pas de plus grand élément. Démontrer : $\forall k \in \mathbb{N}\ ,\ \exists \ell > k\ :\ u_k < u_\ell$. En déduire que, s'il existe un entier p tel que la suite $(u_n)_{n\geqslant p}$ n'admet pas de plus grand élément, alors la suite $(u_n)_{n\geqslant 0}$ admet une sous-suite strictement croissante.
2) On suppose maintenant au contraire que, quel que soit l'entier p, la suite $(u_n)_{n\geqslant p}$ admet un plus grand élément et l'on note v_p ce dernier. Démontrer que la suite $(v_p)_{p\geqslant 0}$ est décroissante. En déduire que la suite $(u_n)_{n\geqslant 0}$ admet une sous-suite décroissante.
3) Quelle conclusion générale peut-on tirer des deux questions précédentes au sujet des suites d'éléments de E ?
4) Selon le module IV.1, toute suite croissante majorée (resp. toute suite décroissante minorée) de réels admet une limite. Que peut-on en déduire pour les suites bornées de réels ?

Solution. 1) Notons $M := \max(u_0, \ldots, u_k)$. Il existe un entier ℓ tel que $M < u_\ell$ (sinon, M serait le plus grand élément de la suite). On a évidemment $\ell > k$ et $u_k < u_\ell$, ce qui démontre la première affirmation. Soit maintenant p un entier tel que la suite $(u_n)_{n\geqslant p}$ n'admet pas de plus grand élément. Posons $i_0 := p$. Appliquons la propriété qui vient d'être démontrée avec $k := i_0$ et $i_1 := \ell$: on

voit qu'il existe $i_1 > i_0$ tel que $u_{i_0} < u_{i_1}$. De même, en l'appliquant avec $k := i_1$ et $i_2 := \ell$, on voit qu'il existe $i_2 > i_1$ tel que $u_{i_1} < u_{i_2}$. En itérant ce procédé, on construit une suite strictement croissante d'indices $i_0 < i_1 < i_2 < \cdots$ telle que $u_{i_0} < u_{i_1} < u_{i_2} < \cdots$. La suite $(u_{i_k})_{k\geqslant 0}$ est une sous-suite strictement croissante de la suite $(u_n)_{n\geqslant 0}$.
2) De manière générale, si $A \subset B$ et si A et B ont chacun un plus grand élément, $\max A \in A$ et $\max B \in B$, la relation $\max A \in B$ implique que $\max A \leqslant \max B$. Puisque $\{u_n \mid n \geqslant p+1\} \subset \{u_n \mid n \geqslant p\}$, on a donc $v_{p+1} \leqslant v_p$ et la suite $(v_p)_{p\geqslant 0}$ est bien décroissante. Puisque $v_p \in \{u_n \mid n \geqslant p\}$, on peut l'écrire $v_p = u_{n_p}$ avec $n_p \geqslant p$. La suite des n_p n'est donc pas majorée, et l'on peut donc en extraire une sous-suite strictement croissante $(n_{p_k})_{k\geqslant 0}$. Posons $m_k = n_{p_k}$. Alors la suite $(u_{m_k})_{k\geqslant 0}$ est une sous-suite décroissante de la suite $(u_n)_{n\geqslant 0}$.
3) Toute suite d'éléments d'un ensemble totalement ordonné E admet donc une sous-suite croissante ou une sous-suite décroissante.
4) Comme toute sous-suite d'une suite bornée est elle-même bornée, on voit que toute suite bornée de réels admet une sous-suite croissante majorée ou décroissante minorée, donc convergente : c'est le théorème de Bolzano-Weierstraß (module IV.1, théorème 47).

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

I.1.1 $*$ Démontrer que $\{\{a\},\{a,b\}\} = \{\{a'\},\{a',b'\}\}$ si, et seulement si, $a = a'$ et $b = b'$.
Cela peut donc servir à construire les couples.

I.1.2 Que dire de $\mathcal{P}(A \cap B)$? de $\mathcal{P}(A \cup B)$ lorsque A et B sont disjoints ?

I.1.3 $*$ 1) On définit par récurrence $E_0 := \varnothing$, puis, pour tout entier k, $E_{k+1} := E_k \cup \{E_k\}$. Démontrer par récurrence sur k que tout élément de E_k en est un sous-ensemble.
2) Démontrer par récurrence sur k que $E_{k+1} \neq E_k$; en déduire que les E_k forment une suite strictement croissante d'ensembles (pour l'inclusion). Combien l'ensemble E_k a-t-il d'éléments ?

I.1.4 $**$ 1) Dans certaines variantes de la théorie des ensembles, on introduit l'*axiome de fondation* :

$$\forall x \neq \varnothing\ ,\ \exists y \in x\ :\ y \cap x = \varnothing.$$

Démontrer que cet axiome équivaut à l'affirmation suivante : il n'existe pas de suite d'ensembles $x_0, x_1, \ldots$ telle que $\forall i \in \mathbb{N}\ ,\ x_{i+1} \in x_i$.
Pour l'implication directe, procéder par l'absurde et considérer $x = \{x_i \mid i \in \mathbb{N}\}$.
2) Démontrer que l'axiome de fondation entraîne qu'il n'existe aucun ensemble qui soit élément de lui-même et donc qu'il n'existe pas d'ensemble de tous les ensembles.
3) On pose $s(x) = x \cup \{x\}$ (« successeur » de x).
Sous la même hypothèse, montrer que $x, s(x), s\big(s(x)\big), \ldots$ sont deux à deux distincts.
4) La relation $(\forall y \in x\ ,\ y \in y)$ est-elle collectivisante ? Et la relation $(\forall y \in x\ ,\ y \notin y)$?

Exercices sur la section 2

I.1.5 Soient $E = \{a,b\}$ et $F = \{c,d,e\}$. On suppose $a \neq b$.
Existe-t-il des surjections, des injections, des bijections de E dans F ?

I.1.6 Soient $f : E \to F$ et $B \subset F$.

Montrer que $B = f\big(f^{-1}(B)\big)$ si, et seulement si, $B \subset \operatorname{Im} f$.

I.1.7 On dit que $A \subset E$ est saturé par $f : E \to F$ si $A = f^{-1}\big(f(A)\big)$, autrement dit, si :

$$\forall x \in A\ ,\ \forall x' \in E\ :\ f(x) = f(x') \Rightarrow x' \in A.$$

Montrer que les parties saturées de E sont les parties de la forme $f^{-1}(B)$.

I.1.8 Que peut-on déduire de l'égalité $g \circ f = \mathrm{Id}_E$?

I.1.9 (i) Soient $f : E \to F$ et $f' : E' \to F$ deux applications.
Démontrer l'équivalence suivante : $\big(\exists u : E \to E'\ :\ f = f' \circ u\big) \Longleftrightarrow \big(\operatorname{Im} f \subset \operatorname{Im} f'\big)$.
(ii) Soient $f : E \to F$ et $f' : E \to F'$ deux applications. On suppose F' non vide.
Démontrer l'équivalence :

$$\big(\exists v : F \to F'\ :\ f' = v \circ f\big) \Longleftrightarrow \big(\forall x_1, x_2 \in E\ :\ f(x_1) = f(x_2) \Rightarrow f'(x_1) = f'(x_2)\big).$$

Pourquoi faut-il supposer F' non vide ?

Exercices sur la section 3

I.1.10 Soit $(E_i)_{i\in I}$ une famille de sous-ensembles de E.
À quelle condition l'application $f \mapsto (f_{|E_i})_{i\in I}$ (exemple 4 de la page 12) de $\mathcal{F}(E,F)$ dans $\prod\limits_{i\in I} \mathcal{F}(E_i, F)$ est-elle injective ? surjective ? En général, quelle est son image ?

Exercices sur la section 4

I.1.11 Quels éléments de $(\mathcal{P}(E), \cup)$, resp. de $(\mathcal{P}(E), \cap)$ sont neutres, inversibles, simplifiables, absorbants, idempotents ?

I.1.12 Montrer que l'image d'un idempotent par tout morphisme est un idempotent, mais que ce n'est pas vrai pour un neutre, un simplifiable, un inversible, un absorbant.

I.1.13 $*$ Montrer que les nombres de Catalan (exemple de la page 26) vérifient la relation de récurrence : $c_n = \sum\limits_{i+j=n-1} c_i c_j$. Calculer c_3, c_4 et c_5.

I.1.14 $*$ Montrer comment l'on peut traduire l'associativité d'une loi sur E à l'aide d'un diagramme commutatif qui met en jeu les ensembles E, $E \times E$ et $E \times E \times E$. Traiter de manière similaire la commutativité.

Exercices sur la section 5

I.1.15 La relation $\complement_E A \subset B$ entre sous-ensembles de E est symétrique. De même pour la relation $B \subset \complement_E A$.

I.1.16 Soient $n_1, \ldots, n_k$ des entiers non nuls et $\mathcal{R}_i$ la relation de congruence $x \equiv y \pmod{n_i}$ dans $\mathbb{Z}$. Calculer l'intersection des $\mathcal{R}_i$.

I.1.17 Soient V un espace vectoriel et W un sous–espace vectoriel de V. Montrer que tout supplémentaire de W est un ensemble de représentants pour la relation d'équivalence $x \mathcal{R} y$ définie par $x - y \in W$.

I.1.18 Sur l'ensemble $E^{\mathbb{N}}$, on définit la relation $\sim$ par :

$$(u_n)_{n\in\mathbb{N}} \sim (v_n)_{n\in\mathbb{N}} \Longleftrightarrow \exists p \in \mathbb{N} \ : \ \forall n \geqslant p \ , \ u_n = v_n .$$

Démontrer que c'est une relation d'équivalence.

I.1.19 $*$ 1) On munit $\{0,1\}^{\mathbb{N}}$ de l'ordre lexicographique, pour lequel $(u_n)_{n\in\mathbb{N}} \leqslant (v_n)_{n\in\mathbb{N}}$ si les deux suites sont égales ou s'il existe $p \in \mathbb{N}$ tel que $u_p < v_p$ et $\forall n < p \ , \ u_n = v_n$. Démontrer que c'est un ordre total.
2) Décrire l'unique relation d'ordre sur $\mathcal{P}(\mathbb{N})$ telle que la bijection $\mathcal{P}(\mathbb{N}) \to \{0,1\}^{\mathbb{N}}$ (théorème 2 de la page 24) est un isomorphisme.
3) Démontrer que l'application $A \mapsto \sum\limits_{a\in A} 2^a$ est une bijection de l'ensemble $\mathcal{P}_f(\mathbb{N})$ des parties finies de $\mathbb{N}$ sur $\mathbb{N}$. (Utiliser l'écriture binaire, module II.1.)
4) Décrire l'unique relation d'ordre $\leqslant$ sur $\mathcal{P}_f(\mathbb{N})$ telle que cette application soit un isomorphisme de $(\mathcal{P}_f(\mathbb{N}), \leqslant)$ sur $(\mathbb{N}, \leqslant)$.

I.1.20 La clôture réflexive transitive d'une relation symétrique $\mathcal{R}$ est une relation d'équivalence (c'est donc la relation d'équivalence engendrée par $\mathcal{R}$).

I.1.21 On appelle cycle pour une relation $\mathcal{R}$ sur E un n-uplet $(x_1, \ldots, x_n) \in E^n$, avec $n \geqslant 2$, les x_i deux à deux distincts et $x_1 \mathcal{R} x_2$, $\ldots$, $x_{n-1} \mathcal{R} x_n$, $x_n \mathcal{R} x_1$. Démontrer que, pour qu'une relation quelconque engendre une relation d'ordre, il faut, et il suffit, qu'elle n'admette pas de cycle ; et la relation d'ordre engendrée est alors sa clôture réflexive transitive.

I.1.22 $*$ Démontrer que la relation $x \in y$ est une relation d'ordre strict sur chaque E_k (exercice I.1.3 de la page 57) ainsi que sur $\bigcup\limits_{k\in\mathbb{N}} E_k$.

Exercices sur la section 6

I.1.23 $*$ Soit $(E_i)_{i\in I}$ une famille d'ensembles ordonnés indexée par un ensemble bien ordonné I. Sur l'ensemble produit $E = \prod\limits_{i\in I} E_i$, on définit la relation $(x_i)_{i\in I} \preceq (y_i)_{i\in I}$ par : $(x_i)_{i\in I} = (y_i)_{i\in I}$, ou bien, i étant le plus petit indice tel que $x_i \neq y_i$, on a $x_i < y_i$. Vérifier que c'est une relation d'ordre. Démontrer que, si les E_i sont totalement ordonnés , leur produit l'est également. Si I est fini et si les E_i sont bien ordonnés, leur produit l'est également.

I.1.24 Soit E un ensemble quelconque. Montrer que l'ensemble des parties finies de E, ordonné par inclusion, est artinien.

I.1.25 Soient E un ensemble ordonné et $f : E \to \mathbb{N}$ une application strictement croissante. Démontrer que E est artinien. Donner un exemple où il n'est pas bien ordonné.

I.1.26 Démontrer à l'aide des résultats sur les cardinaux qu'il n'existe pas d'ensemble de tous les ensembles. (Comparer un tel ensemble à l'ensemble de ses parties.)

I.1.27 Soient $\aleph_1, \aleph_2$ et $\aleph_3$ trois cardinaux. Démontrer les formules :

$$\left(\aleph_1\aleph_2\right)^{\aleph_3} = \aleph_1^{\aleph_3}\aleph_2^{\aleph_3}, \quad \aleph_1^{\aleph_2+\aleph_3} = \aleph_1^{\aleph_2}\aleph_1^{\aleph_3}, \quad \left(\aleph_1^{\aleph_2}\right)^{\aleph_3} = \aleph_1^{\aleph_2\aleph_3}.$$

I.1.28 Démontrer que les opérations sur les cardinaux sont croissantes en chacun de leurs arguments (il y aura cependant une exception pour l'opération d'exponentiation).

I.1.29 (i) Démontrer que chacun des ensembles E_k de l'exercice I.1.3 de la page 57 est fini mais que l'ensemble dont les éléments sont les E_k ne l'est pas.
(ii) Démontrer sans faire appel à l'axiome de l'infini que, si E est infini il existe des applications injectives $f_k : E_k \to E$ qui sont restrictions les unes des autres. En déduire l'existence de $\mathbb{N}$ et d'une application injective de $\mathbb{N}$ dans E.

I.1.30 Soit k un entier naturel non nul. Pour tout $n \in \mathbb{N}$, on note $n \operatorname{div} k$ et $n \mod k$ le quotient et le reste de la division euclidienne de n par k.
Démontrer que l'application $n \mapsto (n \operatorname{div} k, n \mod k)$ est une bijection de $\mathbb{N}$ sur $\mathbb{N}\times\{0,\ldots,k-1\}$. En déduire que $k\aleph_0 = \aleph_0$.

I.1.31 Démontrer que l'ensemble $\mathbb{N}^{(\mathbb{N})}$ des suites d'entiers à support fini (c'est-à-dire nulles à partir d'un certain rang) est dénombrable.

I.1.32 Pour tout cardinal infini $\aleph$, démontrer les égalités $\aleph + \aleph = \aleph\aleph_0 = \aleph$. En déduire qu'aucun cardinal infini n'est simplifiable pour l'addition ou la multiplication des cardinaux. Que dire des cardinaux finis ?

I.1.33 Démontrer que $\mathbb{N}^{\mathbb{N}}$ et $\mathbb{R}^{\mathbb{N}}$ ont la puissance du continu.

Exercices sur la section 7

I.1.34 On pose $P \mid Q = \neg(P \wedge Q)$ (*connecteur de Sheffer*, également appelé NAND, comme « Not And »). Dresser sa table de vérité. Montrer que tous les connecteurs peuvent s'en déduire. Voyez-vous un autre connecteur ayant cette propriété ?

I.1.35 Vérifier que la formule $\big(P \wedge (P \Rightarrow Q)\big) \Rightarrow Q$ est une tautologie et que $P \Leftrightarrow (\neg P)$ est une contradiction. Que dire de la formule $P \Rightarrow (\neg P)$?

I.1.36 Soient F et G deux formules. Démontrer qu'elles sont logiquement équivalentes si, et seulement si, la formule $F \Leftrightarrow G$ est une tautologie.

I.1.37 Démontrer par l'absurde que, pour toute suite de réels, si $\lim\limits_{n\to+\infty} u_n = 1$, alors :

$$\exists n \in \mathbb{N} \ : \ \forall p \geqslant n \ , \ u_p > 0.$$

On utilisera la méthode de l'hypothèse auxiliaire.

I.1.38 Démontrer par récurrence forte que les coefficients binomiaux sont entiers, en utilisant la relation de Pascal (relation (4) de la page 79).

Partie II

Algèbre

L'algèbre au sens moderne a deux visages : d'une part l'étude des structures indépendamment de leurs réalisations concrètes et d'autre part celle d'algorithmes performants permettant des calculs effectifs sur ordinateur.
Elle est issue de l'arithmétique, déjà bien présente chez les anciens Grecs (notamment Euclide), qui a connu de grands développements au XVIIe siècle (Fermat), au XVIIIe (Euler) et au XIXe (Gauß, Hermite), et joue toujours un rôle très important, central dans un certain nombre de conjectures très célèbres comme celle de Riemann. Vers le milieu du XXe siècle, l'algèbre a permis d'unifier arithmétique et géométrie algébrique, et d'établir aujourd'hui de fascinantes convergences entre arithmétique et physique théorique. Algèbre et arithmétique ont aussi d'importantes applications (cryptographie, codes correcteurs d'erreurs, secret bancaire, communications).
Les « recettes » pour les équations du premier et du second degré remontent à une lointaine antiquité : à Babylone (vers 1800 avant Jésus-Christ), en Grèce grâce à Euclide et Diophante, suivis au Moyen Âge par des mathématiciens de langue arabe. Il faudra attendre le XVIe siècle italien pour connaître les formules de résolution par radicaux des équations de degré trois et quatre (Del Ferro, Cardan, Ferrari). Un siècle plus tard, Viète et Descartes essaieront de dépasser ce stade, ce qui conduira ce dernier à inventer au passage la géométrie analytique. Au début du XIXe, Gauß, Abel puis surtout Galois mettront un terme aux recherches vaines de leurs prédécesseurs : « il n'y a plus de formule » à partir du degré cinq. Après avoir joué un très grand rôle dans la construction des mathématiques, le sujet perdait presque tout intérêt ; toutefois les outils inventés pour le clore (groupes, théorie de Galois) restaient centraux dans les mathématiques et certaines de leurs applications.
Dans une perspective plus appliquée, Gauß, au XIXe siècle, a développé des techniques « effectives » d'algèbre linéaire pour l'astronomie (trajectoire de Cérès en 1801). Au XXe siècle, les groupes et l'algèbre linéaire interviennent de manière centrale en mécanique quantique et en chimie (cristallographie, chimie quantique). L'algèbre linéaire est aujourd'hui également omniprésente dans de nombreux problèmes industriels (contrôle, optimisation, robotique...).

Nous avons évoqué plus haut les structures algébriques. Dans ce cours de première année apparaissent celles de groupe, anneau, corps, espace vectoriel. Dans une première approche, il faut apprendre à maîtriser les corps usuels (rationnels, réels, complexes), la notion de groupe abélien et surtout l'algèbre linéaire sous son double aspect théorique et algorithmique, sans oublier ses relations filiales avec la géométrie.

Arithmétique II.1

L'arithmétique est la science des nombres, plus précisément des nombres entiers, qu'ils soient *naturels* : $0, 1, 2, \ldots$ ou *relatifs* : $-2, -1, 0, 1, 2, \ldots$. Il s'agit donc d'étudier l'ensemble $\mathbb{N}$ des entiers naturels et l'ensemble $\mathbb{Z}$ des entiers relatifs. Les outils essentiels sont les quatre opérations : addition, multiplication, soustraction, division, auxquelles il faut ajouter la relation d'ordre $\leqslant$. L'objectif poursuivi dans cet ouvrage n'est pas d'étudier $\mathbb{N}$ et $\mathbb{Z}$ « pour eux-mêmes », mais en liaison avec les nombreux domaines des mathématiques (ou de leurs applications) où ils interviennent. Citons quelques raisons de s'intéresser aux entiers.

1. Les entiers naturels servent d'abord à compter (c'est leur fonction *cardinale*), ou à numéroter des objets (c'est leur fonction *ordinale*). Chacun est familiarisé avec ces deux fonctions depuis son enfance.
2. Dans toutes les branches des mathématiques, l'un des modes de raisonnement fondamentaux est le *raisonnement par récurrence*, dans lequel l'intervention de $\mathbb{N}$ est évidente. À cet égard, l'arithmétique elle-même a le mérite d'offrir de nombreux exemples de tels raisonnements, qu'ils soient très simples ou plus élaborés.
3. Les ordinateurs manient des nombres entiers (écrits en système binaire). Ils ne font même que cela, mais pas n'importe comment : ils sont « programmés ». Lorsqu'ils effectuent un programme, ils passent d'une instruction n à l'instruction suivante $n+1$, démarche proche de celle que l'on suit dans un raisonnement par récurrence. Dans les programmes, il est fait appel à des *algorithmes*, c'est-à-dire des procédés de calcul, et ces algorithmes portent souvent sur des entiers.

Il peut être important de comprendre ou de savoir se servir de ces algorithmes et, si nécessaire, de mettre au point un algorithme répondant à tel ou tel besoin. À l'inverse, savoir écrire un programme (fonctionnant correctement) qui résout un certain problème mathématique prouve au moins que l'on a compris les notions intervenant dans ce problème.

4. En ce qui concerne $\mathbb{Z}$, son étude a ici une motivation directe. D'abord chacun pourra constater que l'*algèbre* (notamment l'algèbre linéaire) occupe une place importante dans l'enseignement des mathématiques. Il s'agit d'un outil puissant, utilisé dans beaucoup de domaines des mathématiques et de leurs applications. L'algèbre met en jeu des objets fondamentaux, tels que les *groupes* ou les *anneaux* (voir le module II.2). Dans chacun de ces cas, $\mathbb{Z}$ offre l'exemple le plus simple du type d'objet envisagé. Par exemple, l'*arithmétique des polynômes* s'avérera très voisine de l'*arithmétique de* $\mathbb{Z}$. L'étude de $\mathbb{Z}$ servira donc de « répétition générale » avant celle des polynômes, qui sont des objets *a priori* moins simples que les entiers.

1 Ensemble des entiers naturels

Cette section utilise librement le module précédent concernant les ensembles. Les notions d'application injective (surjective, bijective) et de relation d'ordre sont supposées connues. À propos de la présentation de $\mathbb{N}$ qui est donnée ici, mentionnons les points suivants :

1. Une approche entièrement « intuitive » ne peut pas suffire. Chacun sait « calculer » sur les entiers, mais il est nécessaire de savoir bâtir un raisonnement correct dans une situation où ces entiers interviennent.
2. Pour autant, nous ne donnons pas une « construction rigoureuse » de $\mathbb{N}$, contrairement à ce qui sera fait plus loin pour les ensembles de nombres $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$. Il est possible de donner une telle construction, mais cela nous a semblé trop lourd dans un ouvrage de ce type.
3. En conséquence, nous *admettrons* qu'il existe un ensemble, que nous appellerons $\mathbb{N}$, possédant trois propriétés simples (voir la section 1.1). Ces propriétés seront prises comme *axiomes* ou règles du jeu. L'important est de comprendre comment l'on *déduit* de ces axiomes les propriétés des entiers qui seront utiles. Bien entendu, les axiomes ne sont pas arbitraires. C'est l'expérience, comme dans d'autres sciences, qui a montré que la théorie obtenue en partant de ces axiomes « fonctionne bien », tant en mathématiques que dans leurs applications.

1.1 Relations d'ordre et entiers naturels

Les relations d'ordre ont déjà été étudiées dans le module I.1 (section 5.3), le lecteur est prié de s'y reporter. Considérons donc un *ensemble ordonné* $\mathcal{N}$, c'est-à-dire un ensemble muni d'une relation d'ordre, notée $\leqslant$. Cette relation binaire entre éléments de $\mathcal{N}$ est donc *réflexive, antisymétrique et transitive* ; on en déduit les relations notées $<$, $\geqslant$ et $>$. Soient A une partie de $\mathcal{N}$ et $t \in \mathcal{N}$. Nous savons ce que signifient les expressions « t est un majorant (ou un minorant) de A », « t est le plus petit élément de A », ce que l'on note $t = \min(A)$, « t est le plus grand élément de A », ce que l'on note $t = \max(A)$. Nous savons aussi ce que veut dire « A est majorée, ou minorée, ou bornée ».

À partir de la relation $\leqslant$ sur $\mathcal{N}$, nous pouvons parler d'*intervalles*. Soient a, b deux éléments de $\mathcal{N}$ donnés. On notera $[a, b]$ l'ensemble des éléments x de $\mathcal{N}$ tels que $a \leqslant x \leqslant b$ et $]a, b[$ l'ensemble des éléments x de $\mathcal{N}$ tels que $a < x < b$. On a des définitions analogues pour $]a, b]$ et $[a, b[$. L'ensemble des éléments x de $\mathcal{N}$ tels que $x \geqslant a$ (resp. $x > a$) sera noté $[a, \rightarrow[$ (resp. $]a, \rightarrow[$). Les définitions de $]\leftarrow, a]$ et $]\leftarrow, a[$ sont analogues. Ces divers ensembles, appelés *intervalles*, peuvent éventuellement être vides.

Cela étant, nous *admettrons l'existence* d'un ensemble ordonné non vide $\mathbb{N}$ vérifiant les trois propriétés suivantes :

(N1) Toute partie non vide de $\mathbb{N}$ possède un plus petit élément.

(N2) Toute partie non vide et *majorée* de $\mathbb{N}$ possède un plus grand élément.

(N3) L'ensemble $\mathbb{N}$ lui-même n'est pas majoré, en particulier il ne possède pas de plus grand élément.

Les éléments de $\mathbb{N}$ sont appelés *entiers naturels*.
Une question se pose ici. Si deux ensembles ordonnés vérifient les conditions ci-dessus, vont-ils conduire à des théories équivalentes ? La réponse est oui, comme l'exercice II.1.1 le montrera. En pratique, le lecteur peut *admettre* que les entiers naturels, qu'il sait manipuler depuis son enfance, forment un ensemble $\mathbb{N}$ qui vérifie les conditions (N1), (N2), (N3).

1.2 Récurrence

Dans cette section, nous allons donner d'importantes conséquences des propriétés (N1), (N2), (N3). Nous parlerons d'« entiers » au lieu d'« entiers naturels », tant qu'il n'y aura pas de risque de confusion. Une première conséquence de (N1) est que $\mathbb{N}$ est *totalement ordonné*, c'est-à-dire que deux éléments x, y quelconques de $\mathbb{N}$ sont toujours *comparables* : on a soit $x \leqslant y$, soit $y \leqslant x$ (soit les deux). C'est évident si $x = y$. Sinon, $\{x, y\}$ étant une partie non vide de $\mathbb{N}$, elle possède un plus petit élément, égal soit à x soit à y, et l'on a $x \leqslant y$ (resp. $y \leqslant x$) dans le premier (resp. le deuxième) cas. Ensuite, l'ensemble $\mathbb{N}$ lui-même n'étant pas vide, la propriété (N1) montre qu'il possède un plus petit élément, on le note 0 ; les entiers autres que 0 sont dits *strictement positifs*, ils forment une partie de $\mathbb{N}$ notée $\mathbb{N}^*$.
Soit $n \in \mathbb{N}$. L'ensemble $A := \{k \in \mathbb{N} \mid k > n\}$ n'est pas vide, car n ne majore pas $\mathbb{N}$, vu la propriété (N3). Ainsi A possède un plus petit élément, noté $n + 1$ et appelé *successeur* de n. Le successeur de 0 est noté 1.
De même, si $n \in \mathbb{N}^*$, $B := \{k \in \mathbb{N} \mid k < n\}$ n'est pas vide, car $0 \in B$. Évidemment, n est un majorant de B ; la propriété (N2) montre alors que B possède un plus grand élément, noté $n-1$ et appelé *prédécesseur* de n. Provisoirement, la notation $n+1$ n'est pas reliée à l'addition, idem pour $n-1$ et la soustraction. Cependant on a $(n-1)+1 = n = (n+1)-1$ pour tout entier $n > 0$: le vérifier en exercice, de même que l'égalité $[0, n[= [0, n-1]$.
Pour deux entiers m, n les équivalences ci-dessous résultent des définitions :

$$(m > n) \Longleftrightarrow (m \geqslant n + 1)\ ,\ \ (m < n) \Longleftrightarrow \big((n > 0) \text{ et } (m \leqslant n - 1)\big).$$

Voici maintenant un résultat essentiel, il fournit la base logique du mode de raisonnement par récurrence. Le raisonnement par récurrence et le raisonnement par l'absurde sont deux outils très importants en mathématiques.

Théorème 1 (Théorème de récurrence).
Soit n_0 un entier. Pour tout entier $n \geqslant n_0$, notons $\mathcal{P}(n)$ une certaine propriété de l'entier n. On fait les hypothèses suivantes :

(R1) La propriété $\mathcal{P}(n_0)$ est vraie.

(R2) Si la propriété $\mathcal{P}(n)$ est vraie pour un certain entier $n \geqslant n_0$, la propriété $\mathcal{P}(n+1)$ l'est aussi.

Sous ces hypothèses, la propriété $\mathcal{P}(n)$ est vraie pour *tout* entier $n \geqslant n_0$.

Démonstration. Raisonnons par l'absurde : supposons qu'il existe au moins un entier $m \geqslant n_0$ tel que $\mathcal{P}(m)$ soit fausse. L'ensemble A des entiers $k \geqslant n_0$ tels que $\mathcal{P}(k)$ soit fausse étant alors non vide, il possède un plus petit élément n. D'après l'hypothèse (R1) $n \neq n_0$, donc $n > n_0$ et par suite $n-1 \geqslant n_0$ (définition de $n-1$). Ainsi $n_0 \leqslant n-1 < n$, donc $n-1 \notin A$ $\big($car $n = \min(A)\big)$, *i.e.* $\mathcal{P}(n-1)$ est vraie. Comme n est le successeur de $n-1$, (R2) fait que $\mathcal{P}(n)$ est vraie, c'est absurde car $n \in A$. Il n'existe donc pas d'entier $m \geqslant n_0$ tel que $\mathcal{P}(m)$ soit fausse. Le théorème en résulte. ■

Dans cette preuve, les hypothèses (R1) et (R2) peuvent être remplacées par l'hypothèse (R3) ci-dessous, sans changer la conclusion du théorème (vérifiez-le) ; c'est ce qu'on appelle la *récurrence forte*. Il est parfois plus facile de vérifier (R3) que (R1) et (R2), *cf.* par exemple le théorème 33 de la page 83.
(R3) Dès qu'un entier $n \geqslant n_0$ est tel que $\mathcal{P}(k)$ soit vraie pour tous les entiers k vérifiant $n_0 \leqslant k < n$, la propriété $\mathcal{P}(n)$ est vraie.
Si $\mathcal{P}(n)$ n'est définie que sur un intervalle $[n_0, n_1]$, on modifie (R2) ainsi : si $\mathcal{P}(n)$ est vraie pour un certain entier $n \in [n_0, n_1-1]$, $\mathcal{P}(n+1)$ est aussi vraie. En gardant (R1), on obtient la conclusion suivante : $\mathcal{P}(n)$ est vraie pour tout $n \in [n_0, n_1]$ (*récurrence finie*). Voici une conséquence utile du théorème de récurrence. Soit A une partie de $\mathbb{N}$ contenant 0. On suppose que le successeur de tout élément de A appartient aussi à A. Alors $A = \mathbb{N}$ (prendre pour $\mathcal{P}(n)$ la relation $n \in A$).
Un autre aspect de la récurrence est la *construction par récurrence*. Soient E un ensemble, g une application de E dans E et a un élément de E donné. Il existe une unique suite (u_n) d'éléments de E telle que $u_0 = a$ et $u_{n+1} = g(u_n)$ pour tout entier n. On dit que l'on a construit cette suite à partir de son premier terme $u_0 := a$ et de la relation de récurrence $u_{n+1} := g(u_n)$. Par exemple la suite (2^n) des puissances de 2 est définie par son premier terme 1 et la relation de récurrence $2^{n+1} := 2 \times 2^n$. De même la suite $(n!)$ des factorielles est définie par son premier terme $0! := 1$ et la relation de récurrence $(n+1)! := (n+1)n!$. Rappelons qu'une suite d'éléments de E est simplement une application de $\mathbb{N}$ dans E. La justification de la construction par récurrence va résulter du théorème qui suit.

Théorème 2. Soient E un ensemble, g une application de E dans E et a un élément de E donné. Il existe une unique application f de $\mathbb{N}$ dans E telle que $f(0) = a$ et $f(n+1) = g\big(f(n)\big)$ pour tout entier n.

Démonstration. Si f existe, sa restriction f_n à $[0,n]$ est telle que :

$$f_n(0) = a \qquad \text{et} \qquad f_n(m+1) = g\big(f_n(m)\big) \quad \text{pour tout} \quad m < n.$$

Montrons inversement, par récurrence, qu'une application telle que f_n existe et est unique. C'est vrai pour $n = 0$. Si c'est vrai pour n, il est nécessaire que f_n soit la restriction de f_{n+1} à $[0,n]$ et que

$$f_{n+1}(n+1) = g\big(f_{n+1}(n)\big) = g\big(f_n(n)\big).$$

Inversement ces deux relations impliquent que f_{n+1} convient.
Par suite, si f existe, elle est unique, car définie pour tout n par $f(n) := f_n(n)$. Cette application convient bien puisque $f(0) = f_0(0) = a$ et, pour tout n :

$$f(n+1) = f_{n+1}(n+1) = g\big(f_n(n)\big) = g\big(f(n)\big).$$

■

Exemples.

1. Soient X un ensemble, $b \in X$ et h une application de $\mathbb{N} \times X$ dans X. On peut de même définir une suite (u_n) par son premier terme $u_0 := b$ et la relation de récurrence $u_{n+1} := h(n, u_n)$. On vérifiera en exercice qu'il suffit d'appliquer le théorème 2 de la page précédente en prenant pour E l'ensemble $\mathbb{N} \times X$, pour g l'application $(n,x) \mapsto \big(n+1, h(n,x)\big)$ de E dans E, et pour a le couple $(0,b)$. L'application f est alors de la forme $n \mapsto (n, u_n)$, et la suite (u_n) répond à la question.
   Prenons par exemple $X := \mathbb{N}$, $b := 1$, et définissons $h : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ par la formule $h(k,x) := (k+1)x$. La suite (u_n) est alors la suite $(n!)$ des factorielles, dont la définition est ainsi justifiée. Nous avons utilisé ici la multiplication des entiers, dont la définition sera donnée dans la section suivante, le lecteur constatera qu'il n'y a pas de cercle vicieux.
2. Soient X un ensemble et $f : X \to X$ une application. On définit la suite (f_n) des itérées de f par son premier terme, qui est l'application identique Id_X de X, et la relation de récurrence $f_{n+1} := f \circ f_n$. Si cela ne prête pas à confusion, f_n sera notée f^n (ainsi $f^0 = \mathrm{Id}_X, f^1 = f$, $f^2 = f \circ f, \dots$). À cause de l'associativité de la composition des applications (*cf.* le module I.1), on a (voir aussi l'exercice 1 de la page 67) :
   $$f^3 = f \circ f^2 = f \circ (f \circ f) = (f \circ f) \circ f = f^2 \circ f.$$

Voici un théorème qui nous servira dans la section suivante.

Théorème 3. Soient X, Y deux ensembles et $u : X \to Y$, $v : Y \to Y$ deux applications. Il existe une unique application F de $X \times \mathbb{N}$ dans Y vérifiant les deux conditions suivantes :

1. Pour tout élément x de X, $F(x,0) = u(x)$.
2. Pour tout $x \in X$ et tout entier n, $F(x, n+1) = v\big(F(x,n)\big)$.

Démonstration. Définissons $F : X \times \mathbb{N} \to Y$ ainsi :

$$F(x,n) := (v^n \circ u)(x) \quad \text{pour tous} \quad x \in X,\ n \in \mathbb{N}.$$

Alors F vérifie la condition 1, car $v^0 \circ u = u$. Si $n \in \mathbb{N}$ et $x \in X$, on a $v^{n+1} \circ u = (v \circ v^n) \circ u = v \circ (v^n \circ u)$ (associativité de la composition des applications), et l'on en déduit que F vérifie la condition 2 :

$$F(x, n+1) = \big(v \circ (v^n \circ u)\big)(x) = v\big((v^n \circ u)(x)\big) = v\big(F(x,n)\big).$$

Soit $F' : X \times \mathbb{N} \to Y$ une autre application vérifiant les conditions 1 et 2. Notons $\mathcal{P}(n)$ la propriété suivante d'un entier n : pour tout $x \in X$, $F(x,n) = F'(x,n)$. Vu la condition 1 (pour F et F'), $\mathcal{P}(0)$ est vraie. Ensuite $\mathcal{P}(n) \implies \mathcal{P}(n+1)$ pour tout n, grâce à la condition 2.
D'après le théorème de récurrence, $\mathcal{P}(n)$ est vraie pour tout entier n, *i.e.* $F = F'$. ■

Théorème 4. Il n'y a pas d'application strictement décroissante de $\mathbb{N}$ dans $\mathbb{N}$.

Démonstration. Raisonnons par l'absurde, en supposant qu'une telle application f existe. Soit $m := \min\big(f(\mathbb{N})\big)$; il existe $k \in \mathbb{N}$ tel que $m = f(k)$. Alors $f(k+1) < f(k) = m$, car f décroît strictement. Mais $f(k+1) \in f(\mathbb{N})$, donc $m \leqslant f(k+1)$, par définition de m. D'où une contradiction. ■

On peut appliquer ce théorème ainsi : soit $\mathcal{P}(n)$ une certaine propriété dépendant d'un entier n. Supposons que l'on sache construire, à partir de tout entier n tel que $\mathcal{P}(n)$ soit vraie, un entier $n' < n$ tel que $\mathcal{P}(n')$ soit vraie. Alors, *pour tout entier k, $\mathcal{P}(k)$ est fausse*. C'est là ce qu'on appelle « principe de descente infinie de Fermat » , Fermat ayant été le premier à l'appliquer pour démontrer certains résultats d'arithmétique.

1.3 Addition et multiplication des entiers naturels

Les résultats précédents permettent de donner une définition rapide des deux opérations de base dans $\mathbb{N}$ et de démontrer leurs principales propriétés. Pour les entiers *intuitifs*, ces propriétés sont connues du lecteur. Contentons-nous donc d'énoncer les résultats et d'esquisser quelques preuves.

Théorème et définition 5.
Il existe une unique application de $\mathbb{N} \times \mathbb{N}$ dans $\mathbb{N}$, notée $+$, telle que :

1. Pour tout entier n, on a $n + 0 = n$, et $n + 1$ est le successeur de n.
2. Pour tous entiers m, n, on a : $m + (n+1) = (m+n) + 1$.

L'application $+$ est appelée *addition* et, si a, b sont deux entiers, $a+b$ est appelé *somme* de a et b.

Démonstration. Notons s l'application de $\mathbb{N}$ dans $\mathbb{N}$ associant à tout entier son successeur. Appliquons le théorème 3 de la page précédente, en prenant $X := \mathbb{N}$, $Y := \mathbb{N}$, u égal à l'application identique de $\mathbb{N}$ et $v := s$. Il existe une unique application F de $\mathbb{N} \times \mathbb{N}$ dans $\mathbb{N}$ telle que, pour tous entiers a, b, on ait $F(a,0) = a$ et $F\big(a, s(b)\big) = s\big(F(a,b)\big)$, d'où $F(a,1) = s(a)$. Il suffit alors de définir $+$ par la formule $m + n = F(m,n)$ pour conclure, le lecteur le vérifiera. ■

Voici des propriétés standard de l'addition :

a) Pour tout entier a, $a+0=a=0+a$.

b) *Associativité* : pour tous entiers a,b,c, on a $(a+b)+c=a+(b+c)$.

c) *Commutativité* : pour tous entiers a,b, on a $b+a=a+b$.

d) *Compatibilité avec l'ordre* : pour tous entiers a,a',b,b' les inégalités $a\leqslant a'$ et $b\leqslant b'$ impliquent la suivante : $a+b\leqslant a'+b'$.

e) Pour tous entiers a,b,c, $(a\leqslant b)\iff(a+c\leqslant b+c)$.

f) *Régularité* : pour tous entiers a,b,c, $(a+c=b+c)\implies(a=b)$.

g) Pour tous entiers a,b, l'égalité $a+b=0$ implique $a=b=0$.

L'assertion a) est immédiate, par récurrence sur a. Pour l'associativité, notons $\mathcal{P}(c)$ la propriété suivante d'un entier c : pour *tous* $a,b\in\mathbb{N}$, $(a+b)+c=a+(b+c)$. La propriété $\mathcal{P}(0)$ est vraie, vu la condition 1 du théorème. Si $\mathcal{P}(c)$ est vraie pour un certain c et si $a,b\in\mathbb{N}$, on voit que $\mathcal{P}(c+1)$ est vraie, en utilisant plusieurs fois la condition 2 du théorème :

$$\begin{aligned}(a+b)+(c+1)&=\big((a+b)+c\big)+1\\&=\big(a+(b+c)\big)+1=a+\big((b+c)+1\big)=a+\big(b+(c+1)\big).\end{aligned}$$

Les démonstrations des autres assertions sont laissées au lecteur, comme exemples de raisonnements par récurrence.

Exercice 1.

Soient X un ensemble et $f:X\to X$ une application. Montrer que $f^m\circ f^n=f^{m+n}$ pour tous entiers m,n.

Solution. Notons $\mathcal{P}(m)$ la propriété suivante d'un entier m :

« pour *tout* $n\in\mathbb{N}$, $f^m\circ f^n=f^{m+n}$. »

La propriété $\mathcal{P}(0)$ est vraie : pour tout $n\in\mathbb{N}$,

$$f^0\circ f^n=\mathrm{Id}_X\circ f^n=f^n=f^{0+n}.$$

Soit m un entier tel que $\mathcal{P}(m)$ soit vraie. Pour tout $n\in\mathbb{N}$, l'associativité de la composition des applications donne :

$$f^{m+1}\circ f^n=(f\circ f^m)\circ f^n=f\circ(f^m\circ f^n)=f\circ f^{m+n}=f^{(m+n)+1}.$$

Mais $(m+n)+1=(m+1)+n$, et donc $\mathcal{P}(m+1)$ est vraie.

La composée de deux applications injectives (resp. surjectives) est injective (resp. surjective), *cf.* le module I.1. Si donc f est injective (resp. surjective), on en déduit, par récurrence sur n, que chaque f^n l'est aussi.

Théorème et définition 6. Soient a,b deux entiers. Pour que l'on ait $a\leqslant b$, il faut et il suffit qu'il existe un entier c tel que $b=a+c$. Un tel entier c est alors unique, on le note $b-a$ et on l'appelle *différence* de b et de a. En d'autres termes, l'application $x\mapsto a+x$ est une bijection de $\mathbb{N}$ sur $[a,\to[$.

Seule la surjectivité de l'application t_a en question (translation de vecteur a) nécessite une preuve, immédiate en raisonnant par récurrence sur a. Noter l'exercice suivant.

Exercice 2.
Pour tout entier $n \geqslant 1$, $n-1$ est le prédécesseur de n.
Solution. Puisque $\mathbb{N}$ est totalement ordonné, il suffit de vérifier qu'il n'y a aucun entier strictement compris entre $n-1$ et n ; mais cela résulte du fait que $n = (n-1)+1$ est le successeur de $n-1$.

Théorème et définition 7. Il existe une unique application de $\mathbb{N} \times \mathbb{N}$ dans $\mathbb{N}$, notée $\times$, vérifiant les propriétés suivantes :

1. Pour tout entier m, on a $m \times 0 = 0$.
2. Pour tous entiers m, n, on a : $m \times (n+1) = (m \times n) + m$.

L'application $\times$ est appelée *multiplication* et, si $a, b \in \mathbb{N}$, $a \times b$ est appelé *produit* de a et b et noté simplement ab si cela ne prête pas à confusion ($2 \times x$ peut être noté $2x$, mais 2×3 ne peut pas être noté 23 !).

Démonstration. Pour tous entiers m, n, posons $m \times n := t_m^n(0)$, où t_m^n est la $n^{ème}$ itérée de la translation de vecteur m. La propriété 1 est évidente. Pour la propriété 2, soient $m, n \in \mathbb{N}$. La commutativité de l'addition donne :

$$m \times (n+1) = t_m^{n+1}(0) = (t_m \circ t_m^n)(0) = t_m(m \times n) = m + (m \times n) = (m \times n) + m\,,$$

d'où la partie existence de l'énoncé.

Le lecteur vérifiera la partie unicité de l'énoncé. Noter qu'une multiplication apparaît ainsi comme une succession d'additions ($x \times 3 = x + (x + x)$, etc). ∎

Voici des propriétés standard de la multiplication, à prouver en exercice :

a) Pour tout entier a, on a $1 \times a = a = a \times 1$.

b) *Associativité* : pour tous entiers a, b, c, on a $(ab)c = a(bc)$.

c) *Commutativité* : pour tous entiers a, b, on a $ba = ab$.

d) *Distributivité de la multiplication par rapport à l'addition* : pour tous entiers a, b, c, on a $a(b+c) = ab + ac$ et $(a+b)c = ac + bc$.

e) *Compatibilité avec l'ordre* : pour tous entiers a, a', b, b' les inégalités $a \leqslant a'$ et $b \leqslant b'$ impliquent la suivante : $ab \leqslant a'b'$. En particulier le produit de deux entiers strictement positifs est strictement positif.

f) Pour tous entiers a, b, c, l'entier c étant *non nul*, les inégalités $a \leqslant b$ et $ac \leqslant bc$ sont équivalentes.

g) *Régularité* : pour tous entiers a, b, c, l'entier c étant *non nul*, l'égalité $ac = bc$ implique $a = b$.

h) Pour tous entiers a, b, l'égalité $ab = 0$ équivaut à $\big((a = 0) \text{ ou } (b = 0)\big)$, et l'égalité $ab = 1$ équivaut à $a = b = 1$.

Proposition 8 (Propriété d'Archimède).
Soient a, b deux entiers, on suppose $a > 0$. Il existe un entier n tel que $na > b$.

Démonstration. D'après la propriété (N3) de $\mathbb{N}$, b n'est pas un majorant de $\mathbb{N}$. Il existe donc un entier n tel que $n > b$. Mais $a > 0$, donc $a \geqslant 1$ (définition de 1), d'où $na \geqslant n \times 1 = n > b$, et ainsi $na > b$. ■

À partir de la multiplication, on définit la suite (a^n) des puissances d'un entier a fixé, par $a^0 := 1$ et la relation de récurrence $a^{n+1} := a^n \times a$.
Outre les égalités évidentes $0^n = 0$ pour tout entier $n > 0$ (mais $0^0 = 1$) et $1^n = 1$ pour tout entier n, les formules classiques à retenir sont les suivantes :

Proposition 9. Pour tous entiers a, b, m, n, on a :

$$a^m \times a^n = a^{m+n} \quad , \quad (a^m)^n = a^{mn} \quad , \quad (ab)^n = a^n \times b^n.$$

Démonstration. Si a est un entier, notons μ_a l'application $x \mapsto ax$ de $\mathbb{N}$ dans $\mathbb{N}$. Pour tout entier n, on a $a^n = \mu_a^n(1)$, d'où (*cf.* l'exercice 1 de la page 67) la première formule. La seconde en résulte, vu la définition de la multiplication à partir de l'addition. La dernière formule se démontre simplement par récurrence sur n, en utilisant la commutativité de la multiplication. ■

Voici un exemple simple de raisonnement par récurrence.

Exercice 3.
Pour tout $n \in \mathbb{N}$, prouver l'inégalité $2^n > n$.
Solution. Ici, 2 est le successeur de 1 $(2 = 1 + 1)$! Soit $\mathcal{P}(n)$ la propriété à établir. Puisque $\mathcal{P}(0)$ s'écrit $1 > 0$, elle est vraie.
Supposons $\mathcal{P}(n)$ vraie pour un certain $n \in \mathbb{N}$: $2^n > n$, soit $2^n \geqslant n + 1$, vu la définition du successeur. Multiplions par 2 : $2^{n+1} \geqslant 2n + 2 \geqslant n + 2 > n + 1$, donc $\mathcal{P}(n + 1)$ est vraie. Le théorème de récurrence montre donc que $\mathcal{P}(n)$ est vraie pour tout $n \in \mathbb{N}$.

Lemme 10.
Soient deux entiers m, t tels que $m > 1$ et $1 \leqslant t \leqslant m$.
Alors il existe une bijection u de l'ensemble $T = [1, m] \setminus \{t\}$ (complémentaire du singleton $\{t\}$ dans $[1, m]$) sur $[1, m - 1]$.

Démonstration.
L'application s de $[1, m]$ dans $[1, m]$ définie par

$$s(t) := m, \qquad s(m) := t \qquad \text{et} \qquad s(k) := k \quad \text{si} \quad k \notin \{m, t\}$$

est une bijection de $[1, m]$ sur $[1, m]$ (c'est par exemple l'identité si $t = m$).
L'application u de T dans $[1, m - 1]$ définie par $u(k) := s(k)$ convient. ■

La proposition ci-dessous sera très utile dans la section suivante.

Proposition 11. Soient n et p deux entiers strictement positifs.

1. Il existe une injection de $[1,n]$ dans $[1,p]$ si, et seulement si, $n \leqslant p$.
2. Il existe une surjection de $[1,n]$ sur $[1,p]$ si, et seulement si, $n \geqslant p$.

Démonstration. L'application f de $[1,n]$ dans $[1,p]$ définie par $f(k) := \min(k,p)$ est injective si $n \leqslant p$ et surjective si $n \geqslant p$, d'où la suffisance en 1 et 2.

Venons-en aux conditions nécessaires. Démontrons par récurrence sur $n \geqslant 1$ la propriété $\mathcal{P}(n)$ selon laquelle l'existence d'une injection de $[1,n]$ dans $[1,m]$ implique $n \leqslant m$.

La propriété $\mathcal{P}(1)$ est évidente. Supposons $\mathcal{P}(n)$ vraie pour un entier $n \geqslant 1$, et démontrons $\mathcal{P}(n+1)$. Soient f une injection de $[1,n+1]$ dans $[1,m]$ et $t = f(n+1)$; la restriction g de f à $[1,n]$ est une injection de $[1,n]$ dans $[1,m] \setminus \{t\} \neq \varnothing$. Par suite $u \circ g$ est une injection de $[1,n]$ dans $[1,m-1]$, d'où $n \leqslant m-1$ et $n+1 \leqslant m$ (ici u est comme dans le lemme précédent).

La preuve de l'autre condition nécessaire est analogue. Soit f une surjection de $[1,n+1]$ sur $[1,m]$; les égalités :

$$[1,m] = f([1,n+1]) = f([1,n]) \cup \{f(n+1)\}$$

font qu'ou bien $f([1,n]) = [1,m]$, alors $n+1 > n \geqslant m$ par hypothèse de récurrence, ou bien il existe un t tel que $f([1,n]) = [1,m] \setminus \{t\}$. Alors $u \circ g$ est une surjection de $[1,n]$ sur $[1,m-1]$, d'où $n \geqslant m-1$ et $n+1 \geqslant m$. ∎

2 Dénombrement

2.1 Ensembles finis, ensembles dénombrables

Les ensembles finis, ainsi que les ensembles dénombrables dont il sera question plus loin, ont déjà été abordés dans le module I.1. Le point de vue adopté ici est un peu différent, il précise la définition intuitive « un ensemble fini est un ensemble qui a un nombre fini d'éléments ».

Définition 1. Un ensemble X est dit *fini* s'il est vide ou s'il est en bijection avec $[1,n]$, pour un certain entier $n \geqslant 1$. S'il n'est pas fini, X est dit *infini*.

Théorème et définition 12. Soit X un ensemble fini. Si X n'est pas vide, il existe un unique entier $n \geqslant 1$ tel que X soit en bijection avec $[1,n]$. Cet entier n est appelé *cardinal* de X, on le note $\operatorname{card}(X)$, et l'on dit que X est un ensemble à n éléments. Si X est vide, on pose $\operatorname{card}(X) := 0$.

Démonstration. Soient $n,p \geqslant 1$ deux entiers, g une bijection de X sur $[1,n]$, et h une bijection de X sur $[1,p]$. Nous devons prouver que $n = p$. Or $v := h \circ g^{-1}$ est une bijection de $[1,n]$ sur $[1,p]$. Ainsi $n \leqslant p$ car v est injective (assertion 1 de la proposition 11) mais aussi $n \geqslant p$ car v est surjective (assertion 2 de la proposition 11). D'où $n = p$. ∎

La proposition 11 de la page précédente peut maintenant s'énoncer ainsi :

Théorème 13. Soient X, Y deux ensembles finis non vides, de cardinaux respectifs n, p.

1. Il existe une injection de X dans Y si, et seulement si, $n \leqslant p$.
2. Il existe une surjection de X dans Y si, et seulement si, $n \geqslant p$.
3. Il existe une bijection de X dans Y si, et seulement si, $n = p$.

Démonstration. Elle est immédiate à partir de la proposition 11 de la page ci-contre. Il existe une bijection u de $[1, n]$ sur X et une bijection v de Y sur $[1, p]$. Soient $f : X \to Y$ et $F : [1, n] \to [1, p]$ deux applications. Les égalités $F = v \circ f \circ u$ et $f = v^{-1} \circ F \circ u^{-1}$ sont équivalentes. Si elles sont vérifiées, f sera injective (resp. surjective) si, et seulement si, F l'est, d'où aussitôt la conclusion. ■

Voici une conséquence simple des résultats précédents. Soient n, p deux entiers tels que $n > p$. Aucune application d'un ensemble fini à n éléments dans un ensemble fini à p éléments n'est injective. En effet, l'existence d'une injection d'un ensemble à n éléments dans un ensemble à p éléments contredirait l'assertion 1 du théorème précédent. Le résultat ci-dessus s'appelle « principe des tiroirs », à cause de l'interprétation suivante. Si l'on distribue n objets dans p tiroirs, et si $n > p$, l'un des tiroirs contiendra au moins deux objets. En effet, soient X l'ensemble des objets, Y l'ensemble des tiroirs et $f : X \to Y$ l'application associant à tout objet x le tiroir $f(x)$ dans lequel x est placé. La conclusion traduit la non injectivité de f.

Théorème 14. Soient X un ensemble fini et A une partie de X. L'ensemble A est fini, $\operatorname{card}(A) \leqslant \operatorname{card}(X)$, et cette inégalité n'est une égalité que si $A = X$.

Démonstration. Raisonnons par récurrence sur $n = \operatorname{card}(X)$, le cas $n = 0$ étant évident. Supposons X de cardinal $n \geqslant 1$, la conclusion étant vraie pour tout ensemble fini de cardinal strictement inférieur à n. Soit A une partie de X. Si $A = X$, il n'y a rien à démontrer. Sinon, soit $x \in X \setminus A$. Il existe un $q \in [1, n]$ et une bijection de $X \setminus \{x\}$ sur $[1, n] \setminus \{q\}$. Comme $[1, n] \setminus \{q\}$ est de cardinal $n - 1$, $X \setminus \{x\}$ l'est aussi. Or $A \subset X \setminus \{x\}$. D'après l'hypothèse de récurrence, A est donc fini, de cardinal au plus $n-1$, d'où la conclusion. ■

Théorème 15. Soient X, Y deux ensembles de *même* cardinal fini n. Une application $f : X \to Y$ est bijective si, et seulement si, elle est injective (resp. surjective).

Démonstration. Au moyen de bijections de $[1, n]$ sur X et Y, on se ramène au cas où $X = Y = [1, n]$. Si f est bijective, elle est injective et surjective. Si f est injective, l'application $x \mapsto f(x)$ est une bijection de $[1, n]$ sur $f([1, n])$. Ainsi $f([1, n])$ est une partie de $[1, n]$ à n éléments, donc $f([1, n]) = [1, n]$, vu le théorème précédent : f est surjective, donc bijective. Supposons f surjective. Si $k \in [1, n]$, notons $g(k)$ le plus petit des antécédents de k par f. On définit ainsi une application g de $[1, n]$ dans lui-même, et $f \circ g$ est, par construction, l'application identique de $[1, n]$. Il en résulte que g est injective, car l'application identique l'est. Vu ce qui précède, g est bijective, donc f l'est aussi (c'est la réciproque de g). ■

Un problème de *dénombrement* consiste à calculer le cardinal (ou nombre d'éléments) d'un ensemble fini X, pouvant être défini par des procédés variés. La première méthode, dite *méthode bijective*, est d'établir une bijection entre X et un ensemble fini Y dont on connaît déjà le cardinal ; si l'on y parvient, on aura $\operatorname{card}(X) = \operatorname{card}(Y)$, en vertu du théorème 13 de la page précédente. Voici un exemple simple : si $a, b \in \mathbb{N}$ et si $a \leqslant b$, le cardinal de $[a, b]$ est $b - a + 1$ (ne pas oublier le 1), car $x \mapsto x + a - 1$ est une bijection de $[1, b - a + 1]$ sur $[a, b]$, comme le lecteur le vérifiera. Voici deux théorèmes fondamentaux, qui seront illustrés dans la section suivante.

Théorème 16. Soient X un ensemble, $n \geqslant 1$ un entier et $A_1, \ldots, A_n$ des parties finies de X deux à deux disjointes. La réunion de ces parties est finie, et :

$$\operatorname{card}(A_1 \cup A_2 \cup \cdots \cup A_n) = \operatorname{card}(A_1) + \operatorname{card}(A_2) + \cdots + \operatorname{card}(A_n).$$

Démonstration. Par récurrence sur n, on se ramène facilement au cas $n = 2$. Posons $m_i := \operatorname{card}(A_i)$, $i = 1, 2$. Soit f_1 une bijection de A_1 sur $[1, m_1]$. Puisque $J := [m_1 + 1, m_1 + m_2]$ est de cardinal m_2, il existe une bijection f_2 de A_2 sur J. Définissons une application $f : A_1 \cup A_2 \to [1, m_1 + m_2]$ comme suit : si $x \in A_1 \cup A_2$, on pose $f(x) := f_1(x)$ si $x \in A_1$ et $f(x) := f_2(x)$ si $x \in A_2$; cette définition est licite parce que $A_1 \cap A_2$ est vide. Il est clair que f est une bijection de $A_1 \cup A_2$ sur $[1, m_1 + m_2]$, d'où la conclusion. ∎

Définition 2. Soient X, Y deux ensembles et $f : X \to Y$ une application. Si $y \in Y$, soit $f^{-1}(\{y\})$ l'image réciproque de $\{y\}$ par f, *i.e.* l'ensemble des antécédents de y par f. Un tel ensemble est appelé une *fibre* de f.

Lorsque cela ne prête pas à confusion, $f^{-1}(\{y\})$ sera noté simplement $f^{-1}(y)$.

Exemple. Soient X, Y deux ensembles finis et $f : X \to Y$ une application. Lorsque y décrit Y, les fibres $f^{-1}(y)$ forment une partition de X : si $x \in X$, le seul élément y de Y tel que $x \in f^{-1}(y)$ est $f(x)$. D'où :

$$\operatorname{card}(X) = \sum_{y \in Y} \operatorname{card}\left(f^{-1}(y)\right). \tag{1}$$

Soient maintenant A, B deux ensembles finis et E une partie de $A \times B$. On peut calculer le cardinal de E de deux façons, par la *formule de Fubini* :

$$\operatorname{card}(E) = \sum_{a \in A} \operatorname{card}\left(\{b \in B \mid (a, b) \in E\}\right) = \sum_{b \in B} \operatorname{card}\left(\{a \in A \mid (a, b) \in E\}\right).$$

Soit en effet $p : E \to A$ l'application définie par $(x, y) \mapsto x$ (ne pas oublier que E est une partie de $A \times B$). Si $a \in A$, la fibre $p^{-1}(a)$ est l'intersection $E \cap (\{a\} \times B)$, en bijection avec l'ensemble des $b \in B$ tels que $(a, b) \in E$. La première égalité résulte donc de l'égalité (1), appliquée à p. La deuxième égalité est analogue (échanger les rôles de A et B).

Théorème 17. La réunion de deux parties finies A, B d'un ensemble X est finie, et :

$$\operatorname{card}(A \cup B) + \operatorname{card}(A \cap B) = \operatorname{card}(A) + \operatorname{card}(B).$$

Démonstration. Posons $A' := A \setminus B$ et $B' := B \setminus A$. Ce sont des parties finies de X (théorème 14 de la page 71). Puisque A est réunion disjointe de $A \cap B$ et A', le théorème précédent montre que $\operatorname{card}(A) = \operatorname{card}(A \cap B) + \operatorname{card}(A')$. De même $\operatorname{card}(A \cup B) = \operatorname{card}(B) + \operatorname{card}(A')$, car $A \cup B$ est réunion disjointe de B et de A'. L'égalité annoncée en résulte :

$$\operatorname{card}(A \cup B) + \operatorname{card}(A \cap B) + \operatorname{card}(A') = \operatorname{card}(A) + \operatorname{card}(B) + \operatorname{card}(A'). \quad \blacksquare$$

Les théorèmes 16 et 17 seront généralisés dans le module II.2 (*cf.* la « formule du crible » à l'exercice 7 de la page 136).

Théorème 18. Soient $n \geqslant 1$ un entier et $X_1, \ldots, X_n$ des ensembles finis. L'ensemble produit $X_1 \times X_2 \times \cdots \times X_n$ est fini, et l'on a :

$$\operatorname{card}(X_1 \times X_2 \times \cdots \times X_n) = \operatorname{card}(X_1) \times \operatorname{card}(X_2) \times \cdots \times \operatorname{card}(X_n).$$

Démonstration. Il suffit de traiter le cas $n = 2$, en observant par exemple que, si X, Y, Z sont trois ensembles, $(x, y, z) \mapsto ((x, y), z)$ est une bijection de $X \times Y \times Z$ sur $(X \times Y) \times Z$. Écrivons X, Y au lieu de X_1, X_2, et posons $m := \operatorname{card}(X)$, $p := \operatorname{card}(Y)$. Si $x \in X$, $A_x := \{x\} \times Y$ est une partie à p éléments de $X \times Y$, car $y \mapsto (x, y)$ est une bijection de Y sur A_x. Lorsque x décrit X, les A_x forment une *partition* de $X \times Y$: ces parties sont deux à deux disjointes et leur réunion est $X \times Y$.

D'après le théorème 16 de la page précédente, le cardinal de $X \times Y$ est donc $\overbrace{p + p + \cdots + p}^{m \text{ termes}}$, c'est-à-dire mp. Noter que la multiplication apparaît ici comme une succession d'additions. Voir une autre preuve à l'exercice II.1.15. $\blacksquare$

Terminons par quelques résultats sur les ensembles infinis. Une étude précise des cardinaux de ces ensembles dépasse le cadre de cet ouvrage. L'*existence* d'ensembles infinis n'est pas évidente ; en fait, dans certaines présentations théoriques, cette existence est prise comme axiome. Ici, elle va résulter des propriétés de $\mathbb{N}$ déjà admises ou établies.

Théorème 19. Pour un ensemble X, les conditions ci-dessous sont équivalentes :

a) Il existe une injection de $\mathbb{N}$ dans X.

b) Il existe une bijection de X sur une partie de X distincte de X.

c) L'ensemble X est infini.

En particulier, l'ensemble $\mathbb{N}$ est infini.

Démonstration. Pour l'implication a)$\Rightarrow$b), soient $f : \mathbb{N} \to X$ une injection et $a = f(0)$. Définissons une application $g : X \to X \setminus \{a\}$ ainsi : si $x \in X \setminus f(\mathbb{N})$, on pose $g(x) := x$; si x appartient à $f(\mathbb{N})$, il existe un unique entier n tel que $x = f(n)$ (injectivité de f), on

pose $g(x) := f(n+1)$. On vérifie sans peine que g est une bijection de X sur $X \setminus \{a\}$. L'implication b)$\Rightarrow$c) résulte trivialement, par contraposition, du théorème 14 de la page 71. Puisque $n \mapsto n+1$ est une bijection de $\mathbb{N}$ sur $\mathbb{N}^*$, une conséquence de cette implication est que $\mathbb{N}$ est infini.
Il reste l'implication c)$\Rightarrow$a), contentons-nous d'en ébaucher la preuve. Supposons donc X infini, et soit $a \in X$ (X n'est pas vide !). Définissons une suite (u_n) d'éléments de X par son premier terme $u_0 := a$ et le « procédé récurrent » suivant : si $n \in \mathbb{N}$ et si $u_0, \ldots, u_n$ ont déjà été définis et sont deux à deux distincts, la partie $\{u_0, \ldots, u_n\}$ de X est finie, donc n'est pas égale à X. On « choisit » donc un élément u_{n+1} de X n'appartenant pas à cette partie. La suite (u_n) étant ainsi construite, l'application $n \mapsto u_n$ de $\mathbb{N}$ dans X est évidemment injective. Cette construction peut être *justifiée* par certains axiomes de théorie des ensembles. ■

Un ensemble X est fini si, et seulement s'il n'existe pas de bijection de X sur une partie de X *distincte de* X. Cette propriété ne fait plus référence à $\mathbb{N}$; dans le module I.1, elle a été prise comme *définition* des ensembles finis. Les ensembles infinis peuvent être plus ou moins « gros ». Comme dans le cas fini, on dit que deux ensembles quelconques ont même cardinal ou sont *équipotents* s'il existe une bijection de l'un sur l'autre. Ainsi $\mathbb{N}^*$ a le même cardinal que $\mathbb{N}$, il est « aussi gros » que $\mathbb{N}$; cela montre les limites de notre intuition concernant la notion d'infini.

Définition 3. Un ensemble est dit *dénombrable* s'il est en bijection avec $\mathbb{N}$.

Si X est un ensemble dénombrable, *choisir* une bijection de $\mathbb{N}$ sur X revient à *numéroter* tous les éléments de X, de façon exhaustive et sans répétitions. Signalons trois résultats importants concernant les ensembles dénombrables.

Théorème 20. Toute partie d'un ensemble dénombrable est finie ou dénombrable.

Démonstration. Il suffit de voir que toute partie infinie A de $\mathbb{N}$ est dénombrable.
D'abord, A n'est pas majorée, sinon on aurait $A \subset [0, \max(A)]$, donc A serait finie. Soit $a := \min(A)$. On définit une suite (u_n) d'entiers par $u_0 := a$ et la relation de récurrence $u_{n+1} := \min\Big(A \cap]u_n, \rightarrow[\Big)$. Le lecteur vérifiera que $n \mapsto u_n$ est une bijection de $\mathbb{N}$ sur A. ■

Une conséquence est qu'un ensemble X est *fini ou dénombrable* si, et seulement s'il existe une injection de X dans $\mathbb{N}$.

Théorème 21. Si X et Y sont deux ensembles dénombrables, $X \times Y$ est dénombrable. Plus généralement, étant donnés des ensembles dénombrables $X_1, \ldots, X_p$ ($p \geqslant 1$), leur produit $X_1 \times X_2 \times \cdots \times X_p$ est dénombrable.

Démonstration. La deuxième assertion résulte de la première (récurrence sur p). Pour la première, il suffit de montrer que $\mathbb{N} \times \mathbb{N}$ est dénombrable. Esquissons deux méthodes pour

y parvenir.

– La première consiste à énumérer ainsi les éléments de $\mathbb{N} \times \mathbb{N}$:

$$(0,0),\ (0,1),\ (1,0),\ (0,2),\ (1,1),\ (2,0),\ (0,3),\ (1,2),\ (2,1),\ (3,0),\ldots$$

On écrit donc successivement, pour $n = 0, 1, 2\ldots$ tous les couples $(a, b) \in \mathbb{N} \times \mathbb{N}$ tels que $a + b = n$, rangés par ordre de a croissant. Nous laissons le lecteur expliciter la bijection de $\mathbb{N}$ sur $\mathbb{N} \times \mathbb{N}$ correspondante et la bijection réciproque, puis donner l'interprétation graphique en termes de quadrillage du plan.

– La deuxième méthode utilise par anticipation la section 3.2. Tout entier strictement positif s'écrit de façon unique comme produit d'une puissance de 2 et d'un entier impair ; l'application $(n, p) \mapsto 2^n(2p + 1) - 1$ est donc une bijection de $\mathbb{N} \times \mathbb{N}$ sur $\mathbb{N}$. ■

Théorème 22. Soient X un ensemble, I un ensemble fini ou dénombrable et $(A_i)_{i\in I}$ une famille de parties finies ou dénombrables de X. La réunion des A_i est finie ou dénombrable.

Démonstration. Esquissons cette preuve. Soit A la réunion des A_i. D'abord, au moyen d'une injection de I dans $\mathbb{N}$, et quitte à rajouter des A_i vides, on se ramène au cas où $I = \mathbb{N}$. Pour chaque $i \in \mathbb{N}$, il existe une injection f_i de A_i dans $\mathbb{N}$. Définissons une application $f : A \to \mathbb{N} \times \mathbb{N}$ ainsi. Soient $x \in A$ et i le plus petit entier tel que $x \in A_i$. On pose : $f(x) := \big(i, f_i(x)\big)$. On montre aisément que f est injective, donc en bijection avec son image $f(A)$. Cette dernière est une partie de $\mathbb{N} \times \mathbb{N}$, qui est dénombrable, donc $f(A)$ est finie ou dénombrable, d'après le théorème 20 de la page précédente. ■

Théorème 23. Soit f une application d'un ensemble X dans un ensemble Y.

1. On suppose f *injective*. Si Y est fini, X l'est aussi. Si Y est dénombrable, X est fini ou dénombrable.
2. On suppose f *surjective*. Si X est fini, Y l'est aussi. Si X est dénombrable, Y est fini ou dénombrable.

Démonstration. Pour l'assertion 1, X étant en bijection avec $f(X)$, la conclusion vient du théorème 14 de la page 71 et du théorème 20 de la page précédente. Pour l'assertion 2, on peut invoquer le fait qu'il existe une injection de Y dans X (*cf.* le module I.1), puis appliquer l'assertion 1. ■

Il apparaîtra plus loin que l'ensemble $\mathbb{Z}$ des entiers relatifs et l'ensemble $\mathbb{Q}$ des nombres rationnels sont dénombrables, mais que l'ensemble $\mathbb{R}$ des nombres réels ne l'est pas, il est « vraiment plus gros » que $\mathbb{N}$.

2.2 Analyse combinatoire

L'objectif de cette section est de résoudre des problèmes de dénombrement classiques, en montrant l'efficacité des résultats de la section précédente. Nous utilisons ici les nombres

rationnels, réels, complexes, bien que la construction « rigoureuse » des ensembles $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ n'ait pas encore été donnée. Rappelons que la suite des factorielles $(n!)$ est définie par son premier terme $0! := 1$ et la relation de récurrence $(n+1)! := n! \times (n+1)$. Ainsi :

$$0! = 1\ ,\ \ 1! = 1\ ,\ \ 2! = 2\ ,\ \ 3! = 6\ ,\ \ 4! = 24\ ,\ \ 5! = 120\ ,\ \ 6! = 720.$$

La suite $(n!)$ croît très rapidement (sa croissance est « ultra-exponentielle »).
Par exemple $10! = 3\,628\,800$ est déjà « grand ».

Théorème 24. Soient X, Y deux ensembles finis de cardinaux respectifs $n, p \geqslant 1$. L'ensemble $\mathcal{F}(X,Y)$ des applications de X dans Y est de cardinal p^n.

Démonstration. Numérotons les éléments de X : $x_1, \ldots, x_n$, ce qui revient à choisir une bijection $i \mapsto x_i$ de $[1,n]$ sur X. On définit alors une application $F : f \mapsto \big(f(x_1), \ldots, f(x_n)\big)$ de $\mathcal{F}(X,Y)$ dans $\overbrace{Y \times Y \times \cdots \times Y}^{n \text{ facteurs}} = Y^n$.
Soit $v = (y_1, \ldots, y_n) \in Y^n$. Le seul antécédent de v par F est l'application $f : X \to Y$ définie par $f(x_i) := y_i$, $i = 1, \ldots, n$. Ainsi F est bijective. D'après le théorème 18 de la page 73, Y^n est de cardinal p^n, d'où la conclusion. ∎

Introduisons maintenant une notion importante, notamment en calcul des probabilités, celle de *fonction caractéristique* d'une partie d'un ensemble.

Définition 4. Soient X un ensemble et A une partie de X. On appelle *fonction caractéristique* de A l'application χ_A de X dans $\{0,1\}$ prenant la valeur 1 sur A et la valeur 0 sur $X \setminus A$, complémentaire de A dans X.

Proposition 25. Soit X un ensemble. L'application $A \mapsto \chi_A$ est une bijection de l'ensemble $\mathcal{P}(X)$ des parties de X sur $\mathcal{F}(X, \{0,1\})$.

Démonstration. À toute fonction $f \in \mathcal{F}\big(X, \{0,1\}\big)$, associons la partie $f^{-1}(\{1\})$ de X. Il est clair que les applications $A \mapsto \chi_A$ de $\mathcal{P}(X)$ dans $\mathcal{F}(X, \{0,1\})$ et $f \mapsto f^{-1}\big(\{1\}\big)$ de $\mathcal{F}(X, \{0,1\})$ dans $\mathcal{P}(X)$ sont deux bijections réciproques (regarder les composées dans les deux sens). ∎

Théorème 26. Soit X un ensemble fini à n éléments. L'ensemble $\mathcal{P}(X)$ des parties de X est fini, de cardinal 2^n.

Démonstration. Compte tenu de la proposition précédente, il suffit d'appliquer le théorème 24, en prenant $Y := \{0,1\}$ et $p := 2$. ∎

Soient n, p deux entiers tels que $0 \leqslant p \leqslant n$. Un *arrangement* de n objets pris p à p est une suite de p objets *distincts* pris parmi les n objets donnés. On note A_n^p le nombre d'arrangements de n objets pris p à p. Ainsi, un arrangement des 26 lettres de l'alphabet prises 3 à 3 est un mot formé de trois lettres *distinctes*. Combien y a-t-il de tels mots ? Il

y a 26 possibilités pour la première lettre ; une fois cette lettre choisie, il reste 25 choix possibles pour la 2ème puis, cette 2ème lettre étant choisie, il reste 24 choix possibles pour la 3ème . Le nombre cherché est donc $A_{26}^3 = 26 \times 25 \times 24$.
Généralisons cet exemple. Soit Y un ensemble fini, de cardinal n. Une suite $(y_1, \ldots, y_p)$ de p éléments de Y est par définition une application $f : [1, p] \to Y$ ($f(i) = y_i$ pour $i = 1, \ldots, p$), et $y_1, \ldots, y_p$ sont distincts si, et seulement si, f est *injective*. Ainsi un arrangement des n éléments de Y pris p à p est simplement une application *injective* de $[1, p]$ dans Y. La détermination des nombres A_n^p résultera ainsi du théorème suivant :

Théorème 27. Soient X, Y deux ensembles finis, $p := \operatorname{card}(X), n := \operatorname{card}(Y)$. Si $p \leqslant n$, le cardinal de l'ensemble $I(X, Y)$ des *injections* de X dans Y est :

$$A_n^p := \overbrace{n(n-1)\cdots(n-p+1)}^{p \text{ facteurs}} = \frac{n!}{(n-p)!} \cdot$$

Si $p > n$, il n'y a pas d'injections de X dans Y.

Démonstration. Prenons ici la première égalité comme *définition* de A_n^p. La deuxième égalité de l'énoncé est évidente. Pour tout entier p, notons $\mathcal{P}(p)$ la propriété suivante : pour *tout* ensemble X de cardinal p et *tout* ensemble fini Y de cardinal $n \geqslant p$, le cardinal de $I(X, Y)$ est A_n^p. La propriété $\mathcal{P}(0)$ est vraie : pour tout ensemble Z, il y a une seule application de l'ensemble vide dans Z, et elle est injective (se référer aux définitions). Soit $p \geqslant 1$ un entier tel que $\mathcal{P}(p-1)$ soit vraie. Il nous suffit de montrer que $\mathcal{P}(p)$ est vraie. Si donc X, Y, n sont comme dans l'énoncé, nous devons démontrer que le cardinal de $I(X, Y)$ est A_n^p. Fixons un élément a de X, et soit $X' := X \setminus \{a\}$, qui est de cardinal $p-1$.
À toute injection f de $I(X, Y)$ on peut associer bijectivement le couple (g, b) où g est la restriction de f à X' et b est $f(a)$, qui n'appartient pas à $g(X') = f(X')$ qui est de cardinal $p-1$. Par suite :

$$\operatorname{card}\big(I(X, Y)\big) = \operatorname{card}\big(I(X', Y)\big)\,(n-p+1) = A_n^{p-1}(n-p+1) = A_n^p.$$

Nous avons ainsi démontré que $\mathcal{P}(p)$ est vraie.
La dernière assertion du théorème résulte du théorème 13 de la page 71. ■

Le nombre A_n^p est donc bien le *nombre d'arrangements de n objets pris p à p*. Voici une conséquence immédiate de ce théorème et du théorème 15 de la page 71.

Théorème 28. Soient X, Y deux ensembles finis de *même* cardinal n. Le nombre de bijections de X sur Y est $n!$. En particulier l'ensemble $\mathfrak{S}_X$ des bijections de X sur X (on les appelle *permutations* de X) est de cardinal $n!$.

Soit p un entier naturel. Pour tout nombre *complexe* a, on pose :

$$\binom{a}{p} := \frac{a(a-1)\cdots(a-p+1)}{p!} \cdot \tag{2}$$

Le numérateur est un produit de p facteurs, il vaut 1, par convention, lorsque $p = 0$.

A priori, $\binom{a}{p}$ est alors un nombre complexe. Par exemple :

$$\binom{a}{0} = 1\,,\quad \binom{a}{1} = a\,,\quad \binom{a}{2} = \frac{a(a-1)}{2}\,,\quad \binom{a}{3} = \frac{a(a-1)(a-2)}{6}\,.$$

Soient n, p deux entiers tels que $0 \leqslant p \leqslant n$ et X un ensemble de cardinal n. On appelle *combinaison* des n éléments de X pris p à p toute partie de X à p éléments. Le théorème suivant va montrer que le nombre de ces combinaisons est $\binom{n}{p}$.

Prenons $p = 3$, X étant l'ensemble des 26 lettres de l'alphabet. Chaque partie de X à 3 éléments, par exemple $\{a, i, r\}$, « correspond » à $3! = 6$ mots formés de trois lettres distinctes (air, ari, iar, ira, rai, ria). Le nombre de parties de X à 3 éléments est « donc » $A_{26}^3/3! = \binom{26}{3}$.

Théorème 29. Soit X un ensemble fini de cardinal n.
Étant donné un entier $p \in [0, n]$, le nombre de parties à p éléments de X est :

$$\binom{n}{p} := \frac{n(n-1)\cdots(n-p+1)}{p!} = \frac{n!}{p!(n-p)!} = \frac{A_n^p}{p!}\,. \tag{3}$$

Démonstration. La première égalité est la définition du nombre $\binom{n}{p}$, les deux dernières viennent du théorème 27 de la page précédente. Nous devons démontrer que l'ensemble $\mathcal{P}_p(X)$ des parties à p éléments de X a pour cardinal $A_n^p/p!$. Si $f : [1, p] \to X$ est une injection, $f([1, p])$ est une partie à p éléments de X, d'où une application $t : f \mapsto f([1, p])$ de $I([1, p], X)$ dans $\mathcal{P}_p(X)$. Soit B une partie à p éléments de X. La fibre $t^{-1}(\{B\})$ est formée des injections $f : [1, p] \to X$ ayant pour image B, c'est-à-dire des bijections de $[1, p]$ sur B. D'après le théorème précédent, cette fibre est donc de cardinal $p!$. La formule (1) de la page 72 et le théorème 27 de la page précédente impliquent alors :

$$A_n^p = \operatorname{card}\Big(I([1, p], X)\Big) = \sum_{B \in \mathcal{P}_p(X)} p! = p!\,\operatorname{card}\Big(\mathcal{P}_p(X)\Big). \qquad \blacksquare$$

Soit X un ensemble de cardinal fini n ($[1, n]$ par exemple).
Les $\mathcal{P}_p(X)$, pour $p = 0, 1, \ldots, n$ forment une partition de $\mathcal{P}(X)$. Le théorème précédent et le théorème 16 de la page 72 donnent donc l'égalité suivante :

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n.$$

À titre d'exercice, donner une autre preuve de cette formule (sans la formule du binôme, bien sûr). Les coefficients $\binom{n}{p}$ sont appelés *coefficients binomiaux*, à cause de la *formule du binôme* (*cf.* le module II.2).

Exemples. Par définition, $\binom{50}{3} = 50!/(47!\,3!)$, mais il ne faut surtout pas calculer $50!$ (qui a 65 chiffres décimaux) : $\binom{50}{3} = (50 \times 49 \times 48)/6 = 50 \times 49 \times 8 = 19\,600$.

Voici un autre exemple. Soient n, p deux entiers naturels. Alors :

$$\begin{aligned}\binom{-(n+1)}{p} &= \frac{(-n-1)(-n-2)\cdots(-n-p)}{p!}\\ &= (-1)^p\,\frac{(n+p)(n+p-1)\cdots(n+1)}{p!} = (-1)^p\binom{n+p}{p}.\end{aligned}$$

Sur la formule (3), il n'est pas évident que le nombre rationnel $\binom{n}{p}$ soit entier. C'est vrai à cause du théorème 29 de la page précédente ! Cela résulte aussi (voir l'exercice II.1.5) de la formule suivante, qui est par ailleurs essentielle :

$$\binom{a+1}{p+1} = \binom{a}{p} + \binom{a}{p+1} \quad (a \in \mathbb{C},\ p \in \mathbb{N}). \tag{4}$$

Prouvons cette formule. En multipliant les deux membres par $(p+1)!$, il suffit (le lecteur s'en convaincra) de voir que la somme

$$(p+1)a(a-1)\cdots(a-p+1) + a(a-1)\cdots(a-p+1)(a-p)$$

vaut $(a+1)a(a-1)\cdots(a-p+1)$, ce qui est évident. En particulier :

$$\binom{n+1}{p+1} = \binom{n}{p} + \binom{n}{p+1} \quad (n, p \in \mathbb{N},\ p \leqslant n-1). \tag{5}$$

Cette égalité est appelée *relation de Pascal*. C'est la règle de formation du *triangle arithmétique de Pascal*, dans lequel on dispose, ligne après ligne, les coefficients binomiaux :

$n \backslash p$	0	1	2	3	4	5
0	1					
1	1	1				
2	1	2	1			
3	1	3	3	1		
4	1	4	6	4	1	
5	1	5	10	10	5	1

Noter la symétrie que présente ce tableau : $\binom{n}{p} = \binom{n}{n-p}$ lorsque $0 \leqslant p \leqslant n$. Vérifier cette propriété de symétrie directement, puis à l'aide du théorème 29 de la page ci-contre.

Exercice 4.
Étant donnés p objets identiques, quel est le nombre $T(n,p)$ de manières qu'il y a de répartir ces p objets entre n boîtes ($n,p \in \mathbb{N}^*$) ?
On dit que $T(n,p)$ est le nombre de *combinaisons avec répétitions de n objets pris p à p*.
Solution. Numérotons les boîtes : $B_1,\dots,B_n$. Il suffit de se donner, pour tout $i \in [1,n]$, le nombre a_i d'objets qu'on met dans la boîte B_i. Ainsi $T(n,p)$ est le cardinal de l'ensemble A des n-uples $(a_1,\dots,a_n) \in \mathbb{N}^n$ tels que $a_1+a_2+\cdots+a_n = p$. Démontrons la formule suivante :

$$T(n,p) \;=\; \binom{n+p-1}{n-1} \;=\; \binom{n+p-1}{p}. \tag{6}$$

Notons B l'ensemble des suites strictement croissantes d'entiers $(b_1,\dots,b_n)$ telles que $b_n = n+p$. Si $(a_1,\dots,a_n) \in A$, posons $b_i := a_1+\cdots+a_i+i$ pour $i=1,\dots,n$. Il est clair que $(b_1,\dots,b_n) \in B$, et que l'application $(a_1,\dots,a_n) \mapsto (b_1,\dots,b_n)$ ainsi définie est une bijection de A sur B : les a_i sont donnés en fonction des b_i par $a_1 := b_1 - 1$ et $a_i := b_i - b_{i-1} - 1$ pour $i = 2,\dots,n$. Ainsi $T(n,p) = \operatorname{card}(B)$. Si $(b_1,\dots,b_n) \in B$, les b_i sont distincts, donc $\{b_1,\dots,b_{n-1}\}$ est une partie à $n-1$ éléments de $[1,n+p-1]$ (car $b_n = n+p$, donc $b_i < n+p$ pour $i=1,\dots,n-1$).
Enfin $(b_1,\dots,b_n) \mapsto \{b_1,\dots,b_{n-1}\}$ est une bijection de B sur l'ensemble des parties à $n-1$ éléments de $[1,n+p-1]$, comme on le vérifiera, donc $\operatorname{card}(B) = \binom{n+p-1}{n-1}$, en vertu du théorème 29 de la page 78. D'où la conclusion.

3 Divisibilité

La divisibilité sera étudiée deux fois : d'abord dans $\mathbb{N}$ puis dans $\mathbb{Z}$. Quelques résultats seront démontrés de deux façons ; cela n'était pas indispensable, mais nous a paru instructif. Ce sera le cas pour deux théorèmes fondamentaux (et liés entre eux) : le *théorème de Gauß* (théorème 40 de la page 88 et théorème 54 de la page 97), et le *lemme d'Euclide* (théorème 36 de la page 85 et théorème 55 de la page 97). Dans $\mathbb{N}$, les preuves sont peut-être plus naturelles, pas forcément plus simples ; dans $\mathbb{Z}$ elles sont plus puissantes et mettent en lumière le caractère « linéaire » de la notion de divisibilité. Ce sont les méthodes utilisées pour $\mathbb{Z}$ qui seront généralisées dans le module portant sur les anneaux de polynômes.

3.1 Division euclidienne — Numération

En général, étant donnés deux entiers a,b, $b \neq 0$, il n'est pas possible de *diviser* a par b, c'est-à-dire de trouver un entier c tel que $a = bc$. C'est pour cela qu'on introduit les nombres rationnels (fractions). En tous cas, si un tel c existe, il est unique (régularité de la multiplication). Dans ce cas, on dit que b *divise* a, ou est un *diviseur* de a, ou encore que a est *multiple* de b. Un entier est dit *pair* si c'est un multiple de 2 et *impair* sinon. On notera que, si b divise deux entiers a,a', il divise $na+n'a'$, ceci pour tous entiers n,n', ainsi que $a-a'$ si $a \geqslant a'$. Sur $\mathbb{N}^*$, la relation « x divise y » est une relation d'ordre, nous la noterons $x \mid y$ pour la distinguer de l'ordre usuel. Noter que 1 est le plus petit

élément de $\mathbb{N}^*$ mais que, par exemple, 2 et $3 = 2 + 1$ ne sont pas comparables pour la divisibilité : l'ordre obtenu n'est pas total. Si la division « exacte » n'est pas toujours possible, la *division euclidienne*, définie par le théorème suivant, la « remplace ». C'est une opération importante en arithmétique, mais aussi en informatique. Une première application en sera l'écriture d'un entier dans une base donnée.

Théorème et définition 30. Soient a, b deux entiers, on suppose $b > 0$. Il existe un unique couple d'entiers (q, r) vérifiant les conditions suivantes :

$$\text{(D1)} : \ a = bq + r \qquad \text{et} \qquad \text{(D2)} : \ 0 \leqslant r < b.$$

Les entiers q et r sont appelés respectivement *quotient* et *reste* de la division de a par b, on les note $a \operatorname{div} b$ et $a \bmod b$ respectivement.

Démonstration. Soit E l'ensemble des entiers n tels que $nb > a$. D'après la propriété d'Archimède, E n'est pas vide, soit $m := \min(E)$; bien sûr $m \neq 0$. Posons $q := m - 1$. Par définition de m, on a $qb \leqslant a < (q+1)b$. Nous pouvons donc poser $r := a - qb$, d'où $a = bq + r$. De plus $qb + r < qb + b$, donc $r < b$, vu la propriété d) de l'addition : le couple (q, r) vérifie (D1) et (D2).
Soit (q', r') un autre couple d'entiers tel que $a = bq' + r'$ et $r' < b$. Alors $bq' \leqslant a$ mais $b(q'+1) > bq' + r' = a$, autrement dit $q' + 1 \in E$ mais $q' \notin E$. Ainsi $q' + 1 = m$, soit $q' = m - 1 = q$, d'où l'on déduit $r' = r$. ∎

Exercice 5.
Cette preuve n'est pas « constructive ». Notons ici $q(a, b)$ et $r(a, b)$ le quotient et le reste de la division de a par b. Vérifier que $\big(q(a, b), r(a, b)\big)$ vaut $(0, a)$ si $a < b$ et $\big(1 + q(a - b, b), r(a - b, b)\big)$ si $a \geqslant b$. Donner une preuve « constructive » de l'existence du couple (q, r), puis écrire un algorithme permettant le calcul de q et r à partir de a et b.
Solution. Si $a < b$, on constate qu'en posant $q := 0$ et $r := a$, on a bien $a = qb + r$ et $0 \leqslant r < b$. Puisque l'unicité est acquise, on voit que $q(a, b) = 0$ et $r(a, b) = a$, comme annoncé.
Pour l'aspect « constructif » et l'algorithme qui l'incarne, se référer au module II.8.

Théorème 31 (Numération en base b). Soit b un entier, $b > 1$. Tout entier non nul x peut s'écrire de manière unique sous la forme

$$x = a_n b^n + a_{n-1} b^{n-1} + \cdots + a_1 b + a_0, \tag{7}$$

où n est un entier, $a_0, a_1, \ldots, a_n$ sont des entiers appartenant à $[0, b-1]$ et $a_n \neq 0$. On dit alors que $x = a_n a_{n-1} \ldots a_1 a_0$ est *l'écriture de x en base b*. Si la *base* b doit être précisée, on écrit $x = (a_n a_{n-1} \ldots a_1 a_0)_b$.

Démonstration. Notons $\mathcal{P}(x)$ la propriété : « x possède une écriture du type (7) ». Si $x < b$, $\mathcal{P}(x)$ est vraie : prendre $n := 0$ et $a_0 := x$. Supposons que $\mathcal{P}(y)$ soit vraie pour tous les entiers $y < x$, x étant supérieur à b. Soit $a_0 := x \bmod b$: il existe un entier q tel que $x = bq + a_0$, et $q \neq 0$ puisque $x \geqslant b > a_0$. De plus $q < bq$ (car $b > 1$

et $q > 0$), donc $q < bq \leqslant x$. Ainsi $\mathcal{P}(q)$ est vraie : q s'écrit $q = a_n b^{n-1} + \cdots + a_2 b + a_1$, où n est un entier non nul et $a_1, \ldots, a_n \in [0, b-1]$, $a_n \neq 0$. L'égalité (7) résulte alors de l'égalité $x = bq + a_0$. En conclusion, $\mathcal{P}(x)$ est vraie, et le théorème de récurrence montre que $\mathcal{P}(t)$ est vraie pour tout entier $t > 0$.

Pour l'unicité, la conclusion est évidente si $x < b$. Sinon, partant d'une écriture du type (7), posons $q := a_n b^{n-1} + \cdots + a_2 b + a_1$. Alors $x = bq + a_0$ et $0 \leqslant a_0 < b$, donc $a_0 = x \mod b$. On conclut comme ci-dessus, par récurrence sur x, grâce à la régularité de l'addition et de la multiplication. ∎

Le système décimal ($b = 10$) est celui que nous employons, le système binaire ($b = 2$) sert aux ordinateurs actuels qui utilisent aussi le système hexadécimal ($b = 16$), dans lequel les *chiffres* sont notés $0, 1, \ldots, 9$ (pourquoi pas 10 ?), puis A, B, C, D, E, F.

Exercice 6.

Soit $x := 13_{10}$ (c'est le nombre 13 en base 10). Il s'écrit $x = 1101_2$ en système binaire, car $13 = 8 + 4 + 1 = 2^3 + 2^2 + 2^0$. Écrire x en base 3 et en base 7.

Solution. On a $x = 111_3 = 16_7$.

Exercice 7.

Avec les hypothèses et notations du théorème 31 de la page précédente, montrer que n est le seul entier tel que $b^n \leqslant x < b^{n+1}$.

Solution. En effet $x \geqslant b^n$ parce que $a_n \geqslant 1$ et $a_k \geqslant 0$ pour $k = 0, \ldots, n-1$. Utilisons maintenant l'identité fondamentale :

$$z^m - 1 = (z-1)\left(z^{m-1} + z^{m-2} + \cdots + z + 1\right), \tag{8}$$

où z, m sont deux entiers et $m \geqslant 1$ (cela reste vrai si $z \in \mathbb{C}$). Puisque $a_k \leqslant b - 1$ pour $k = 0, \ldots, n$, il vient :

$$x \leqslant (b-1)\Big[b^n + b^{n-1} + \cdots + b + 1\Big] = b^{n+1} - 1 < b^{n+1}.$$

D'où $b^n \leqslant x < b^{n+1}$, ce qui *caractérise* évidemment l'entier n.

Exemple. Une application de l'écriture binaire d'un entier, importante en pratique, consiste à calculer « rapidement » une puissance x^n. On suppose que $n > 0$ est un entier ; quant à x, c'est un entier, ou un élément d'un *anneau* quelconque, par exemple l'anneau $\mathbb{Z}/m\mathbb{Z}$ où $m > 0$ est un entier (voir le module II.2). L'important est l'associativité de la multiplication et l'égalité $x^0 = 1$. *A priori*, il faut $n - 1$ multiplications pour calculer x^n, ainsi $x^4 = x \times \big(x \times (x \times x)\big)$. Il est possible de faire mieux. Par exemple :

$$x^2 = x \times x\,,\ x^4 = x^2 \times x^2\,,\ x^8 = x^4 \times x^4\,,\ x^{13} = x^8 \times (x^4 \times x)\,,$$

donc cinq multiplications suffisent pour calculer x^{13}. Dans cet exemple, nous avons utilisé, sans le dire, l'égalité $13 = 1101_2$. Généralisons. D'abord, pour tout entier k, k multiplications (élévations au carré successives) permettent de calculer x^{2^k}. Ensuite, dans l'écriture binaire de n, si nous ne conservons que les chiffres 1, nous obtenons une égalité du type :

$$n \;=\; 2^{k_p} + 2^{k_{p-1}} + \cdots + 2^{k_1} + 2^{k_0},$$

où $(k_0, k_1, \ldots, k_p)$ est une suite d'entiers strictement croissante. D'où le calcul de x^n avec $m := k_p + p$ multiplications : k_p pour calculer $x^{2^{k_p}}$ (ce qui donne au passage les autres $x^{2^{k_j}}$), puis p multiplications pour calculer le produit des $x^{2^{k_j}}$ pour $j = 0, 1, \ldots, p$. Évidemment $p \leqslant k_p$ (pourquoi ?), donc $m \leqslant 2k_p$. Par ailleurs $2^{k_p} \leqslant n$, soit $k_p \leqslant \log_2(n)$ (voir le module IV.3 pour la notation $\log_2$). Ainsi $m \leqslant 2\log_2(n)$, c'est nettement mieux que $n - 1$: vérifiez que le calcul de x^{10007} peut se faire avec 20 multiplications ! Une variante de cette méthode n'utilise pas vraiment l'écriture binaire de n, mais seulement la division (euclidienne) par 2. Par exemple, les égalités :

$$x^{13} = x \times x^{12}\,,\ x^{12} = (x^6)^2\,,\ x^6 = (x^3)^2\,,\ x^3 = x \times x^2\,,\ x^2 = x \times x$$

donnent à nouveau x^{13} à l'aide de cinq multiplications.

3.2 Nombres premiers — Factorisation des entiers

Il y a plusieurs raisons de s'intéresser aux nombres *premiers*. D'abord leur définition est naturelle (elle date de plus de 2400 ans). Ensuite ces nombres permettent d'obtenir, par multiplication, tous les autres entiers (théorème 37 de la page 86). Les polynômes *irréductibles* (*cf.* le module II.6) joueront un rôle analogue. Troisième raison : les nombres premiers tiennent aujourd'hui une place centrale en mathématiques. Quatrième raison : ces nombres sont utilisés aujourd'hui en *cryptographie* (science du codage) ; en ce sens ils influencent notre vie pratique (transmission de données, cartes bancaires, etc).

Définition 5. On appelle *nombre premier* tout entier $p > 1$ possédant exactement deux diviseurs : 1 et p. L'ensemble des nombres premiers sera noté $\mathbb{P}$.

Proposition 32. Pour qu'un entier $p > 1$ soit premier, il faut et il suffit qu'il ne soit pas produit de deux entiers strictement plus grands que 1.

Démonstration. Supposons p non premier : il possède un diviseur $a \neq 1, p$. Soit b l'entier tel que $p = ab$; $b \neq 1, p$, vu l'hypothèse sur a. Ainsi $p = ab$, et $a > 1$, $b > 1$. Réciproquement, supposons que p s'écrive $p = xy$, où x, y sont deux entiers strictement supérieurs à 1. Alors $x \mid p$, et $x \neq p$ (sinon $y = 1$), *i.e.* x est un diviseur de p distinct de 1 et p : p n'est pas premier. ∎

Théorème 33 (Euclide). Tout entier $n > 1$ est un produit (fini) de nombres premiers. En particulier, n possède au moins un diviseur premier.

Démonstration. Pour un entier $n > 1$, notons $\mathcal{P}(n)$ la propriété : « n est produit de nombres premiers ». Soit $n > 1$ un entier tel que $\mathcal{P}(k)$ soit vraie pour tout entier k tel que $1 < k < n$. Il nous suffit de prouver que $\mathcal{P}(n)$ est vraie (c'est un exemple de *récurrence forte*, *cf.* la condition (R3) de la page 64). Si n est premier, c'est un produit d'un seul nombre premier : lui-même. Sinon, $n = ab$, où a, b sont deux entiers strictement supérieurs à 1. Comme $1 < a, b < ab = n$, chacun des entiers a et b est, par hypothèse, produit de nombres premiers. Alors $n = ab$ est aussi produit de nombres premiers (« juxtaposer » les facteurs premiers de a et de b). ∎

Proposition 34.
Soit $n > 1$. Si n n'est pas premier, il possède un facteur premier p tel que $p^2 \leqslant n$.

Démonstration. Puisque $n \notin \mathbb{P}$, il s'écrit $n = ab$, où a, b sont deux entiers strictement supérieurs à 1, on peut supposer que $a \leqslant b$. Soit p un facteur premier de a (l'existence en est assurée par le théorème précédent). Alors $p \mid n$, et $p^2 \leqslant ap \leqslant ab = n$. ■

Le nombre 1 n'est pas premier, mais $2, 3, 5, 7, 11$ sont des nombres premiers. Noter que 2 est le seul nombre premier pair (pourquoi ?). Voici deux questions naturelles. Comment savoir si un entier $N > 1$ est premier ? La proposition ci-dessus donne une réponse : on dresse la liste des nombres premiers p tels que $p^2 \leqslant N$ (c'est-à-dire, si l'on utilise les réels, $p \leqslant E(\sqrt{N})$, en notant $E(x)$ la partie entière d'un réel x, voir page 518). Alors N est premier si, et seulement s'il n'est multiple d'aucun des nombres premiers obtenus.

Exemple. Le nombre $1\,789$ est premier. D'abord $E(\sqrt{1\,789}) = 42$ puisque $42^2 = 1\,764$ et $43^2 = 1\,849$. Les nombres premiers inférieurs à 42 sont :

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41.$$

Par division euclidienne, on vérifie qu'aucun de ces nombres ne divise $1\,789$.

Cette méthode marche pour $1\,789$, mais n'est pas praticable, avec ou sans ordinateur, pour un nombre de cent chiffres. On connaît aujourd'hui des méthodes beaucoup plus rapides, mais elles dépassent le cadre de cet ouvrage (*cf.* cependant, dans le module II.2, le « petit théorème de Fermat », page 147).
Une deuxième question est la suivante : si $N > 1$ est un entier donné, comment trouver *tous* les nombres premiers inférieurs à N ? Le *crible d'Ératosthène* fournit une méthode. Dans la liste des entiers de 2 à N, on supprime tous les multiples de 2, puis tous les multiples de 3, etc. Après avoir supprimé tous les multiples d'un certain entier, le plus petit des entiers qui restent dans la liste, s'il y en a, est premier, et tous les nombres premiers entre 2 et N sont obtenus successivement par ce procédé.

Exemple. Certains nombres premiers sont « remarquables ».
1) C'est le cas des *nombres de Mersenne*. Étant donnés deux entiers $a, n \geqslant 2$, le nombre $M := a^n - 1$ peut-il être premier ?
Certainement pas si $a > 2$, car la formule (8) de la page 82 montre alors que $a - 1 \mid M$, et $a-1$ est distinct de 1 (évident) et de M ($a-1 < a^n-1$). De même M n'est pas premier lorsque n lui-même n'est pas premier. En effet, on écrit alors $n = uv$, où u, v sont deux entiers strictement supérieurs à 1, et la même formule montre que a^u-1 est un diviseur de M distinct de 1 et de M. Il reste le cas des nombres $M_p := 2^p - 1$, où p est un nombre premier ; M_p est appelé *nombre de Mersenne* d'indice p. Ces nombres peuvent être premiers, comme $M_7 = 2^7 - 1 = 127$, ou non premiers, comme $M_{11} = 2^{11} - 1 = 2\,047 = 23 \times 89$. On connaît beaucoup de nombres de Mersenne premiers, mais on ne sait pas s'il y en a une infinité. Voir l'exercice II.1.12 pour une propriété de ces nombres. L'écriture binaire de M_p en base 2 ne comporte que des 1, c'est bien adapté au calcul par ordinateur !

2) De même, étant donné un entier $m \geqslant 1$, le nombre $F := 2^m + 1$ peut-il être premier ? Si m possède un facteur premier $p > 2$, F n'est pas premier. En effet, écrivons $m = pr$. La formule (8) donne, en calculant ici dans $\mathbb{Z}$:

$$2^m + 1 = -\left[(-2^r)^p - 1\right] = (2^r + 1)\left[(2^r)^{p-1} - (2^r)^{p-2} + \cdots - 2^r + 1\right],$$

ce qui fait apparaître F comme produit de deux entiers strictement supérieurs à 1 (le vérifier). Si m ne possède aucun facteur premier autre que 2, m est une puissance de 2 (théorème 33 de la page 83). Si $n \in \mathbb{N}$, le nombre $F_n := 2^{2^n} + 1$ est appelé *nombre de Fermat d'indice* n. Les nombres F_n suivants sont premiers :

$$\begin{cases} F_0 = 2^{2^0} + 1 = 3, F_1 = 2^{2^1} + 1 = 2^2 + 1 = 5, F_2 = 2^{2^2} + 1 = 2^4 + 1 = 17 \\ F_3 = 2^{2^3} + 1 = 2^8 + 1 = 257,\ F_4 = 2^{2^4} + 1 = 2^{16} + 1 = 65\,537. \end{cases}$$

En revanche $F_5 = 2^{32} + 1$ n'est pas premier (*cf.* l'exemple de la page 146), et l'on ne connaît à ce jour aucun nombre de Fermat F_n premier avec $n > 4$.

Théorème 35 (Euclide). L'ensemble $\mathbb{P}$ des nombres premiers est infini.

Démonstration. Supposons le contraire ; soit $n \geqslant 1$ le cardinal de $\mathbb{P}$. Numérotons les éléments de $\mathbb{P}$: ainsi $\mathbb{P} = \{p_1, p_2, \ldots, p_n\}$. L'entier $N := 1 + p_1 p_2 \cdots p_n$ est strictement plus grand que 1. D'après le théorème 33 de la page 83, il possède un facteur premier p. Mais $p \in \mathbb{P}$ doit être l'un des p_i, donc p divise $p_1 p_2 \cdots p_n = N - 1$. Comme p divise aussi N, il divise $1 = N - (N - 1)$. C'est absurde car $p \neq 1$, et le seul diviseur de 1 est 1. Ce raisonnement très simple est sans doute dû à Euclide (IVe siècle avant J.-C.) lui-même. ∎

Le résultat important ci-dessous sera redémontré dans la section 4.2.

Théorème 36 (Lemme d'Euclide). Si un nombre premier divise le produit de deux entiers (strictement positifs), il divise au moins l'un de ces deux entiers.
Plus généralement, si un nombre premier divise un produit d'entiers strictement positifs, il divise l'un de ces entiers.

Démonstration. Soit p premier, divisant ab mais pas a. L'ensemble A des entiers $n > 0$ tels que p divise an contient p, b et $m := \min A > 0$, mais pas 1, donc $m > 1$.

Pour tout $n \in A$, effectuons la division euclidienne $n = mq + r$, avec $0 \leqslant r < m$; alors p divise $an - (am)q = ar$. Comme $r < m$, on a $r \notin A$, d'où $r = 0$, ce qui montre que m divise n. En particulier m divise p et b. Or p est premier et $m > 1$, donc $p = m$, qui divise ainsi b.

La seconde assertion résulte de la première, en raisonnant par récurrence sur le nombre de facteurs. ∎

Exercice 8.

Soient p un nombre premier et $k \in [1, p-1]$. Montrer que $\binom{p}{k}$ est multiple de p.

Solution. Comme $\binom{p}{k} = px$, où $x = (p-1)(p-2)\cdots(p-k+1)/k!$, on pourrait dire que $p \mid \binom{p}{k}$ parce que $\binom{p}{k} = px$: p est « en facteur » dans $\binom{p}{k}$. C'est tentant, mais faux : cela reviendrait à dire que $2 \mid 3$ parce que $3 = 2\times(3/2)$! La divisibilité concerne $\mathbb{N}$ (ou $\mathbb{Z}$), mais pas $\mathbb{Q}$. Restons donc dans $\mathbb{N}$. Écrivons $k!\binom{p}{k} = p(p-1)(p-2)\cdots(p-k+1) = pt$, où cette fois $t \in \mathbb{N}$. Ainsi p divise $k!\binom{p}{k}$. Le lemme d'Euclide montre que p divise $\binom{p}{k}$ ou $k!$. La seconde possibilité est exclue : si p divise $k!$, produit de $1, 2, \ldots, k$, ce même lemme entraîne l'existence d'un entier $j \in [1, k]$ multiple de p. C'est absurde car $j \leqslant k \leqslant p-1$. Ainsi $p \mid \binom{p}{k}$.

Voici un théorème essentiel, parfois appelé « théorème fondamental de l'arithmétique ».

Théorème 37 (Factorisation). Soit n un entier strictement supérieur à 1. On peut *factoriser* n, *i.e.* l'écrire sous la forme :

$$n = p_1 p_2 \cdots p_r, \tag{9}$$

où r est un entier strictement positif et $p_1, p_2, \ldots, p_r$ sont des nombres premiers (pas nécessairement distincts). Une telle factorisation est *unique, à l'ordre des facteurs près*, on dit aussi qu'elle est *essentiellement unique*.

Démonstration. L'existence d'une factorisation de n du type (9) vient du théorème 33 de la page 83. Prouvons l'unicité par récurrence (forte). Soit $n > 1$ un entier tel que l'assertion d'unicité soit vraie pour tout entier m, $1 < m < n$. Considérons deux factorisations de n : $n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$. Il s'agit de voir que ce sont essentiellement les mêmes. Si n est premier, $r = s = 1$ et $p_1 = n = q_1$. Supposons $r \geqslant 2$, $s \geqslant 2$. Puisque le nombre premier p_1 divise $n = q_1 q_2 \cdots q_s$, le lemme d'Euclide entraîne l'existence d'un indice $j \in [1, s]$ tel que $p_1 \mid q_j$. En réordonnant les q_k, on peut supposer que $j = 1$. Les nombres p_1, q_1 sont premiers et $p_1 \mid q_1$, donc $p_1 = q_1$. En simplifiant, il vient $p_2 \cdots p_r = q_2 \cdots q_s$, notons n' ce produit. Comme $n' < n$ (car $n = p_1 n'$), les deux factorisations de n' dont nous disposons sont essentiellement les mêmes. Autrement dit d'une part $r = s$, d'autre part $p_2, \ldots, p_r$ et $q_2, \ldots, q_s$ sont les mêmes, à l'ordre près. Puisque $p_1 = q_1$, les deux factorisations de n dont nous sommes partis ne diffèrent qu'à l'ordre près. ∎

Dans une factorisation de n du type (9) on obtient, en regroupant les facteurs premiers égaux, une nouvelle écriture de n de la forme :

$$n = q_1^{e_1} q_2^{e_2} \cdots q_s^{e_s}, \tag{10}$$

où cette fois $q_1, \ldots, q_s$ sont des nombres premiers *deux à deux distincts* et les e_i sont des entiers strictement positifs. Une telle factorisation de n, ou *décomposition de n en facteurs premiers*, est unique, à l'ordre des q_i près, elle devient vraiment unique si l'on

range les q_i par ordre croissant. Ainsi la factorisation de $9\,000$ est : $9\,000 = 2^3 \times 3^2 \times 5^3$. Pour factoriser un entier n, on commence par en déterminer un facteur premier p (voir plus haut), on écrit $n = pn'$, où n' est un entier tel que $n' < n$, il reste à factoriser n', etc. En fait, factoriser un entier pose des problèmes algorithmiques non triviaux. Ce problème de factorisation est utile, notamment en cryptographie.
Voici une notation commode : si n est écrit comme en (10), et si $i \in [1, s]$, l'entier e_i est appelé *exposant du nombre premier q_i dans la décomposition de n en facteurs premiers* et noté $\nu_{q_i}(n)$. Si p est un nombre premier distinct de $q_1, \ldots, q_s$, on pose $\nu_p(n) := 0$. L'égalité (10) s'écrit alors :

$$n \;=\; \prod_{p \,\in\, \mathbb{P}} p^{\nu_p(n)}. \tag{11}$$

Le second membre de cette formule est en fait un produit *fini* (voir la section 1 du module II.2) : si $p \in \mathbb{P}$ n'est pas l'un des q_i, ce qui revient à dire que p ne divise pas n (pourquoi ?), on omet $p^{\nu_p(n)} = p^0 = 1$ dans ce produit. La notation ν_p est complétée en posant $\nu_p(1) := 0$ pour tout $p \in \mathbb{P}$. Si $n \in \mathbb{N}^*$ et $p \in \mathbb{P}$, l'entier $\nu_p(n)$ est aussi appelé *valuation de n en p*.
Soient $a, b \in \mathbb{N}^*$. En écrivant $a = \prod_{p \in \mathbb{P}} p^{\nu_p(b)}$ et $b = \prod_{p \in \mathbb{P}} p^{\nu_p(b)}$, il est clair que $ab = \prod_{p \in \mathbb{P}} p^{\nu_p(a) + \nu_p(b)}$. D'où la formule importante suivante, valable pour tous $a, b \in \mathbb{N}^*$ et tout nombre premier p :

$$\nu_p(ab) \;=\; \nu_p(a) + \nu_p(b). \tag{12}$$

Cette formule vient de l'unicité d'une factorisation du type (11) pour ab.

Théorème 38. Soient x, y deux entiers strictement positifs. Pour que x divise y, il faut et il suffit que $\nu_p(x) \leqslant \nu_p(y)$ pour *tout* nombre premier p.

Démonstration. Si $x \mid y$, écrivons $y = xz$, où $z \in \mathbb{N}^*$. Pour tout $p \in \mathbb{P}$, la formule (12) donne : $\nu_p(y) = \nu_p(x) + \nu_p(z) \geqslant \nu_p(x)$. Inversement, supposons que $\nu_p(x) \leqslant \nu_p(y)$ pour tout $p \in \mathbb{P}$. Si $p \in \mathbb{P}$, écrivons $\nu_p(y) = \nu_p(x) + t_p$, où $t_p \in \mathbb{N}$ (soustraction dans $\mathbb{N}$). Alors $y = xz$, où $z = \prod_{p \in \mathbb{P}} p^{t_p}$, donc x divise y (le produit définissant z est fini : pourquoi ?). ∎

3.3 Plus grand commun diviseur, algorithme d'Euclide

La notion de *plus grand commun diviseur* est essentielle dans $\mathbb{N}$ et $\mathbb{Z}$, comme elle le sera dans les anneaux de polynômes. *Dans ce qui suit, lorsqu'il s'agira de divisibilité, les entiers concernés seront supposés strictement positifs*. Noter que, pour deux entiers a, b, la relation $a \mid b$ implique $a \leqslant b$. Étant donnés des entiers $a_1, \ldots, a_n$, un entier est appelé *diviseur commun à $a_1, \ldots, a_n$* s'il divise chacun des a_i. L'ensemble de ces diviseurs communs est majoré par a_1, il a donc un plus grand élément (propriété (N2) de $\mathbb{N}$), qu'on pourrait appeler « pgcd » des a_i. On préfère donner la définition suivante :

Théorème et définition 39. Soient $a_1, a_2, \ldots, a_n$ $(n \geqslant 1)$ des entiers strictement positifs. Il existe un *unique* entier $d > 0$ vérifiant les deux conditions suivantes :

1. L'entier d est un diviseur commun à $a_1, a_2, \ldots, a_n$.
2. Tout autre diviseur commun à $a_1, a_2, \ldots, a_n$ divise d.

Cet entier d, appelé *plus grand commun diviseur* ou en abrégé *pgcd* de $a_1, a_2, \ldots, a_n$, est noté $\operatorname{pgcd}(a_1, \ldots, a_n)$ ou $a_1 \wedge a_2 \wedge \cdots \wedge a_n$. Il est donné explicitement, en fonction des factorisations des a_i, par la formule :

$$d = \prod_{p \in \mathbb{P}} p^{m_p}, \text{ où } m_p := \min\big(\nu_p(a_1), \ldots, \nu_p(a_n)\big) \ \forall p \in \mathbb{P}. \tag{13}$$

Démonstration. Soit d le second membre de la formule (13), c'est bien un produit fini : si $p \in \mathbb{P}$ ne divise pas a_1, $m_p = 0$. D'après le théorème 38 de la page précédente, d divise chacun des a_i. Soit de même t un diviseur commun aux a_i. Pour tout $p \in \mathbb{P}$, $\nu_p(t) \leqslant \nu_p(a_i)$ pour $i = 1, \ldots, n$, d'où $\nu_p(t) \leqslant m_p$. Ainsi $t \mid d$: d vérifie 1 et 2. Soit enfin d' un entier vérifiant les mêmes propriétés 1 et 2 que d, montrons que $d' = d$. Puisque d' est un diviseur commun aux a_i, la propriété 2 (pour d) montre que $d' \mid d$. De même $d \mid d'$, donc $d = d'$. ■

Retenons qu'un entier t divise deux entiers a et b si, et seulement s'il divise leur pgcd.

Exemple. Le pgcd de $9\,000 = 2^3 \times 3^2 \times 5^3$ et $1\,575 = 3^2 \times 5^2 \times 7$ est $3^2 \times 5^2 = 225$.

Définition 6. Deux entiers sont dits *premiers entre eux* si leur pgcd vaut 1, *i.e.* s'ils n'ont pas d'autre diviseur commun que 1. De même des entiers $a_1, a_2, \ldots, a_n$ $(n \geqslant 1)$ sont dits premiers entre eux si leur pgcd est 1.

Pour trois entiers a, b, c, il ne faut pas confondre les conditions $\operatorname{pgcd}(a, b, c) = 1$ et $a \wedge b = b \wedge c = c \wedge a = 1$. Dans le deuxième cas, on dit que a, b, c sont *premiers entre eux deux à deux*. Dans le premier cas, a, b, c sont « premiers entre eux » ; pour éviter toute confusion, on dit aussi qu'ils sont *premiers entre eux dans leur ensemble*. Ainsi, $4, 6, 9$ sont premiers entre eux dans leur ensemble, mais pas deux à deux ($4 \wedge 6 = 2$).

Sur $\mathbb{N}^*$, l'opération pgcd : $(a, b) \mapsto a \wedge b$ est associative et commutative, comme il résulte de la formule (13). Pour la même raison, le lecteur pourra vérifier que la multiplication est distributive par rapport au pgcd : $a(b \wedge c) = ab \wedge ac$ pour tous entiers a, b, c. Voici une application essentielle :

Théorème 40 (Gauß, 1777-1855). Soient a, b, c trois entiers strictement positifs. On suppose que a divise bc et que a est premier avec b. Alors a divise c.

Démonstration. Par hypothèse $a \wedge b = 1$. Ainsi $c = c(a \wedge b) = ca \wedge cb$. Or a divise ca et aussi cb, il divise donc le pgcd c de ca et cb. ■

Exercice 9.

Pour tout entier $m \geqslant 1$, notons $\varphi(m)$ le nombre d'entiers $k \in [1, m]$ premiers avec m. La fonction $\varphi : \mathbb{N}^* \to \mathbb{N}$ ainsi définie s'appelle *indicatrice d'Euler*. Démontrer, pour tout entier $n \geqslant 1$, l'égalité suivante :

$$\sum_{d|n} \varphi(d) \;=\; n,$$

la somme étant étendue à tous les diviseurs $d > 0$ de n.

Solution. L'idée est de trier les entiers $k \in [1, n]$ selon la valeur de $t = k \wedge n$. Soit E l'ensemble des couples d'entiers (d, h) tels que $d > 0$ soit un diviseur de n et $h \in [1, d]$ soit premier avec d. D'après la formule de Fubini (exemple de la page 72), la somme $\sum_{d|n} \varphi(d)$ est égale à $\operatorname{card}(E)$. Soit $f : E \to [1, n]$ l'application associant à tout couple $(d, h) \in E$ le produit th, où t est le seul entier tel que $n = td$. Il suffit de montrer que f est bijective. Soit $k \in [1, n]$. Un couple $(d, h) \in E$, avec $n = td$ comme ci-dessus, est un antécédent de k par f si, et seulement si, $k = th$. Si c'est le cas, $n \wedge k = td \wedge th = t$, d'où l'unicité de t, donc de d et h. Inversement, posons $t := k \wedge n$. On peut écrire $k = th$, $n = td$, où $h, d \in \mathbb{N}^*$ et $h \wedge d = 1$. Mais $th = k \leqslant n = td$, donc $h \in [1, d]$; alors $(d, h) \in E$ est un antécédent de k par f. Ainsi k a un unique antécédent par f, donc f est bijective.

Voir le module II.2 pour d'autres propriétés de la fonction d'Euler.

Le calcul à l'aide de la formule (13) du pgcd de deux entiers « grands », par exemple à cent chiffres chacun, est très long : c'est la factorisation de ces entiers qui est longue. Il existe une méthode beaucoup plus rapide, remontant à Euclide, elle est basée sur la *division euclidienne*.

Théorème 41 (Algorithme d'Euclide). Soient $a, b \in \mathbb{N}^*$. Il existe un unique entier $n \geqslant 1$ et deux suites d'entiers $(q_1, \ldots, q_n)$ et $(r_1, \ldots, r_{n+2})$, elles aussi uniques, qui vérifient les conditions ci-après :

1. (initialisation) : $r_1 := a$ et $r_2 := b$.
2. Pour $k = 1, \ldots, n$, on a : $r_k \;=\; q_k r_{k+1} + r_{k+2}$, égalité notée ($\mathrm{D}_k$).
3. On a : $r_2 > r_3 > \cdots > r_{n+1} > r_{n+2} := 0$.

Le pgcd de a et b est alors r_{n+1}, *dernier reste non nul*.

Démonstration. Prouvons l'existence par récurrence *sur* b. Si $b \mid a$, on écrit $a = bq_1$, avec $q_1 \in \mathbb{N}^*$, et il suffit de prendre $n := 1$, $r_1 := a$, $r_2 := b$ et $r_3 := 0$. Supposons que b ne divise pas a et que l'existence ait été démontrée pour tout entier strictement inférieur à b. On effectue la division euclidienne de a par b : $a = bq + r$, où $q \in \mathbb{N}$ et $0 \leqslant r < b$. En fait $r > 0$ car b ne divise pas a. Remplaçons le couple (a, b) par (b, r). Appliquons l'hypothèse de récurrence à (b, r) : il existe un entier $n \geqslant 2$ et deux suites $(q_2, \ldots, q_n)$, $(r_2, \ldots, r_{n+2})$ pour lesquels $r_2 = b, r_3 = r$, $0 = r_{n+2} < r_{n+1} < \cdots < r_3$, et les égalités ($\mathrm{D}_k$) soient vraies pour $k = 2, \ldots, n$. Noter le décalage effectué sur les indices. Posons alors $r_1 := a$ et $q_1 := q$. L'égalité (D_1) est donc satisfaite, et $r_3 = r < b = r_2$. Les conditions requises sont donc vérifiées pour (a, b).

Supposons maintenant que les suites $(q_1, \ldots, q_n)$ et $(r_1, \ldots, r_{n+2})$ vérifient les conditions 1, 2, 3. Soit $d := a \wedge b$, montrons que $d = r_{n+1}$. Il suffit de montrer que $d = r_k \wedge r_{k+1}$ pour $k = 1, \ldots, n$: le cas $k = n$ donnera alors $d = r_n \wedge r_{n+1}$. Or (D_n) s'écrit $r_n = q_n r_{n+1}$, donc $r_n \wedge r_{n+1} = r_{n+1}$, et ainsi $d = r_{n+1}$. Procédons par récurrence sur k. D'après 1, $r_1 \wedge r_2 = a \wedge b = d$. Supposons que $d = r_k \wedge r_{k+1}$ pour un certain entier $k \in [1, n-1]$. Soit t un diviseur de r_{k+1}. Comme $r_k = q_k r_{k+1} + r_{k+2}$, t divise r_k si, et seulement s'il divise r_{k+2}. Ainsi les diviseurs communs à r_k et r_{k+1} sont les mêmes que les diviseurs communs à r_{k+1} et r_{k+2}, d'où $r_{k+1} \wedge r_{k+2} = r_k \wedge r_{k+1} = d$: notre assertion est vraie au rang $k+1$. Pour terminer la preuve, il faut voir que les suites $(q_1, \ldots, q_n)$ et $(r_1, \ldots, r_{n+2})$ sont uniques, nous laisserons cette vérification au lecteur. ■

L'entier n est le nombre de divisions euclidiennes à effectuer pour calculer $a \wedge b$ par la méthode que cette preuve suggère (*cf.* l'exercice II.1.16). Voir une autre méthode de calcul de $a \wedge b$ dans l'exercice II.1.17.

Exercice 10.
Reprendre le calcul du pgcd de $9\,000$ et $1\,575$.
Solution. On effectue les divisions euclidiennes voulues :

$$\begin{cases} 9\,000 = 5 \times 1575 + 1125 \\ 1\,575 = 1 \times 1125 + 450 \\ 1\,125 = 2 \times 450 + \mathbf{225} \\ \quad 450 = 2 \times \mathbf{225} + 0. \end{cases}$$

Le pgcd cherché est donc $\mathbf{225}$, dernier reste non nul.

Introduisons enfin la notion de ppcm. La preuve du théorème ci-après est analogue à celle du théorème et définition 39 de la page 88, nous la laissons au lecteur.

Théorème et définition 42. Soient $a_1, a_2, \ldots, a_n$ $(n \geqslant 1)$ des entiers strictement positifs. Il existe un *unique* entier $m > 0$ vérifiant les deux conditions suivantes :

1. L'entier m est un multiple commun à $a_1, a_2, \ldots, a_n$.
2. Tout autre multiple commun à $a_1, a_2, \ldots, a_n$ est un multiple de m.

Cet entier m est appelé *plus petit commun multiple* ou *ppcm* de $a_1, a_2, \ldots, a_n$, on le note $\mathrm{ppcm}(a_1, \ldots, a_n)$ ou $a_1 \vee a_2 \vee \cdots \vee a_n$. De plus m est donné explicitement, en fonction des factorisations des a_i, par la formule :

$$m = \prod_{p \in \mathbb{P}} p^{t_p}, \text{ où } t_p := \max\big(\nu_p(a_1), \ldots, \nu_p(a_n)\big) \ \ \forall p \in \mathbb{P}. \tag{14}$$

Dans le cas de *deux* entiers, le pgcd et le ppcm sont liés (vérifier en exercice que le résultat ci-dessous ne marche plus pour trois entiers).

Proposition 43. Soient a, b deux entiers et m leur ppcm, d leur pgcd.
Alors $dm = ab$.

Démonstration. Observons que, si x, y sont deux entiers, ou d'ailleurs deux réels, alors :

$$x + y = \max(x, y) + \min(x, y).$$

Cette simple remarque va nous suffire. Soit en effet $p \in \mathbb{P}$. Vu les égalités (12), (13) et (14), il vient :

$$\begin{aligned}\nu_p(ab) = \nu_p(a) + \nu_p(b) &= \min\big(\nu_p(a), \nu_p(b)\big) + \max\big(\nu_p(a), \nu_p(b)\big)\\ &= \nu_p(d) + \nu_p(m) = \nu_p(dm).\end{aligned}$$

Ainsi, pour tout $p \in \mathbb{P}$, les exposants de p dans ab et dm sont égaux. D'où la conclusion, compte tenu de la formule (11), appliquée à ab et dm. ■

Théorème 44. Soient $a_1, \ldots, a_n \in \mathbb{N}^*(n \geqslant 2)$ des entiers premiers entre eux *deux à deux*. Leur ppcm est alors égal à leur produit.

Démonstration. Soient m le ppcm et M le produit des a_i. Il suffit de voir que, pour tout $p \in \mathbb{P}$, $\nu_p(m) = \nu_p(M)$. Or $\nu_p(M)$ est la somme des $\nu_p(a_i), i = 1, \ldots, n$ par l'égalité (12) et $\nu_p(m)$ est le maximum des $\nu_p(a_i)$ d'après (14). L'idée est que, parmi les entiers $\nu_p(a_i)$, un seul au plus est non nul (alors le maximum et la somme sont évidemment égaux). Supposons en effet le contraire, et soient $i, j \in [1, n]$ tels que $i \neq j$ et $\nu_p(a_i) \geqslant 1$, $\nu_p(a_j) \geqslant 1$: p divise a_i et a_j, contrairement à l'hypothèse $a_i \wedge a_j = 1$. ■

Le ppcm de $4, 6, 9$ est 36 et leur produit est 216. Pourtant $4, 6, 9$ sont premiers entre eux. Quelle explication donner ?

4 Entiers relatifs

4.1 Opérations sur les entiers relatifs

Dans $\mathbb{N}$, l'opération inverse de l'addition (soustraction) n'est pas toujours définie : $a - b$ n'a un sens dans $\mathbb{N}$ que si $a \geqslant b$. L'introduction des entiers relatifs $(\ldots, -2, -1, 0, 1, 2, \ldots)$ pallie ce défaut. Les entiers relatifs forment un ensemble noté $\mathbb{Z}$, qui contient $\mathbb{N}$. Le lecteur sait déjà effectuer sur les entiers relatifs les trois opérations standard : addition, soustraction et multiplication, et il sait comparer deux entiers relatifs, *i.e.* utiliser la relation d'ordre usuelle sur $\mathbb{Z}$. Précisons alors les objectifs de la présente section :

1. Rappeler sans démonstration les principales propriétés des opérations standard et de la relation d'ordre sur $\mathbb{Z}$. Ces propriétés sont indispensables, elles seront librement utilisées dans cet ouvrage.
2. Proposer une *construction rigoureuse* de l'ensemble $\mathbb{Z}$ à partir de $\mathbb{N}$, puis des trois opérations et de la relation d'ordre sur $\mathbb{Z}$.

La construction de $\mathbb{Z}$ est basée sur la notion, assez abstraite, de quotient d'un ensemble par une relation d'équivalence (*cf.* le module I.1). Cette *construction* ne sera pas utilisée dans le reste de cet ouvrage. Le lecteur peut donc s'il le désire omettre en première lecture les passages concernant cette construction. Précisons d'abord la terminologie :

Les éléments de $\mathbb{Z}$ sont appelés entiers relatifs, ou rationnels. Lorsqu'il n'y aura pas de confusion possible, nous dirons simplement « entier » au lieu de « entier relatif ». Par contre les éléments de $\mathbb{N}$ seront désormais appelés entiers naturels (ou positifs).

L'addition de $\mathbb{Z}$ est une application $(a, b) \mapsto a + b$ de $\mathbb{Z} \times \mathbb{Z}$ dans $\mathbb{Z}$. Ses principales propriétés sont rappelées ci-après ; certaines (par exemple l'associativité) seront étudiées de façon plus générale dans le module suivant.

Proposition 45. L'addition de $\mathbb{Z}$ a les propriétés suivantes (rappel : $\mathbb{Z}$ *contient* $\mathbb{N}$) :

1. Elle est *associative* et commutative.
2. Pour tout $t \in \mathbb{Z}$, $0 + t = t$.
3. Soit $t \in \mathbb{Z}$. Il existe un unique élément $t' \in \mathbb{Z}$ tel que $t + t' = 0$.
   Cet élément t' est noté $-t$ et appelé *opposé* de t.
4. Pour tous $x, y \in \mathbb{Z}$, $-(-x) = x$ et $-(x + y) = (-x) + (-y)$.
5. L'application $n \mapsto -n$ est une bijection de $\mathbb{N}^*$ sur le complémentaire de $\mathbb{N}$ dans $\mathbb{Z}$. En clair, tout élément de $\mathbb{Z}$ est soit un élément de $\mathbb{N}$, soit de la forme $-n$, où $n \in \mathbb{N}^*$.
6. L'addition de $\mathbb{Z}$ prolonge celle de $\mathbb{N}$: si $x, y \in \mathbb{N}$, la somme $x + y$ est la même, qu'on la calcule dans $\mathbb{N}$ ou dans $\mathbb{Z}$.
7. *Régularité* : pour tous $a, b, c \in \mathbb{Z}$, $a + c = b + c$ implique $a = b$.

C'est la propriété 3 qui distingue $\mathbb{Z}$ de $\mathbb{N}$ en ce qui concerne l'addition.
Les propriétés 1, 2, 3 se traduisent en disant que $\mathbb{Z}$*, muni de l'addition, est un groupe commutatif*, *cf.* le module II.2.
À partir de l'addition, on peut définir la *soustraction* dans $\mathbb{Z}$: si $x, y \in \mathbb{Z}$, on pose : $x - y := x + (-y)$. D'où par exemple $-(x - y) = y - x$. Tout entier relatif x peut s'écrire sous la forme $x = a - b$, où $a, b \in \mathbb{N}$ (pourquoi ?), mais cette écriture n'est pas unique ($-3 = 2 - 5 = 5 - 8$). Voici une application :

Exercice 11.
Montrer que l'ensemble $\mathbb{Z}$ est dénombrable.
Solution. L'application $(a, b) \mapsto a - b$ de $\mathbb{N}^2$ dans $\mathbb{Z}$ étant surjective, $\mathbb{Z}$ est fini ou dénombrable parce que $\mathbb{N}^2$ est dénombrable (voir le théorème 23 de la page 75). Mais $\mathbb{Z}$ n'est pas fini, car il contient $\mathbb{N}$.

La multiplication de $\mathbb{Z}$ est une application $(a, b) \mapsto ab = a \times b$ de $\mathbb{Z} \times \mathbb{Z}$ dans $\mathbb{Z}$. Voici ses principales propriétés :

Proposition 46. La multiplication de $\mathbb{Z}$ possède les propriétés suivantes :

1. Elle est *associative* et *commutative*.
2. Elle est *distributive* par rapport à l'addition :
$$a(b + c) = ab + ac \qquad \text{pour tous } a, b, c \in \mathbb{Z}.$$
3. Pour tout $t \in \mathbb{Z}$, $1 \times t = t$.
4. La multiplication de $\mathbb{Z}$ prolonge celle de $\mathbb{N}$: si $x, y \in \mathbb{N}$, le produit xy est le même, qu'on le calcule dans $\mathbb{N}$ ou dans $\mathbb{Z}$.
5. Soient $a, b \in \mathbb{Z}$. Alors $ab = 0$ si, et seulement si, $a = 0$ ou $b = 0$.
6. Soient $a, b \in \mathbb{Z}$. Alors $ab = 1$ si, et seulement si, $a = b = \pm 1$.

Les propriétés 1, 2, 3, jointes aux propriétés de l'addition, signifient que $\mathbb{Z}$, *muni de l'addition et de la multiplication, est un anneau commutatif*, *cf.* le module II.2. La propriété 5 signifie que $\mathbb{Z}$ *est un anneau intègre*.
Voici enfin les propriétés de la relation d'ordre usuelle sur $\mathbb{Z}$:

Théorème 47. La relation d'ordre sur $\mathbb{Z}$ a les propriétés suivantes.

1. Elle est *totale* : si $x, y \in \mathbb{Z}$, on a soit $x \leqslant y$ soit $y \leqslant x$.
2. Pour tout $x \in \mathbb{Z}$, on a $0 \leqslant x$ si, et seulement si, $x \in \mathbb{N}$.
3. Elle prolonge celle de $\mathbb{N}$: si $x, y \in \mathbb{N}$, $x \leqslant y$ a le même sens dans $\mathbb{Z}$ et dans $\mathbb{N}$.
4. Elle est *compatible avec l'addition* : si $a, a', b, b' \in \mathbb{Z}$ et si $a \leqslant a'$ et $b \leqslant b'$, on a : $a + b \leqslant a' + b'$.
5. Pour deux éléments quelconques x, y de $\mathbb{Z}$, $(x \leqslant y) \Longleftrightarrow (-y \leqslant -x)$.
6. Soient $a, a', b \in \mathbb{Z}$. Si $a \leqslant a'$ *et* $b \geqslant 0$, on a : $ba \leqslant ba'$.
   Si de plus $b > 0$, alors $(a \leqslant a') \Longleftrightarrow (ba \leqslant ba')$ et $(a < a') \Longleftrightarrow (ba < ba')$.
7. Toute partie non vide *majorée* de $\mathbb{Z}$ a un plus grand élément.
8. Toute partie non vide *minorée* de $\mathbb{Z}$ a un plus petit élément.

Esquissons maintenant la construction de $\mathbb{Z}$. Sur l'ensemble $\mathbb{N}^2 = \mathbb{N} \times \mathbb{N}$, définissons une relation, notée $\sim$, comme suit : $(a, b) \sim (c, d)$ si $a + d = b + c$. Il s'agit d'une relation d'équivalence sur $\mathbb{N}^2$. Seule la transitivité mérite d'être vérifiée. Supposons que $(a, b) \sim (c, d)$ et $(c, d) \sim (e, f)$ $(a, b, c, d, e, f \in \mathbb{N})$. Ainsi $a + d = b + c$ et $c + f = d + e$. D'où :

$$(a + f) + c = a + (c + f) = a + (d + e) = (a + d) + e = (b + c) + e = (b + e) + c.$$

La régularité de l'addition dans $\mathbb{N}$ donne donc : $a + f = b + e$, soit $(a, b) \sim (e, f)$.
L'ensemble $\mathbb{Z}$ est *par définition* le quotient de $\mathbb{N}^2$ par cette relation d'équivalence. Notons provisoirement $[a, b]$ la classe d'un couple $(a, b) \in \mathbb{N}^2$, elle jouera plus tard le rôle de « différence » de a et b. L'application $(a, b) \mapsto [a, b]$ de $\mathbb{N}^2$ dans son quotient $\mathbb{Z}$ est donc surjective.
Que devient $\mathbb{N}$ dans cette construction ? L'application $n \mapsto [n, 0]$ de $\mathbb{N}$ dans $\mathbb{Z}$ est *injective* : si $n, n' \in \mathbb{N}$ et si $[n, 0] = [n', 0]$, on a $n + 0 = n' + 0$, soit $n = n'$. Il est alors naturel d'*identifier* chaque $n \in \mathbb{N}$ avec $[n, 0] \in \mathbb{Z}$. *Nous considérons donc désormais* $\mathbb{N}$ *comme partie de* $\mathbb{Z}$. En particulier, $0 \in \mathbb{N}$ est identifié à $[0, 0] \in \mathbb{Z}$.
Comment définir l'addition sur $\mathbb{Z}$? L'idée est simple : soient $t, t' \in \mathbb{Z}$. Si $(a, b), (a', b') \in \mathbb{N}^2$ sont tels que $t = [a, b]$ et $t' = [a', b']$, on a envie de poser :

$$t + t' \ := \ [a + a', b + b']. \qquad (*)$$

Pour justifier cette formule, il faut vérifier que le second membre ne dépend que de t, t', et pas du choix de $(a, b), (a', b')$. Soit donc (c, d) $\big($resp. $(c', d')\big)$ un autre représentant de t (resp. t'). Ainsi $(a, b) \sim (c, d)$, soit $a + d = b + c$, de même $a' + d' = b' + c'$. Alors :

$$(a + a') + (d + d') = (a + d) + (a' + d') = (b + c) + (b' + c') = (b + b') + (c + c'),$$

donc $(a + a', b + b') \sim (c + c', d + d')$. Cela justifie la formule $(*)$, qui est donc valable *pour tout* choix du représentant (a, b) $\big($resp. $(a', b')\big)$ de t (resp. t'). D'où une *addition* : $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ qui est, comme celle de $\mathbb{N}$, associative et commutative. Cette addition prolonge bien celle de $\mathbb{N}$ car, si $n, n' \in \mathbb{N}$, on a $[n, 0] + [n', 0] = [n + n', 0]$.

Soit $t \in \mathbb{Z}$, écrivons-le $t = [a, b]$, où $a, b \in \mathbb{N}$, et posons $t' = [b, a] \in \mathbb{Z}$. Alors $t + t' = [a + b, b + a] = [0, 0] = 0$, d'où la propriété 3 de la proposition 45 de la page 92. Si $a \geqslant b$, $a - b$ a un sens dans $\mathbb{N}$, et $(a - b) + b = a$, d'où $[a, b] = [a - b, 0] \in \mathbb{N}$. Si $a < b$, $b - a$ a un sens dans $\mathbb{Z}$, et $[a, b] = [0, b - a] = -[b - a, 0]$, d'où la propriété 5 de la proposition 45 de la page 92. Le lecteur vérifiera les autres propriétés de cette proposition.
Comment définir la multiplication de $\mathbb{Z}$? Soient $t, t' \in \mathbb{Z}$. Si $(a, b), (a', b') \in \mathbb{N}^2$ sont tels que $t = [a, b]$ et $t' = [a', b']$, nous avons ici envie de poser :

$$tt' := [aa' + bb', ab' + ba']. \qquad (**)$$

Le lecteur montrera que la formule $(**)$ est licite, le principe est le même que pour $(*)$. D'où une *multiplication* : $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$, qui prolonge bien celle de $\mathbb{N}$ car, si $n, n' \in \mathbb{N}$, on a $[n, 0] \times [n', 0] = [nn', 0]$.
Vérifions, à titre d'exemple, la propriété 5 de la proposition 46 de la page 92, laissant le lecteur consciencieux vérifier les autres. La suffisance est claire. Supposons a, b non nuls. Si $a, b \in \mathbb{N}^*$, le produit ab dans $\mathbb{Z}$ est le même que dans $\mathbb{N}$, il est donc non nul. Si $b \in \mathbb{N}^*$ et $a = -n$, $n \in \mathbb{N}^*$, on a $ab = (-n)b = -(nb)$ (conséquence facile de la propriété 2), donc $ab \neq 0$. Il en est de même si $a \in \mathbb{N}^*$ et $b = -n'$, $n' \in \mathbb{N}^*$. Enfin, si $a = -n$ et $b = -n'$, où $n, n' \in \mathbb{N}^*$, $ab = (-n)(-n') = nn' \neq 0$.
Comment définir la relation d'ordre sur $\mathbb{Z}$? Soient $x, y \in \mathbb{Z}$. On dit que $x \leqslant y$ si $y - x \in \mathbb{N}$. Si $x \in \mathbb{Z}$, $x - x = x + (-x) = 0 \in \mathbb{N}$, donc $x \leqslant x$: la réflexivité est vérifiée. Si $x, y, z \in \mathbb{Z}$ et $x \leqslant y \leqslant z$, $y - x$ et $z - y$ sont dans $\mathbb{N}$, leur somme $(y - x) + (z - y) = z - x$ aussi (propriété 6 de l'addition), donc $x \leqslant z$: la transitivité est vérifiée. Enfin soient $x, y \in \mathbb{Z}$ tels que $x \leqslant y$ et $y \leqslant x$: $a = y - x$ et $b = x - y$ sont deux éléments de $\mathbb{N}$ de somme nulle, ils sont donc tous les deux nuls. Ainsi $x = y$, d'où l'antisymétrie. Nous avons bien défini une relation d'ordre sur $\mathbb{Z}$.
Dans le théorème 47 de la page précédente, les propriétés 1 à 6 sont immédiates, comme le lecteur le constatera. Prouvons 8. Soient donc A une partie non vide de $\mathbb{Z}$ et m un minorant de A. Pour tout $a \in A$, on a $m \leqslant a$, soit $a - m \in \mathbb{N}$. Posons $B := \{a - m \mid a \in A\}$. C'est une partie non vide de $\mathbb{N}$, soit $b_0 := \min(B)$. Comme $b_0 \in B$, il existe un élément a_0 de A tel que $b_0 = a_0 - m$. Soit alors $a \in A$. Par définition de B, $a - m \in B$, donc $b_0 \leqslant a - m$. La propriété 4 donne : $b_0 + m \leqslant (a - m) + m = a$, soit $a_0 \leqslant a$. Ainsi a_0 est le plus petit élément de A. La propriété 7 se déduit de 8 au moyen de la bijection $x \mapsto -x$ de $\mathbb{Z}$ sur $\mathbb{Z}$ (le vérifier).

En exercice, le lecteur vérifiera que $\mathbb{Z}$ ne possède ni plus grand ni plus petit élément ; en particulier $\mathbb{Z}$ ne vérifie pas la propriété (N1) de $\mathbb{N}$.

4.2 Sous-groupes de $\mathbb{Z}$, divisibilité dans $\mathbb{Z}$

La division euclidienne dans $\mathbb{N}$ s'étend sans difficulté à $\mathbb{Z}$. Voici l'analogue du théorème et définition 30 de la page 81.

Théorème et définition 48. Soient a, b deux entiers relatifs, on suppose $b > 0$. Il existe un unique couple d'entiers relatifs (q, r) vérifiant les conditions suivantes :

$$\text{(D1)} : \ a = bq + r \qquad \text{et} \qquad \text{(D2)} : \ 0 \leqslant r < b.$$

Les entiers q et r sont appelés respectivement *quotient* et *reste* de la division de a par b, on les note $a \operatorname{div} b$ et $a \bmod b$ respectivement.

Démonstration. Prouvons l'existence de (q, r).

Le théorème et définition 30 de la page 81 l'assure si $a \geqslant 0$. Supposons donc $a < 0$. Alors $-a \in \mathbb{N}$, donc il existe $(q', r') \in \mathbb{N}^2$ tel que $-a = bq' + r'$ et $0 \leqslant r' < b$. Si $r' = 0$, il suffit de prendre $q := -q'$, $r := 0$. Si $r' > 0$, on prend $r := b - r'$ et $q := -q' - 1$.
Supposons que $bq + r = a = bq' + r'$, où q, q', r, r' sont des entiers et $0 \leqslant r, r' < b$. Si par exemple $r' \leqslant r$, on a $0 \leqslant b(q' - q) = r - r' < b$ et donc $0 \leqslant q' - q < 1$. Ainsi $q = q'$, d'où $r = r'$. ∎

Comme dans $\mathbb{N}$, étant donnés deux entiers a, b, nous dirons que a *divise* b, ou que b *est un multiple de* a, s'il existe un entier c tel que $b = ac$ (notation : $a \mid b$). Si de plus $b \neq 0$, on aura $|a| \leqslant |b|$ (voir plus loin un rappel sur les valeurs absolues). Observer que 1 a maintenant deux diviseurs : 1 et -1. Lorsque $n \in \mathbb{Z}$ est donné, l'ensemble des multiples de n sera noté $n\mathbb{Z}$:

$$n\mathbb{Z} := \Big\{ kn \mid k \in \mathbb{Z} \Big\}.$$

Bien sûr $0\mathbb{Z} = \{0\}$. Si $n \in \mathbb{Z}$ n'est pas nul, $k \mapsto kn$ est une bijection de $\mathbb{Z}$ sur $n\mathbb{Z}$. La proposition suivante est importante, bien qu'évidente.

Proposition 49. Étant donnés $a, b \in \mathbb{Z}$, a divise b si, et seulement si, $b\mathbb{Z} \subset a\mathbb{Z}$ (faire attention à l'ordre). De plus $a\mathbb{Z} = b\mathbb{Z}$ si, et seulement si, $a = \pm b$.

Démonstration. Si $a \mid b$, soit $c \in \mathbb{Z}$ tel que $b = ca$. Pour tout $k \in \mathbb{Z}$, $kb = (kc)a \in a\mathbb{Z}$, d'où $b\mathbb{Z} \subset a\mathbb{Z}$. Si $b\mathbb{Z} \subset a\mathbb{Z}$, en particulier $b = b \times 1 \in a\mathbb{Z}$. Il existe donc un entier k tel que $b = ka$, donc $a \mid b$. Cela prouve la première équivalence. Nous laissons la seconde au lecteur. ∎

Nous verrons que les parties $n\mathbb{Z}$ de $\mathbb{Z}$ jouent un rôle important. Donnons d'abord une définition, qui sera généralisée dans le module II.2.

Définition 7. Une partie C de $\mathbb{Z}$ est appelée un *sous-groupe* (additif) de $\mathbb{Z}$ si elle vérifie les conditions ci-dessous.

1. Elle n'est pas vide.
2. Elle est *additivement stable* : la somme de deux éléments quelconques de C appartient à C.
3. L'opposé de tout élément de C appartient aussi à C.

Théorème 50.

a) Pour tout entier n, $n\mathbb{Z}$ est un sous-groupe de $\mathbb{Z}$.

b) Réciproquement, soit C un sous-groupe de $\mathbb{Z}$. Il existe un *unique* entier $n \geqslant 0$ tel que $C = n\mathbb{Z}$.

Démonstration. L'assertion a) est triviale. Pour b), soit C un sous-groupe de $\mathbb{Z}$. Il contient 0 : soit en effet $a \in C$. D'après la propriété 3, $-a \in C$, donc $0 = a + (-a) \in C$ (propriété 2). Si $C = \{0\}$, il suffit de prendre $n := 0$. Sinon, soit $b \in C, b \neq 0$.

Quitte à remplacer b par $-b$, qui appartient aussi à C, nous pouvons supposer $b > 0$. Ainsi $C \cap \mathbb{N}^*$ est une partie non vide de $\mathbb{N}^*$, elle possède un plus petit élément n. Montrons que $C = n\mathbb{Z}$. L'inclusion $n\mathbb{Z} \subset C$ est claire : soit $k \in \mathbb{Z}$. Si $k = 0$, $kn = 0 \in C$. Si $k > 0$, $kn = n + n + \cdots + n$ appartient à C par application répétée de la stabilité de C par addition. Si $k < 0$, kn est l'opposé de $(-k)n$, qui appartient à C (cas précédent), donc $kn \in C$ (propriété 3). L'inclusion inverse $C \subset n\mathbb{Z}$ résulte du théorème et définition 48 de la page 94. Soit en effet $a \in C$. Effectuons la division euclidienne de a par n : $a = nq + r$, où $q, r \in \mathbb{Z}$ et $0 \leqslant r < n$. D'abord $nq \in n\mathbb{Z} \subset C$, et $a \in C$, donc $r = a - qn \in C$ (propriétés 2 et 3). Si r n'était pas nul, il appartiendrait à $C \cap \mathbb{N}^*$, on aurait donc $n = \min(C \cap \mathbb{N}^*) \leqslant r$, ce qui est faux. Ainsi $r = 0$, d'où $a = nq \in n\mathbb{Z}$. ■

Définition 8. Soient A, B deux parties de $\mathbb{Z}$. On appelle *somme* de A et B l'ensemble des entiers de la forme $a + b$, où $a \in A$ et $b \in B$.

Cette « addition » de parties de $\mathbb{Z}$ est clairement associative et commutative.

Proposition 51. La somme de deux sous-groupes de $\mathbb{Z}$ est un sous-groupe de $\mathbb{Z}$.

Démonstration. Soient C, C' deux sous-groupes de $\mathbb{Z}$. D'abord $0 \in C$ et $0 \in C'$, donc $0 = 0 + 0 \in C + C'$. Soit $x \in C + C'$. Il existe $(a, a') \in C \times C'$ tel que $x = a + a'$. Alors $-x = (-a) + (-a') \in C + C'$, car $-a \in C$ (propriété 3) et de même $-a' \in C'$. La stabilité de $C + C'$ par addition est évidente. ■

Noter que $C + C'$ contient C (et C') : si $x \in C$, $x = x + 0 \in C + C'$ car $0 \in C'$. Inversement, tout sous-groupe D de $\mathbb{Z}$ contenant C et C' contient $C + C'$. Soit en effet $x \in C + C'$: $x = a + a'$, où $a \in C, a' \in C'$. Alors a et a' appartiennent à D, qui est stable par addition, donc $x = a + a' \in D$:

PROPRIÉTÉ. *La somme $C + C'$ est, au sens de l'inclusion, le plus petit sous-groupe de $\mathbb{Z}$ contenant C et C'.*

Voici un résultat capital, montrant que la notion de pgcd est plus simple à utiliser dans $\mathbb{Z}$ que dans $\mathbb{N}^*$, à cause de son caractère « linéaire ».

Théorème 52 (Plus grand commun diviseur dans $\mathbb{Z}$).
Soient a et b deux entiers relatifs *quelconques*.

1. Il existe un unique entier $d \geqslant 0$ tel que $a\mathbb{Z} + b\mathbb{Z}$ soit égal à $d\mathbb{Z}$.
2. L'entier d divise a et b. Réciproquement, tout diviseur commun à a et b divise d.
3. Il existe deux entiers u, v vérifiant la *relation de Bézout* suivante :
$$au + bv = d. \tag{15}$$
4. Si $a, b \in \mathbb{N}^*$, d est le pgcd de a et b au sens du théorème et définition 39 de la page 88.

Démonstration. L'assertion 1 résulte de la proposition précédente et du théorème 50 de la page précédente. Ensuite $a\mathbb{Z} \subset a\mathbb{Z} + b\mathbb{Z} = d\mathbb{Z}$, donc $d \mid a$, de même $d \mid b$. Soit t un divi-

seur commun à a et b. Ainsi $t\mathbb{Z}$ *contient* $a\mathbb{Z}$ et $b\mathbb{Z}$, donc aussi leur somme $d\mathbb{Z}$, autrement dit $t \mid d$: d'où 2. Puisque $d \in d\mathbb{Z} = a\mathbb{Z} + b\mathbb{Z}$, d est somme d'un élément de $a\mathbb{Z}$ et d'un élément de $b\mathbb{Z}$, d'où 3. Supposons enfin $a, b > 0$. D'abord $d > 0$ (car $a\mathbb{Z} + b\mathbb{Z} \neq \{0\}$). Ensuite il existe $c \in \mathbb{Z}$ tel que $a = cd$, d'où $c > 0$, de sorte que $d \mid a$ *dans* $\mathbb{N}$, de même $d \mid b$ dans $\mathbb{N}$. Soit t un diviseur commun à a et b dans $\mathbb{N}$. D'après 2, t divise d dans $\mathbb{Z}$, donc dans $\mathbb{N}$ (regarder les signes). Ainsi 4 résulte du théorème et définition 39 de la page 88. ∎

Compte tenu de 4, il est logique de donner la définition suivante :

Définition 9. Soient a, b deux entiers quelconques. On appelle *plus grand commun diviseur* (ou pgcd) de a et b l'unique entier $d \geqslant 0$ tel que $a\mathbb{Z} + b\mathbb{Z} = d\mathbb{Z}$, on le note aussi $a \wedge b$. On dit que a et b sont *premiers entre eux* si $a \wedge b = 1$.

Ainsi $a \wedge b$ est caractérisé par la propriété suivante :

PROPRIÉTÉ. $a \wedge b \geqslant 0$ et $a\mathbb{Z} + b\mathbb{Z} = (a \wedge b)\mathbb{Z}$.

Noter que, si $a \in \mathbb{Z}$, le pgcd de 0 et a est $|a|$. Cela dit, il est maintenant facile d'obtenir les trois résultats fondamentaux ci-dessous.

Théorème 53 (Bézout, 1731-1783). Pour que deux entiers a et b soient premiers entre eux, il faut et il suffit qu'il existe deux entiers u, v tels que $au + bv = 1$.

Démonstration. Si $a \wedge b = 1$, l'assertion 3 du théorème 52 de la page précédente donne l'existence de deux entiers u, v tels que $au + bv = 1$. Inversement, supposons qu'il existe $u, v \in \mathbb{Z}$ tels que $au + bv = 1$, et soit $d := a \wedge b$. Alors $d \mid au$ et $d \mid bv$, donc d divise $1 = au + bv$. Comme $d \geqslant 0$, nécessairement $d = 1$. ∎

Théorème 54 (Gauß). Soient $a, b, c \in \mathbb{Z}$. Si a divise bc et est premier avec b, il divise c.

Démonstration. Comme $a \wedge b = 1$, il existe $u, v \in \mathbb{Z}$ tels que $au + bv = 1$. Or $a \mid a(cu)$, $a \mid v(bc)$ (car $a \mid bc$), donc $a \mid a(cu) + v(bc) = c(au + bv) = c$. ∎

Théorème 55 (Lemme d'Euclide). Si un nombre *premier* divise le produit de deux entiers relatifs, il divise l'un (au moins) de ces deux entiers.

Démonstration. Soit p un nombre premier, supposons que p divise un produit ab ($a, b \in \mathbb{Z}$). Supposons que p ne divise pas b, et montrons que p divise c. Soit $d := p \wedge b$. Alors $d > 0$, et d divise p dans $\mathbb{N}$, donc d vaut 1 ou p, car p est premier. Le cas $d = p$ est exclu : p diviserait alors b. Ainsi $d = 1$, *i.e.* p est premier avec b. On conclut grâce au théorème de Gauß. ∎

Remarquer la simplicité de ces démonstrations. Dans $\mathbb{Z}$, le lemme d'Euclide découle du théorème de Gauß, alors que dans $\mathbb{N}$ c'était l'inverse.

Soient $a_1, \ldots, a_n$ $(n \geqslant 3)$ des entiers relatifs. La définition de leur pgcd est simple : c'est l'unique entier $d \geqslant 0$ tel que $a_1\mathbb{Z} + \cdots + a_n\mathbb{Z} = d\mathbb{Z}$. L'associativité de l'opération pgcd résulte alors de l'associativité de l'addition de sous-groupes de $\mathbb{Z}$. Nous n'insisterons pas sur cette généralisation évidente. Comme dans $\mathbb{N}$, il faut distinguer les notions « premiers entre eux (dans leur ensemble) » et « premiers entre eux deux à deux ».

Proposition 56. Soient $a_1, \ldots, a_n (n \geqslant 2)$ des entiers non tous nuls et d leur pgcd. Pour $i = 1, \ldots, n$, écrivons $a_i = db_i$, où $b_i \in \mathbb{Z}$ (d divise a_i). Alors $b_1, \ldots, b_n$ sont premiers entre eux (dans leur ensemble).

Démonstration. Il est clair que $d \neq 0$. Soit $t \geqslant 0$ le pgcd des b_i. Pour tout i, td divise donc a_i. Ainsi td est un diviseur commun aux a_i, et par suite $td \mid d$. Comme $d \neq 0$, t divise 1, *i.e.* $t = \pm 1$, et $t = 1$ car $t \geqslant 0$. ■

Voici un problème naturel. Soient $a, b \in \mathbb{Z}$ et $d := a \wedge b$. Il existe donc deux entiers u, v (*coefficients de Bézout*) tels que $au + bv = d$. Comment trouver *un* tel couple (u, v) ? Voyons un exemple (*cf.* aussi le module II.8) :

Exercice 12.
Nous avons vu (exemple de la page 88) que le pgcd de $9\,000$ et $1\,575$ est 225. Chercher *une* solution (u, v) de l'équation :

$$9\,000u + 1\,575v = 225 \quad (u, v \in \mathbb{Z}). \tag{$*$}$$

Solution. L'algorithme d'Euclide (exercice 10 de la page 90) donne, en partant du bas :

$$\begin{cases} 225 = 1\,125 - 2 \times 450 \\ 450 = 1\,575 - 1 \times 1\,125 \\ 1\,125 = 9\,000 - 5 \times 1\,575, \quad \text{d'où} \end{cases}$$

$$\begin{aligned} 225 &= 1\,125 - 2\big[1\,575 - 1 \times 1\,125\big] = 3 \times 1\,125 - 2 \times 1\,575 \\ &= 3\big[9\,000 - 5 \times 1\,575\big] - 2 \times 1\,575 = 3 \times 9\,000 - 17 \times 1\,575. \end{aligned}$$

Le couple $(3, -17)$ est donc solution de l'équation $(*)$.

Ce qui précède est lié à la résolution de certaines équations du premier degré à coefficients entiers. Plus précisément, on a le résultat suivant :

Proposition 57. Soient $a, b, c \in \mathbb{Z}$ trois entiers, $a \neq 0$, $b \neq 0$, et $d \geqslant 1$ le pgcd de a et b. Écrivons $a = da'$, $b = db'$, où $a', b' \in \mathbb{Z}$. Considérons l'équation :

$$ax + by = c. \tag{16}$$

1. Cette équation a une solution $(x, y) \in \mathbb{Z}^2$ si, et seulement si, $d \mid c$.
2. Si l'équation (16) a une solution $(x_0, y_0) \in \mathbb{Z}^2$, ses autres solutions sont données par $(x, y) = (x_0 + kb', y_0 - ka')$, où $k \in \mathbb{Z}$ est arbitraire.

Démonstration. Soit $(x, y) \in \mathbb{Z}^2$ une solution de (16).

Alors $d \mid c$, car $c = ax + by = d(a'x + b'y)$. Supposons inversement que $c = dc'$, où $c' \in \mathbb{Z}$. Il existe $(u, v) \in \mathbb{Z}^2$ tel que $au + bv = d$, d'où $c = c'd = a(c'u) + b(c'v)$. Ainsi $(c'u, c'v) \in \mathbb{Z}^2$ vérifie (16), d'où l'assertion 1.

Pour l'assertion 2, soit d'abord $k \in \mathbb{Z}$. Puisque $ab' = da'b' = ba'$, il vient :

$$a(x_0 + kb') + b(y_0 - ka') = (ax_0 + by_0) + k(ab' - ba') = ax_0 + by_0 = c\,,$$

donc $(x_0 + kb', y_0 - ka') \in \mathbb{Z}^2$ est solution de (16).

Soit inversement $(x, y) \in \mathbb{Z}^2$ une solution de (16) : $ax + by = c = ax_0 + by_0$, d'où :

$$a(x - x_0) = b(y_0 - y) \qquad \text{soit} \qquad da'(x - x_0) = db'(y_0 - y),$$

ou encore $a'(x - x_0) = b'(y_0 - y)$, car $d \neq 0$. Comme $a' \wedge b' = 1$, b' divise $x - x_0$ (théorème de Gauß) : il existe $k \in \mathbb{Z}$ tel que $x - x_0 = kb'$, soit $x = x_0 + kb'$.
Alors $b'(y_0 - y) = a'(x - x_0) = ka'b'$, donc $y_0 - y = ka'$.
En conclusion, $(x, y) = (x_0 + kb', y_0 - ka')$, comme désiré. ∎

Passons au ppcm. Si C et C' sont deux sous-groupes de $\mathbb{Z}$, il est clair que $C \cap C'$ est un sous-groupe de $\mathbb{Z}$. Le théorème suivant est alors évident.

Théorème et définition 58 (Plus petit commun multiple dans $\mathbb{Z}$).
Soient a et b deux entiers relatifs *quelconques*.

1. Il existe un unique entier $m \geqslant 0$ tel que $a\mathbb{Z} \cap b\mathbb{Z}$ soit égal à $m\mathbb{Z}$.
   Cet entier est appelé *ppcm* de a et b.
2. L'entier m est un multiple de a et de b.
   Réciproquement, tout multiple commun à a et b est un multiple de m.
3. Si $a, b \in \mathbb{N}^*$, m est le ppcm de a et b au sens de la page 90.

On généralise sans difficulté au cas de n éléments de $\mathbb{Z}$ quelconques $(n \geqslant 2)$.

5 Nombres rationnels

L'opération inverse de la multiplication dans $\mathbb{Z}$ (division) n'est pas toujours définie : $\dfrac{2}{3}$ n'a pas de sens dans $\mathbb{Z}$. L'introduction des nombres rationnels pallie ce défaut. Adoptons ici le même point de vue que pour $\mathbb{Z}$: la *construction* de $\mathbb{Q}$ donnée ci-après peut être omise en première lecture, l'important est de garder à l'esprit les principales propriétés de $\mathbb{Q}$. L'ensemble $\mathbb{Q}$ est donc supposé connu (ou construit), il contient $\mathbb{Z}$, et ses éléments sont appelés *nombres rationnels*. On pose $\mathbb{Q}^* := \mathbb{Q} \setminus \{0\}$ et $\mathbb{Z}^* := \mathbb{Z} \setminus \{0\}$. À tout couple $(a, b) \in \mathbb{Z} \times \mathbb{Z}^*$ correspond un nombre rationnel écrit sous forme de *fraction* $\dfrac{a}{b}$, et tout nombre rationnel s'écrit de cette manière ; par exemple, si $a \in \mathbb{Z}^*$, $a = \dfrac{a}{1}$.
Une telle écriture n'est pas unique, vu la propriété suivante :

PROPRIÉTÉ. *Si $a, c \in \mathbb{Z}$ et $b, d \in \mathbb{Z}^*$,* $\left(\dfrac{a}{b} = \dfrac{c}{d}\right) \iff (ad = bc)$.

L'addition de $\mathbb{Q}$ est une application $(x, y) \mapsto x + y$ de $\mathbb{Q} \times \mathbb{Q}$ dans $\mathbb{Q}$, en voici les principales propriétés :

Proposition 59. L'addition de $\mathbb{Q}$ a les propriétés suivantes (rappel : $\mathbb{Q}$ contient $\mathbb{Z}$) :

1. Elle est *associative* et *commutative*.
2. Pour tout $t \in \mathbb{Q}$, $0 + t = t$.
3. Si $a, c \in \mathbb{Z}$ et $b, d \in \mathbb{Z}^*$, on a : $\dfrac{a}{b} + \dfrac{c}{d} = \dfrac{ad + bc}{bd}$.
4. Soit $t \in \mathbb{Q}$. Il existe un unique $t' \in \mathbb{Q}$ tel que $t + t' = 0$. Il est noté $-t$ et appelé *opposé* de t. Si $t = \dfrac{a}{b}$ ($a \in \mathbb{Z}$, $b \in \mathbb{Z}^*$), on a $-t = \dfrac{-a}{b}$.
5. L'addition de $\mathbb{Q}$ prolonge celle de $\mathbb{Z}$: si $x, y \in \mathbb{Z}$, la somme $x + y$ est la même, qu'on la calcule dans $\mathbb{Z}$ ou dans $\mathbb{Q}$.
6. *Régularité* : pour tous $x, y, z \in \mathbb{Q}$, $x + z = y + z$ implique $x = y$.

Ainsi $\mathbb{Q}$, *muni de l'addition, est un groupe commutatif*. On définit aussi la *soustraction* sur $\mathbb{Q}$, nous n'en détaillerons pas les propriétés.
La multiplication de $\mathbb{Q}$ est une application $(x, y) \mapsto xy = x \times y$ de $\mathbb{Q} \times \mathbb{Q}$ dans $\mathbb{Q}$, dont les principales propriétés sont données ci-après.

Proposition 60. La multiplication de $\mathbb{Q}$ possède les propriétés suivantes :

1. Elle est *associative* et *commutative*.
2. Elle est *distributive* par rapport à l'addition : $a(b + c) = ab + ac$ pour tous $a, b, c \in \mathbb{Q}$.
3. Pour tout $t \in \mathbb{Q}$, $1 \times t = t$.
4. Si $a, c \in \mathbb{Z}$ et $b, d \in \mathbb{Z}^*$, on a : $\dfrac{a}{b} \times \dfrac{c}{d} = \dfrac{ac}{bd}$.
5. La multiplication de $\mathbb{Q}$ prolonge celle de $\mathbb{Z}$: si $x, y \in \mathbb{Z}$, le produit xy est le même, qu'on le calcule dans $\mathbb{Z}$ ou dans $\mathbb{Q}$.
6. Soit $t \in \mathbb{Q}^*$. Il existe un unique $t' \in \mathbb{Q}^*$, appelé *inverse* de t, tel que $tt' = 1$. Il est aussi noté t^{-1} ou $\dfrac{1}{t}$. Si $t = \dfrac{a}{b}$ $(a, b \in \mathbb{Z}^*)$, $t^{-1} = \dfrac{b}{a}$.
7. *Régularité* : soient $x, y, z \in \mathbb{Q}$. Si $xz = yz$ *et si* $z \neq 0$, on a $x = y$.

La propriété essentielle est 6. Nous dirons que $\mathbb{Q}^*$, *muni de la multiplication, est un groupe commutatif*. Nous dirons aussi (*cf.* le module II.2) que $\mathbb{Q}$, muni des lois $+$ et $\times$, *est un corps commutatif*. La définition de la *division*, à partir de la multiplication, résulte de la propriété suivante :

PROPRIÉTÉ. *Soient $u \in \mathbb{Q}$ et $t \in \mathbb{Q}^*$. L'équation $tx = u$ possède une unique solution $x \in \mathbb{Q}$, appelée* quotient *de u par t et notée $\dfrac{u}{t}$, ou u/t, ou encore ut^{-1}.*
Si $u = \dfrac{a}{b}$ et $t = \dfrac{c}{d}$, où $a \in \mathbb{Z}$ et $b, c, d \in \mathbb{Z}^*$, on a $\dfrac{u}{t} = \dfrac{ad}{bc}$.

Le théorème ci-après rappelle les propriétés principales de la relation d'ordre sur $\mathbb{Q}$.

Théorème 61. La relation d'ordre sur $\mathbb{Q}$ a les propriétés suivantes.

1. Elle est *totale* : si $x, y \in \mathbb{Q}$, on a soit $x \leqslant y$ soit $y \leqslant x$.
2. Elle prolonge celle de $\mathbb{Z}$: si $x, y \in \mathbb{Z}$, $x \leqslant y$ a le même sens dans $\mathbb{Z}$ et dans $\mathbb{Q}$.
3. Elle est *compatible avec l'addition* : si $x, x', y, y' \in \mathbb{Q}$ et si $x \leqslant x'$ et $y \leqslant y'$, on a $x + y \leqslant x' + y'$.
4. Pour deux éléments quelconques x, y de $\mathbb{Q}$, $(x \leqslant y) \iff (-y \leqslant -x)$.
5. *Comportement vis-à-vis de la multiplication* : soient $t, x, y \in \mathbb{Q}$.
   - Si $t > 0$: $(x \leqslant y) \iff (tx \leqslant ty)$ et $(x < y) \iff (tx < ty)$.
   - Si en revanche $t < 0$: $(x \leqslant y) \iff (tx \geqslant ty)$ et $(x < y) \iff (tx > ty)$.
6. Elle vérifie la « règle des signes » : si $x, y \in \mathbb{Q}^*$, chacun des nombres xy, $\dfrac{x}{y}$ est strictement positif si x, y sont de même signe et strictement négatif sinon.

Lorsqu'on parle d'intervalles, il faut préciser s'il s'agit d'intervalles de $\mathbb{Z}$, de $\mathbb{Q}$ (ou plus tard de $\mathbb{R}$). À cet égard, voici une notation commode, désormais utilisée dans tout le livre : *Lorsque $a, b \in \mathbb{Q}$, l'intervalle $[a, b]$ (par exemple) désigne l'ensemble des nombres rationnels x tels que $a \leqslant x \leqslant b$ (il en sera de même avec les nombres réels). Si $a, b \in \mathbb{Z}$, l'ensemble des entiers x tels que $a \leqslant x \leqslant b$ sera noté $[\![a, b]\!]$, pour le distinguer de l'intervalle $[a, b]$ de $\mathbb{Q}$.*

L'ordre de $\mathbb{Q}$ est beaucoup plus complexe que celui de $\mathbb{Z}$. Pour s'en convaincre, il suffit de réfléchir à l'assertion 1 du théorème ci-après. Les exercices II.1.22 et II.1.23 donnent aussi une idée de la complexité de cet ordre.

Théorème 62.

1. Soient $a, b \in \mathbb{Q}$ tels que $a < b$. L'intervalle $]a, b[$ est infini ; il est majoré et minoré, et il n'a ni élément maximum ni élément minimum.
2. Soient $a, b \in \mathbb{Q}$ avec $b > 0$.
   Il existe un entier $n \in \mathbb{N}$ tel que $nb > a$ (*propriété d'Archimède*).

Démonstration. D'abord $]a, b[$ n'est pas vide : il contient $(a+b)/2$ (c'est l'intérêt de la division : dans $\mathbb{Z}$, $]\!]0, 1[\![$ est vide !). Si cet intervalle était fini, il aurait un élément maximum c, car l'ordre de $\mathbb{Q}$ est total. C'est absurde, car $a < c < (b + c)/2 < b$.

De même $]a, b[$ n'a pas d'élément minimum, d'où l'assertion 1.

Pour 2, la conclusion est claire si $a \leqslant 0$: $n := 0$ convient. Supposons $a > 0$. On peut écrire $a = u/c$ et $b = v/c$, où $u, v \in \mathbb{Z}$ et $c \in \mathbb{Z}^*$ (réduire au même dénominateur). Quitte à changer c de signe, on peut supposer $c > 0$, d'où $u > 0, v > 0$ car $a, b > 0$. D'après la propriété d'Archimède dans $\mathbb{N}$, il existe un $n \in \mathbb{N}^*$ tel que $nv > u$. Puisque $c \geqslant 1$, la propriété 5 du théorème 61 implique $nb > a$, d'où l'assertion 2. ∎

Esquissons ici la *construction* de $\mathbb{Q}$ puis de l'addition, de la multiplication et de la relation d'ordre sur $\mathbb{Q}$. Les idées sont les mêmes que pour la construction de $\mathbb{Z}$ à partir de $\mathbb{N}$. L'ensemble $\mathbb{Z}^*$ des entiers relatifs non nuls est *stable par multiplication* : le produit de deux entiers non nuls est non nul. Sur l'ensemble $\mathbb{Z} \times \mathbb{Z}^*$, définissons une relation, notée $\sim$, comme suit : $(a, b) \sim (c, d)$

si $ad = bc$. C'est une relation d'équivalence sur $\mathbb{Z}\times\mathbb{Z}^*$ (le vérifier). L'ensemble $\mathbb{Q}$ est *par définition* le quotient de $\mathbb{Z} \times \mathbb{Z}^*$ par la relation $\sim$, et les éléments de $\mathbb{Q}$ sont appelés *nombres rationnels*. La classe d'un couple $(a, b) \in \mathbb{Z} \times \mathbb{Z}^*$ sera notée $\dfrac{a}{b}$, elle jouera le rôle de « quotient » de a par b. L'application $(a, b) \mapsto \dfrac{a}{b}$ de $\mathbb{Z} \times \mathbb{Z}^*$ dans son quotient $\mathbb{Q}$ est surjective. La propriété précédant la proposition 59 de la page 100 est vraie, par définition.
Comment définir l'addition et la multiplication sur $\mathbb{Q}$? Soient t, t' deux nombres rationnels, représentés par $\dfrac{a}{b}$, $\dfrac{a'}{b'}$ respectivement ($a, a' \in \mathbb{Z},\ b, b' \in \mathbb{Z}^*$). Posons :

$$t + t' \ := \ \frac{a}{b} + \frac{a'}{b'} \ = \ \frac{ab' + a'b}{bb'} \quad \text{et} \quad tt' \ := \ \frac{a}{b} \times \frac{a'}{b'} \ = \ \frac{aa'}{bb'}, \qquad (*)$$

après avoir vérifié que ces formules sont licites. Ces formules sont alors valables *quel que soit* le choix du représentant $\dfrac{a}{b}$ (resp. $\dfrac{a'}{b'}$) de t (resp. t'). D'où une *addition* : $+ : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$ et une multiplication $\times : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$.

Que devient $\mathbb{Z}$ dans cette construction ? L'application $a \mapsto \dfrac{a}{1}$ de $\mathbb{Z}$ dans $\mathbb{Q}$ est injective : si $a, a' \in \mathbb{Z}$ et si $\dfrac{a}{1} = \dfrac{a'}{1}$, on a $(a, 1) \sim (a', 1)$, soit $a = a'$. Nous identifions donc chaque $a \in \mathbb{Z}$ avec $\dfrac{a}{1} \in \mathbb{Q}$ et *nous considérons désormais* $\mathbb{Z}$ *comme partie de* $\mathbb{Q}$. En particulier 0 et 1 sont identifiés à $\dfrac{0}{1}$ et $\dfrac{1}{1}$ respectivement.
Les *preuves* de la proposition 59 de la page 100 et de la proposition 60 de la page 100 sont constituées de vérifications nombreuses mais triviales. Nous conseillons cependant au lecteur de faire ces vérifications, afin de se familiariser avec l'usage de la relation $\sim$. À titre d'exemple, si $a \in \mathbb{Z}$ et $b \in \mathbb{Z}^*$, on a :

$$\frac{a}{b} + \frac{-a}{b} = \frac{ab + b(-a)}{b^2} = \frac{0}{b^2} = \frac{0}{1} = 0 \quad \text{et} \quad \frac{a}{b} \times \frac{b}{a} = \frac{ab}{ba} = \frac{1}{1} = 1 (a \neq 0).$$

La *définition* de l'ordre sur $\mathbb{Q}$ est plus délicate. Soit d'abord $x \in \mathbb{Q}$. On dit que x est positif ($x \geqslant 0$) s'il s'écrit $x = \dfrac{a}{b}$, où $a \in \mathbb{Z}$, $b \in \mathbb{Z}^*$ et $ab \geqslant 0$. Dans ce cas, si $x = \dfrac{c}{d}$ où $c \in \mathbb{Z}$, $d \in \mathbb{Z}^*$, on a $ad = bc$, d'où $(ab)(cd) = a^2d^2 \geqslant 0$ et par suite $cd \geqslant 0$. Ainsi le fait que x soit positif se lit sur *n'importe quel* représentant de x. On pose $\mathbb{Q}_+ := \{x \in \mathbb{Q} \mid x \geqslant 0\}$, et aussi $\mathbb{Q}_+^* := \{x \in \mathbb{Q} \mid x > 0\}$. Il est facile de vérifier que $\mathbb{Q}_+$ est stable par addition et multiplication : si $x, y \in \mathbb{Q}_+$, on a $x + y \in \mathbb{Q}_+$ et $xy \in \mathbb{Q}_+$. Par ailleurs, pour tout $x \in \mathbb{Q}^*$, on a soit $x \in \mathbb{Q}_+^*$, soit $-x \in \mathbb{Q}_+^*$, et ces deux possibilités s'excluent mutuellement (vérifier ces assertions).
Soient maintenant $x, y \in \mathbb{Q}$. On dit que $x \leqslant y$ si $y - x \in \mathbb{Q}_+$ (c'est cohérent car alors $0 \leqslant x$ si, et seulement si, $x \in \mathbb{Q}_+$). Ce qui précède montre qu'on obtient ainsi sur $\mathbb{Q}$ une relation d'ordre. La vérification des diverses assertions du théorème 61 de la page précédente est triviale, là encore, elle sera laissée au lecteur. À titre d'exemple, soient $x, y \in \mathbb{Q}$, supposons $x > 0$ et $y < 0$. On peut donc écrire $x = \dfrac{a}{b}$, $y = \dfrac{c}{d}$, où $a, b, c, d \in \mathbb{Z}^*$ et $ab > 0$, $cd < 0$. Alors $xy = \dfrac{ac}{bd} < 0$ car $(ac)(bd) = (ab)(cd) < 0$ (dans $\mathbb{Z}$).

Passons à d'autres propriétés de $\mathbb{Q}$.

Proposition 63. L'ensemble $\mathbb{Q}$ est dénombrable.

Démonstration. L'application $(a,b) \mapsto \dfrac{a}{b}$ de $\mathbb{Z} \times \mathbb{Z}^*$ dans $\mathbb{Q}$ est surjective et $\mathbb{Z} \times \mathbb{Z}^*$ est dénombrable puisque $\mathbb{Z}$ l'est. Ainsi $\mathbb{Q}$ est fini ou dénombrable (théorème 23 de la page 75) ; mais $\mathbb{Q}$ n'est pas fini, car il contient $\mathbb{Z}$. ∎

Les résultats qui vont suivre montrent comment l'arithmétique (qui porte sur les *entiers*) peut intervenir à propos de nombres rationnels. Rappelons d'abord une notion importante, celle de *partie entière*.

Théorème et définition 64. Soit x un nombre rationnel.

1. Il existe un unique entier relatif n tel que $n \leqslant x < n+1$. Cet entier est appelé *partie entière* de x, il est noté $E(x)$ ou $\lfloor x \rfloor$. Ainsi :
$$E(x) \leqslant x < E(x)+1 \text{ et } E(x) \in \mathbb{Z}, \tag{17}$$
et ces propriétés caractérisent $E(x)$.
2. Écrivons x sous la forme $x = \dfrac{a}{b}$, où $b \in \mathbb{N}^*$ et $a \in \mathbb{Z}$. Alors $E(x)$ est le quotient dans la division euclidienne de a par b.

Démonstration. Si x est écrit sous la forme a/b comme dans l'énoncé, effectuons la division euclidienne de a par b : $a = bq + r$, où $q, r \in \mathbb{Z}$ et $0 \leqslant r < b$. Alors $bq \leqslant bq + r < b(q+1)$ d'où, en divisant par b : $q \leqslant x < q+1$. L'entier q est le seul pour lequel un tel encadrement soit vrai. En effet soit q' un autre entier tel que $q' \leqslant x < q'+1$. Alors $q \leqslant x < q'+1$, donc $q - q' \leqslant 0$, *parce que* $q - q'$ *est un entier*. Symétriquement $q' - q \leqslant 0$, d'où $q = q'$. Cela prouve à la fois 1 et 2. ∎

Ainsi $E(13/5) = 2$, mais $E(-13/5) = -3$ (le vérifier !).

Exercice 13.
Soient a, b deux nombres rationnels tels que $b - a > 1$.
Montrer que l'intervalle $]a, b[$ contient au moins un entier.
Solution. Si b est entier, $b-1$ convient. Sinon, soit $m := E(b)$. Par définition de la partie entière, $m > b-1 > a$, d'où $m \in]a, b[$ (et m est entier). Dans les deux cas, $-E(-b)-1$ convient.

Remarque. Soient a, b deux entiers, $b \neq 0$. Pour que b divise a (dans $\mathbb{Z}$), il faut et il suffit que le nombre rationnel $\dfrac{a}{b}$ soit en fait entier. C'est nécessaire, car si $b \mid a$, $a = kb$ où $k \in \mathbb{Z}$, et $\dfrac{a}{b} = \dfrac{k}{1} = k \in \mathbb{Z}$. Inversement, si $\dfrac{a}{b} = n \in \mathbb{Z}$, on a $a = nb$, donc b divise a.

Théorème et définition 65. Soit x un nombre rationnel. Il existe un unique couple d'entiers $(a, b) \in \mathbb{Z} \times \mathbb{N}^*$ tel que :

- Le nombre x soit égal à $\dfrac{a}{b}$.
- Les entiers a et b soient *premiers entre eux*.

On dit alors que $x = \dfrac{a}{b}$ est la *forme irréductible* de x. De plus, les autres représentants de x sont les $\dfrac{ka}{kb}$, où k est un entier non nul quelconque.

Démonstration. Prouvons l'existence d'un couple (a, b) ayant les propriétés requises. Il existe en tous cas des entiers $s \in \mathbb{Z}$ et $t \in \mathbb{N}^*$ tels que $x = s/t$ (après changement de signe du numérateur et du dénominateur, au besoin). Soit $d := s \wedge t$, écrivons $s = da$ et $t = db$, où $a \in \mathbb{Z}$ et $b \in \mathbb{N}^*$. Nous savons que $a \wedge b = 1$ (*cf.* la proposition 56 de la page 98). De plus $x = \dfrac{s}{t} = \dfrac{da}{db} = \dfrac{a}{b}$: le couple (a, b) convient.
Soit $(a', b') \in \mathbb{Z} \times \mathbb{Z}^*$ un autre couple représentant x : $x = a'/b' = a/b$. Ainsi $ab' = a'b$. Comme $a \wedge b = 1$ et $a \mid a'b$, le théorème de Gauß montre que $a \mid a'$. Soit donc $k \in \mathbb{Z}$ tel que $a' = ka$. Il vient $ab' = a'b = kab$, donc $b' = kb$. Cela prouve la dernière assertion du théorème. Supposons enfin qu'on ait aussi $a' \wedge b' = 1$ et $b' > 0$. Avec les notations ci-dessus, k vaut nécessairement 1, d'où $(a', b') = (a, b)$. ■

Exemple. Nous savons que $1\,575 \wedge 9\,000 = 225$. Comme $1\,575 = 7 \times 225$ et $9\,000 = 40 \times 225$, la forme irréductible de $1\,575/9\,000$ est $7/40$.

Comme autre exemple, soit p un nombre premier. Une fois les réels introduits, la proposition suivante dira que $\sqrt{p}$ est un nombre *irrationnel*.

Proposition 66. Un nombre premier p n'est jamais le carré d'un nombre rationnel.

Démonstration. Supposons le contraire : $p = x^2$, où $x \in \mathbb{Q}$. Soit a/b la forme irréductible de x. Alors $p = (a/b)^2$. De $a \wedge b = 1$, il résulte que $a^2 \wedge b^2 = 1$ (le vérifier de plusieurs façons). Ainsi a^2/b^2 est une forme irréductible de p. Mais $p = p/1$ est aussi une forme irréductible, donc $b = 1$ et $p = a^2$.
C'est absurde, car a serait un diviseur de p distinct de 1 et p. ■

Rappelons enfin ce qu'est la *valeur absolue* d'un nombre rationnel.

Définition 10.
On appelle *valeur absolue* d'un nombre rationnel x le nombre $|x| := \max(x, -x)$. Ainsi $|x| = x$ si $x \geqslant 0$ et $|x| = -x$ sinon.

Les propriétés suivantes sont connues, et elles s'étendront aux nombres réels :

- Pour tout $x \in \mathbb{Q}$, $-|x| \leqslant x \leqslant |x|$ et $(|x| = 0) \Longleftrightarrow (x = 0)$.
- Pour tous $x, y \in \mathbb{Q}$, $|xy| = |x||y|$.
- Pour tous $x, y \in \mathbb{Q}$, $|x + y| \leqslant |x| + |y|$ (*inégalité triangulaire*).
- Pour tous $x, y \in \mathbb{Q}$, $|x - y| \geqslant \big||x| - |y|\big|$.

La dernière propriété se déduit de l'inégalité triangulaire comme suit.
L'égalité $x = (x-y)+y$ donne $|x| \leqslant |x-y|+|y|$, d'où $|x-y| \geqslant |x|-|y|$. En échangeant x et y, $|x - y| \geqslant |y| - |x|$. Alors $|x - y| \geqslant \max(|x| - |y|, |y| - |x|) = \big||x| - |y|\big|$.

Exercice type corrigé

II.1.0 Soient $m, p, q \in \mathbb{N}$, $m \leqslant \min(p, q)$. Prouver la formule :

$$\sum_{k=0}^{m} \binom{p}{k} \binom{q}{m-k} = \binom{p+q}{m}.$$

Prendre un ensemble E à $p + q$ éléments, le partager en deux parties de cardinaux respectifs p et q. Dénombrer alors les parties à m éléments de E. Comment modifier la formule ci-dessus en remplaçant l'hypothèse $m \leqslant \min(p, q)$ par $m \leqslant p + q$? Application : soit n un entier. Calculer la somme $\sum\limits_{p=0}^{n} \binom{n}{p}^2$.

Solution. Il sera commode d'utiliser la notation suivante : si W est un ensemble fini et si $k \in \mathbb{N}$, l'ensemble des parties à k éléments de W sera noté $W^{|k|}$. Cela étant, soit donc E un ensemble de cardinal $p + q$, et considérons une partie A de E de cardinal p. Soit $B := E \setminus A$. Les parties A et B sont fixées. Posons $F := E^{|m|}$. Ensuite, pour tout $k \in [\![0, m]\!]$, notons F_k l'ensemble des parties $Z \in F$ telles que $A \cap Z$ soit de cardinal k. Les F_k, $k \in [\![0, m]\!]$ forment une partition de F, de sorte que :

$$\binom{p+q}{m} = \operatorname{card}(F) = \sum_{k=0}^{m} \operatorname{card}(F_k).$$

Soit $k \in [\![0, m]\!]$ un entier fixé. Il est clair que, si $X \in A^{|k|}$ et $Y \in B^{|m-k|}$, $X \cup Y \in F_k$; en outre $(X, Y) \mapsto X \cup Y$ est une bijection de $A^{|k|} \times B^{|m-k|}$ sur F_k, la bijection réciproque étant définie par $Z \mapsto (A \cap Z, B \cap Z)$. En vertu du théorème 18 de la page 73, il vient alors :

$$\operatorname{card}(F_k) = \operatorname{card}(A^{|k|}) \times \operatorname{card}(B^{|m-k|}) = \binom{p}{k}\binom{q}{m-k},$$

la dernière égalité résultant du théorème 29 de la page 78. D'où l'égalité annoncée :

$$\binom{p+q}{m} = \sum_{k=0}^{m} \binom{p}{k}\binom{q}{m-k}. \qquad (*)$$

Remplaçons l'hypothèse $m \leqslant \min(p, q)$ par l'hypothèse plus faible $m \leqslant p + q$. Le raisonnement précédent doit être modifié comme suit. L'entier k doit être tel que $A^{|k|}$ et $B^{|m-k|}$ ne soient pas

vides, c'est-à-dire $0 \leqslant k \leqslant p$ et $0 \leqslant m-k \leqslant q$, soit $\max(0, m-q) \leqslant k \leqslant \min(m,p)$. On obtient ceci :

$$\binom{p+q}{m} = \sum_{\max(0,m-q)}^{\min(m,p)} \binom{p}{k}\binom{q}{m-k}.$$

Appliquons la formule $(*)$ en prenant $m = p = q = n$. On obtient ceci :

$$\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k} = \sum_{k=0}^{n} \binom{n}{k}^2,$$

la dernière égalité venant de la propriété de symétrie $\binom{n}{k} = \binom{n}{n-k}$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

II.1.1 Soit $\mathcal{N}$ un ensemble ordonné non vide vérifiant lui aussi les propriétés (N1), (N2), (N3). Montrer, à l'aide du théorème 2 de la page 64, qu'il existe une unique bijection g de $\mathbb{N}$ sur $\mathcal{N}$ respectant l'ordre : pour tous entiers m, n, l'inégalité $m \leqslant n$ est équivalente à l'inégalité (dans $\mathcal{N}$) $g(m) \leqslant g(n)$.

II.1.2 Soit $f : \mathbb{N} \to \mathbb{N}$ une application strictement croissante. Montrer que $f(n) \geqslant n$ pour tout entier n. Que dire lorsque f est surjective ?

II.1.3 Sur l'ensemble $\mathbb{N}^2$ on définit une relation comme suit :

$$\Big((x,y) \leqslant (x',y')\Big) \Longleftrightarrow \Big((x < x') \text{ ou } \Big((x = x') \text{ et } (y \leqslant y')\Big)\Big).$$

Montrer qu'il s'agit d'une relation d'ordre. Montrer que l'ensemble $\mathbb{N}^2$ ainsi ordonné vérifie les conditions (N1) et (N3) qui ont servi à caractériser $\mathbb{N}$, mais pas la condition (N2).
Tout élément de $\mathbb{N}^2$ a-t-il un successeur ? Un prédécesseur ?
L'ordre défini ainsi sur $\mathbb{N}^2$ s'appelle l'ordre *lexicographique* .

II.1.4 *Démontrer* l'associativité de l'addition dans $\mathbb{N}$.

II.1.5 $*$

1. Montrer que les coefficients binomiaux sont des entiers, sans utiliser le théorème 29 de la page 78.
2. Soient p, q deux entiers, $p \leqslant q$. Démontrer la formule :

$$\sum_{n=p}^{q} \binom{n}{p} = \binom{q+1}{p+1}.$$

Exercices sur la section 3

II.1.6 Définir explicitement une bijection de $\mathbb{N}$ sur $\mathbb{Z}$ ainsi que la bijection réciproque.

II.1.7 Soit $n > 1$ un entier, écrit comme en (10) page 86. Montrer que le nombre de diviseurs de n est $(e_1+1)(e_2+1)\cdots(e_s+1)$.

II.1.8 Chacun sait effectuer les opérations usuelles en base 10. Écrire les tables d'addition et de multiplication en base 2. Effectuer ensuite les opérations $1011_2 + 101_2$ et $1011_2 \times 11_2$, sans passer par le système décimal.

II.1.9 En système décimal, comment voit-on simplement qu'un entier est multiple de 4 ? Montrer qu'un entier strictement positif est divisible par 9 si, et seulement si, la somme de ses chiffres décimaux l'est.

II.1.10 Soit $n \in \mathbb{N}$. Combien y a-t-il d'entiers $k \in [\![1, 10^n[\![$ dont l'écriture décimale ne comporte pas le chiffre 7 ?

II.1.11 $*$ Pour tout $k \in \mathbb{N}$, posons $F_k := 2^{2^k} + 1$ (*cf.* l'exercice 3.2 de la page 85). Montrer que, si $n, m \in \mathbb{N}$ et $n < m$, F_n divise $F_m - 2$. En déduire une nouvelle preuve du théorème 35 de la page 85.

II.1.12 $*$ Pour tout entier $n > 0$, on note $s(n)$ la somme de tous les diviseurs (positifs) de n.

1. Soient $a, b \in \mathbb{N}^*$. Si $a \wedge b = 1$, montrer que $s(ab) = s(a)s(b)$. À cette fin, établir que tout diviseur de ab s'écrit d'une manière et d'une seule comme produit d'un diviseur de a et d'un diviseur de b.
2. Soit p un nombre premier tel que $M_p = 2^p - 1$ soit premier (*cf.* l'exercice 3.2 de la page 84). On pose $N := 2^{p-1}M_p$. Vérifier que $s(N) = 2N$.
3. Réciproquement, soit $N > 0$ un entier *pair* tel que $s(N) = 2N$ (un tel nombre est dit *parfait*). On écrit $N = 2^r m$, où m est un entier impair et $r \in \mathbb{N}^*$. Montrer qu'il existe un entier t tel que $m = (2^{r+1}-1)t$ et $s(m) = 2^{r+1}t$. Prouver ensuite que $t = 1$, que $p = r+1$ est premier ainsi que $M_p = m$. Quelle est la conclusion ?

II.1.13 Soient $a, b \in \mathbb{N}^*$ tels que $a \wedge b = 1$. Si ab est un carré (d'entier !), montrer que a et b sont des carrés. Soient x, y deux entiers impairs. Montrer que $x^2 + y^2$ est divisible par 2 mais pas par 4.

II.1.14 $*$ Un triplet $(u, v, w) \in (\mathbb{N}^*)^3$ est appelé *triangle pythagoricien* si $u^2 + v^2 = w^2$. Ainsi $(3, 4, 5)$ est pythagoricien : $3^2 + 4^2 = 5^2$. D'où vient l'adjectif « pythagoricien » ? Montrer que tous les triangles pythagoriciens s'obtiennent par le procédé suivant, quitte à échanger u, v. On part de trois entiers d, a, b tels que $d \geqslant 1$ et $a > b \geqslant 1$, $a \wedge b = 1$, et l'on pose :

$$u := d(a^2 - b^2)\,, \quad v := 2dab\,, \quad w := d(a^2 + b^2). \qquad (*)$$

Se ramener au cas de triangles formés d'entiers premiers entre eux. Pour un tel triangle (u, v, w), quitte à échanger u, v, montrer que v est pair et u, w impairs, puis utiliser l'égalité $u^2 = (w+v)(w-v)$ (*cf.* l'exercice précédent).

Il est possible de traiter cet exercice géométriquement, et de comprendre ainsi d'où sort le procédé ci-dessus. Voici l'idée. Dans le plan, soit C le cercle de centre l'origine et de rayon 1 (« cercle-unité »), ayant pour équation $x^2 + y^2 = 1$. On sait « paramétrer rationnellement » ce cercle : en posant :

$$x(t) \;:=\; \frac{t^2-1}{t^2+1}\,,\;\; y(t) \;:=\; \frac{2t}{t^2+1}\,,$$

l'application $t \mapsto \big(x(t), y(t)\big)$ est une bijection de $\mathbb{R}$ sur le cercle C privé du point $(1,0)$ (il faut le vérifier). Cela étant, si $(u,v,w) \in \mathbb{N}^{*3}$, l'égalité $u^2+v^2=w^2$ équivaut à $(u/w)^2+(v/w)^2=1$, et les formules $(*)$ impliquent les suivantes : $u/w = x(t)$, $v/w = y(t)$, où $t = a/b$. Par ailleurs, il revient au même de connaître un nombre $t \in \mathbb{Q}$ ou sa forme irréductible a/b.

II.1.15 $*$ Soient $m, n \in \mathbb{N}^*$. Donner une nouvelle preuve du théorème 18 de la page 73, en montrant que l'application $(a,b) \mapsto a + (b-1)m$ est une bijection de $[\![1,m]\!] \times [\![1,n]\!]$ sur $[\![1,mn]\!]$ (la division euclidienne intervient).

II.1.16 On définit une suite d'entiers $(F_n)_{n\geqslant 0}$, appelée *suite de Fibonacci*, par $F_0 := 0$, $F_1 := 1$, et la relation de récurrence $F_{n+2} := F_{n+1} + F_n$.

1. Justifier la construction de cette suite en considérant l'application $(x,y) \mapsto (y, x+y)$ de $\mathbb{N}^2$ dans $\mathbb{N}^2$, et en appliquant le théorème 2 de la page 64.
2. Soit $m \geqslant 2$ un entier. Décrire toutes les étapes de l'algorithme d'Euclide (théorème 41 de la page 89), appliqué à $a := F_m$ et $b := F_{m-1}$.
3. Soient a, b deux entiers, $a > b \geqslant 1$. On applique l'algorithme d'Euclide à a et b, en reprenant toutes les notations du théorème 41 de la page 89. Ainsi n est le nombre de divisions euclidiennes à faire. Montrer que $r_{n+2-k} \geqslant F_{k+1}$ pour $k = 1, \ldots, n$. En particulier $b \geqslant F_{n+1}$.
4. Démontrer pour tout entier $k \geqslant 0$ la formule suivante (dans $\mathbb{R}$!) :
$$F_k \;=\; \frac{t^k - u^k}{\sqrt{5}},\;\; \text{où}\;\; t = \frac{1+\sqrt{5}}{2}\,,\;\; u = \frac{1-\sqrt{5}}{2}\,\cdot$$
5. En reprenant les hypothèses et les notations de la question 3, en déduire l'inégalité suivante :
$$n+1 \;\leqslant\; \frac{\ln\big(1+b\sqrt{5}\big)}{\ln t}\,\cdot$$
Soit m le nombre de chiffres décimaux de b. Établir l'inégalité $n \leqslant 5m$ (« théorème de Lamé »).

II.1.17 $*$ (Algorithme binaire de calcul du pgcd).
Soit f l'application $(x,y) \mapsto (x',y')$ de $\mathbb{N}^{*2}$ dans $\mathbb{N}^{*2}$ définie comme suit (dans la liste ci-dessous, chaque cas est supposé exclure les précédents) :

a) si $x = y$, $(x',y') := (x,y)$,

b) si x est pair, $(x',y') := (x/2, y)$,

c) si y est pair, $(x',y') := (x, y/2)$,

d) si $x < y$, $(x',y') := (y,x)$,

e) dans les cas restants, $(x', y') := \big(y, (x-y)/2\big)$.

On donne un couple $(a, b) \in \mathbb{N}^{*2}$.

1. Soit $\big((a_n, b_n)\big)$ la suite de $\mathbb{N}^{*2}$ définie par $(a_0, b_0) := (a, b)$ et la relation de récurrence $(a_{n+1}, b_{n+1}) := f\big((a_n, b_n)\big)$. Montrer qu'il existe un entier $k \geqslant 0$ tel que $a_k = b_k$.
   On note n le plus petit de ces entiers.
2. Montrer que $a \wedge b = 2^h a_n$, où $h \geqslant 0$ est un entier que l'on précisera.
3. Exemple numérique : $a = 9\,000$, $b = 1\,575$.
4. Écrire un programme mettant en œuvre l'algorithme suggéré.
5. Si n est comme dans la question 2, montrer que $n \leqslant 4\log_2\big(\max(a, b)\big)$, où $\log_2$ désigne le logarithme en base 2.

L'avantage essentiel de cet algorithme par rapport à l'algorithme d'Euclide, est qu'il ne comporte pas de *divisions* (opérations longues), sauf des divisions par 2 qui, pour un ordinateur, sont de simples opérations de « décalage ».

Exercices sur la section 5

II.1.18 $*$ Soient $n \geqslant 1$ et p un nombre premier. On se propose de calculer $m := \nu_p(n!)$, exposant de p dans la décomposition de $n!$ en facteurs premiers. Pour cela, on écrit d'abord n en base p :

$$n \;=\; a_r p^r + a_{r-1}p^{r-1} + \cdots + a_1 p + a_0 \;\; (a_r \neq 0).$$

1. Démontrer la formule suivante (on désigne par $E(x)$ la partie entière d'un rationnel x) :

$$m \;=\; E\left(\frac{n}{p}\right) + E\left(\frac{n}{p^2}\right) + \cdots + E\left(\frac{n}{p^r}\right).$$

   On pourra appliquer la formule de Fubini (*cf.* l'exemple de la page 72) à l'ensemble des couples $(k, t) \in [\![1, r]\!] \times [\![1, n]\!]$ tels que p^k divise t.
2. Soit $k \in [\![1, r]\!]$. Calculer $E(n/p^k)$ en fonction des chiffres a_i.
3. En déduire, à l'aide de la formule (8) page 82, la *formule de Legendre* :

$$m \;=\; \frac{n - S}{p - 1} \;\;,\;\; S \;:=\; a_r + a_{r-1} + \cdots + a_1 + a_0.$$

   Exemple : calculer m lorsque $n = 1\,000$ et $p = 2$.

II.1.19 Soient $a, b \in \mathbb{N}^*$ tels que $a \wedge b = 1$ et $t \in \mathbb{N}$. On suppose $t > ab$. Montrer que l'équation $au + bv = t$ possède au moins une solution $(u, v) \in \mathbb{N}^2$ (on pourra utiliser l'exercice 13 de la page 103).

II.1.20 $*$ Soit n un entier au moins égal à 2. On considère le nombre rationnel

$$S \;:=\; 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\cdot$$

Montrer que S n'est pas un entier. Indication : soit $m \geqslant 1$ le seul entier tel que $2^m \leqslant n < 2^{m+1}$. Mettre S sous forme irréductible $S = \dfrac{A}{B}$, puis considérer l'exposant de 2 dans la décomposition de B en facteurs premiers.

II.1.21 Soient x un nombre rationnel et $n > 0$ un entier. On suppose que x^n est entier. Montrer qu'alors x est lui-même entier.

II.1.22 $*$ Soit $n \in \mathbb{N}^*$. Considérons tous les nombres rationnels *mis sous forme irréductible*, appartenant à $[0,1]$, et dont le dénominateur est au plus égal à n. En les rangeant par ordre croissant, on obtient une suite $\mathcal{F}_n$, appelée *suite de Farey d'ordre* n. Voici par exemple $\mathcal{F}_7$:

$$\frac{0}{1}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{2}{7}, \frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \frac{1}{2}, \frac{4}{7}, \frac{3}{5}, \frac{2}{3}, \frac{5}{7}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \frac{1}{1}.$$

Montrer que, si $x = \dfrac{a}{b}$ et $y = \dfrac{c}{d}$ sont deux termes consécutifs de $\mathcal{F}_n$ $(x < y)$, on a $bc - ad = 1$. Pour cela, montrer qu'il existe un unique couple $(u,v) \in \mathbb{Z}^2$ tel que $ub - av = 1$ et $n - b < v \leqslant n$. Posant $t := \dfrac{u}{v}$, montrer ensuite que t appartient à $\mathcal{F}_n$ et $t \geqslant y$. Montrer enfin que $t = y$, en raisonnant par l'absurde ; on évaluera les différences $y - x$, $t - y$, $t - x$.
Déduire de ce qui précède que, si $x = \dfrac{a}{b}$, $y = \dfrac{c}{d}$, $z = \dfrac{e}{f}$ sont trois termes consécutifs de $\mathcal{F}_n$, on a : $y = \dfrac{a+e}{b+f}$.

II.1.23 $**$ Soit E un ensemble *dénombrable*, muni d'une relation d'ordre total, notée $\leqslant$. On fait les hypothèses suivantes :

a) L'ensemble E ne possède ni élément maximum ni élément minimum.

b) Pour tous $a, b \in E$ tels que $a < b$, l'intervalle $]a,b[$ est non vide (on dit que E « n'a pas de trous »).

Montrer qu'il existe une bijection f de $\mathbb{Q}$ sur E respectant l'ordre, *i.e.* telle que, pour tous $x, y \in \mathbb{Q}$, $(x \leqslant y) \iff \big(f(x) \leqslant f(y)\big)$ (« lemme de Cantor »).
Indication : soit D l'ensemble des nombres rationnels de la forme $a/2^n$, où $a \in \mathbb{Z}$ et $n \in \mathbb{N}$ (nombres « dyadiques ») . À partir d'une numérotation des éléments de E (i.e. une bijection de $\mathbb{N}$ sur E), on pourra construire une bijection de D sur E respectant l'ordre.

Groupes, anneaux, corps

II.2

En algèbre élémentaire, on « calcule », *i.e.* on effectue des « opérations ». Ainsi $2 + 3 = 5$ est une *addition*. Quant à *l'addition des entiers naturels*, c'est l'*application* associant à *tout* couple d'entiers naturels leur somme (et non pas l'addition de deux entiers particuliers) : l'addition devient une *application* de $\mathbb{N} \times \mathbb{N}$ dans $\mathbb{N}$, *i.e.* une *loi de composition interne sur* $\mathbb{N}$. Il y aura plus loin des *lois de composition externes* (*cf.* les espaces vectoriels).
Considérons des ensembles munis d'une ou de plusieurs lois de composition (comme l'addition et la multiplication sur $\mathbb{Z}$). En imposant à ces lois de vérifier certaines propriétés (appelées « axiomes »), on obtient une « structure algébrique ». De telles structures sont celles de « groupe », d'« anneau », de « corps » et d'« espace vectoriel ».
La structure de groupe met en jeu une loi de composition interne. Les « axiomes » des groupes ont été choisis à cause des nombreuses applications de la notion obtenue, notion introduite par E. Galois (1811-1832) afin de résoudre des problèmes portant sur la résolution des équations algébriques. Les groupes jouent un rôle central dans de nombreuses branches des mathématiques et de la physique. Ils interviennent dans les structures d'anneau, de corps et d'espace vectoriel.
Le prototype de groupe est $\mathbb{Z}$, muni de l'addition. L'important dans un groupe n'est ni la *nature* de ses éléments, ni la notation employée, mais les *règles de calcul*, résultant des axiomes des groupes. Ainsi les éléments du « groupe des permutations d'un ensemble » ne sont pas du tout des nombres.

Un anneau est muni de deux lois de composition. Le prototype d'anneau est $\mathbb{Z}$; l'anneau des polynômes (*cf.* le module II.6) sera voisin de $\mathbb{Z}$ pour son « arithmétique », mais nous verrons des exemples d'anneaux très différents.
La structure de corps, importante en algèbre linéaire, se rattache à celle d'anneau. Contrairement aux anneaux, les corps les plus utiles dans cet ouvrage sont les ensembles de nombres $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, avec leurs lois usuelles.

1 Lois de composition internes

Les lois de composition internes (dans ce module, nous dirons simplement *lois*), ont été introduites dans le module I.1. Dans cette section, nous regroupons les propriétés de ces lois qui seront utiles pour l'étude des groupes, anneaux et corps. Une loi sur un ensemble E est donc une application de $E \times E$ dans E. À tout couple $(x, y) \in E \times E$, cette loi associe un élément z de E, noté de diverses façons. Les notations les plus courantes sont :

- La notation additive : l'élément z est noté $x+y$ (*somme* de x et y), la loi est appelée *addition* et notée $+ : E \times E \to E$.
- La notation multiplicative : l'élément z est noté xy ou $x \times y$ (*produit* de x et y), la loi est appelée *multiplication* et notée $\times : E \times E \to E$.

On peut utiliser aussi des notations telles que $x \star y$, $x \top y$ ou $x \circ y$. Lorsque nous parlerons de $(E, \top)$, cela signifiera que $\top$ est une loi sur l'ensemble E.

Exemples.

1. L'addition (resp. multiplication) est une loi sur $\mathbb{N}$ (ou $\mathbb{Z}$ ou $\mathbb{Q}$).
2. Soient E un ensemble et $\mathcal{F}(E, E)$ l'ensemble des applications de E dans E. Associons à tout couple (f, g) d'éléments de $\mathcal{F}(E, E)$ la *composée* $f \circ g$.

   Cette loi sur $\mathcal{F}(E, E)$ est la *composition des applications*.
3. Soit $\mathcal{P}(X)$ l'ensemble des parties d'un ensemble X. L'*intersection* $(A, B) \mapsto A \cap B$ et la *réunion* $(A, B) \mapsto A \cup B$ sont des lois sur $\mathcal{P}(X)$.
4. Sur l'ensemble $\mathbb{N}^*$, nous disposons de la loi *pgcd* : $(a, b) \mapsto \mathrm{pgcd}(a, b)$ et de la loi *ppcm* : $(a, b) \mapsto \mathrm{ppcm}(a, b)$.
5. Soit $\star$ une loi sur X. À tout couple (A, B) de parties de X, associons $A \star B := \{x \star y \mid (x, y) \in A \times B\}$. Cela définit une loi $(A, B) \mapsto A \star B$ sur $\mathcal{P}(X)$. Si la loi de X est $+$ (resp. $\times$), on écrit $A + B$ (resp. AB).

Certaines des propriétés que peut posséder une loi ont été envisagées dans le module I.1, notamment l'*associativité* et la *commutativité*. Dans les exemples 1 à 4, les lois sont associatives. Dans l'exemple 5, si la loi $\star$ sur X est associative, la loi $\star$ sur $\mathcal{P}(X)$ l'est aussi. Dans $\mathbb{Z}$, la *soustraction* n'est pas associative (le vérifier).

Exemple.
On peut définir une loi sur un ensemble fini par sa *table* : sur $E = \{a, b\}$, la

	a	$\underline{b}$
$\underline{a}$	b	$\underline{\underline{a}}$
b	a	a

$$\underline{a} \times \underline{b} = \underline{\underline{a}}$$

table ci-contre définit une multiplication *non associative*. En effet :

$$aa = b,\ ab = a,\ ba = a,\ bb = a, \text{ d'où } (aa)b = bb = a \neq b = aa = a(ab).$$

Si une loi est *associative*, on peut *supprimer les parenthèses* : si $x, y, z \in E$, l'élément $(x \star y) \star z = x \star (y \star z)$ peut sans ambiguïté être noté $x \star y \star z$ (ou xyz, ou $x + y + z$, etc). Par récurrence sur n, on définit en général des *applications* $(x_1, \ldots, x_n) \mapsto x_1 \star \cdots \star x_n$ de E^n dans E : pour $n = 1$, on part de l'application identique de E, puis on utilise la formule récurrente :

$$x_1 \star x_2 \star \cdots \star x_n \ := \ (x_1 \star x_2 \star \cdots \star x_{n-1}) \star x_n \quad n \geqslant 2\,.$$

Si $x_1, \ldots, x_n \in E$ et $y_1, \ldots, y_m \in E$ $(n, m \geqslant 1)$, on a :

$$(x_1 \star x_2 \star \cdots \star x_n) \star (y_1 \star y_2 \star \cdots \star y_m) \ = \ x_1 \star x_2 \star \cdots \star x_n \star y_1 \star y_2 \star \cdots \star y_m, \quad (1)$$

on le montre par récurrence sur m. En notation additive, $x_1 + x_2 + \cdots + x_n$ est aussi noté $\sum\limits_{i=1}^{n} x_i$, en notation multiplicative, $x_1 x_2 \cdots x_n$ est aussi noté $\prod\limits_{i=1}^{n} x_i$. Le cas où tous les x_i sont égaux à un même élément x est important :

Définition 1. Soient E un ensemble muni d'une multiplication et $x \in E$.
La suite $(x^n)_{n \geqslant 1}$ des *puissances* de x est définie par $x^1 := x$ et la relation de récurrence $x^{n+1} := x^n \times x$. En notation additive, la suite $(nx)_{n \geqslant 1}$ des *multiples* de x est définie par $1x := x$ et la relation $(n+1)x := nx + x$.

Pour une loi associative, $nx = \overbrace{x + x + \cdots + x}^{n \text{ termes}}$, de même $x^n = \overbrace{x \times x \times \cdots \times x}^{n \text{ facteurs}}$.

Proposition 1. Soient E un ensemble muni d'une loi *associative* et $x \in E$. Pour tous entiers $m, n \geqslant 1$, on a :

$$x^m x^n = x^{m+n} \quad \text{et} \quad (x^m)^n = x^{mn} \ \text{ en notation multiplicative,} \qquad (2)$$

$$mx + nx = (m+n)x \quad \text{et} \quad m(nx) = (mn)x \ \text{ en notation additive.} \qquad (3)$$

Démonstration. La première formule (3) (par exemple) vient de l'égalité (1), et la deuxième en résulte (raisonner par récurrence sur n). ■

Partons de $(E, \star)$. On dit que $x, y \in E$ *commutent* si $x \star y = y \star x$. Si c'est vrai pour tous x, y, la loi est *commutative*.

Remarque. Considérons une loi associative, notée multiplicativement par exemple, sur un ensemble E. Soient $x, y_1, \ldots, y_n \in E$ $(n \geqslant 1)$. Si x commute avec chacun des y_i, il commute avec $y_1 y_2 \cdots y_n$.
C'est vrai si $n = 1$ et, si c'est vrai pour $n - 1$ $(n \geqslant 2)$, on a :

$$\begin{aligned} x(y_1 y_2 \cdots y_n) &= (xy_1)(y_2 \cdots y_n) = (y_1 x)(y_2 \cdots y_n) = y_1\big(x(y_2 \cdots y_n)\big) \\ &= y_1\big((y_2 \cdots y_n)x\big) = \big(y_1(y_2 \cdots y_n)\big)x = (y_1 y_2 \cdots y_n)x\,. \end{aligned}$$

On préfère réserver la notation additive à des lois *commutatives*. Dans les exemples 1, 3, 4, les lois sont commutatives. Dans l'exemple 5, si la loi $\star$ sur X est commutative, la loi $\star$ sur $\mathcal{P}(X)$ l'est aussi. La soustraction de $\mathbb{Z}$ n'est pas commutative. La loi $\circ$ (exemple 2) non plus, dès que E a deux éléments distincts a, b : les applications $x \mapsto a$ et $x \mapsto b$ ne commutent pas.

Dans cet ouvrage, les lois seront associatives, pas toujours *commutatives* (*cf.* le groupe symétrique, ou l'anneau des matrices carrées).

Proposition 2. Soit E un ensemble muni d'une multiplication (resp. addition) *associative*. Si $x, y \in E$ *commutent* (par exemple si la loi est *commutative*),

$$(xy)^n = x^n y^n \ \ (\text{resp. } n(x+y) = nx + ny) \ \ \text{pour tout entier } n \geqslant 1. \tag{4}$$

Démonstration. Dans le cas multiplicatif, $xy^k = y^k x$ pour tout $k \geqslant 1$, vu la remarque ci-dessus. Démontrons la formule (4) par récurrence sur n. Le cas $n = 1$ est évident. Si la formule est vraie au rang n, il vient :

$$(xy)^{n+1} = (xy)^n(xy) = (x^n \overbrace{y^n)(x}\, y) = (x^n \overbrace{x)(y^n}\, y) = x^{n+1}y^{n+1}. \qquad \blacksquare$$

Pour une addition associative et commutative, la somme de plusieurs éléments *ne dépend pas de l'ordre* dans lequel on effectue les additions :

Proposition 3. Soient E un ensemble muni d'une addition associative et commutative, $x_1, \ldots, x_n \in E$ $(n \geqslant 1)$ et s une *permutation* de $[\![1, n]\!]$. Alors :

$$x_1 + x_2 + \cdots + x_n \ = \ x_{s(1)} + x_{s(2)} + \cdots + x_{s(n)}. \tag{5}$$

Démonstration. Raisonnons par récurrence sur n (le cas $n = 1$ est trivial).
Supposons $n \geqslant 2$ et le résultat vrai à l'ordre $n-1$. Si $s(n) = n$, l'hypothèse de récurrence s'applique. Sinon, $n = s(k)$ pour un certain $k \in [\![1, n-1]\!]$. Vu la remarque de la page précédente, le membre de droite vaut :

$$x_{s(1)} + \cdots + x_{s(k-1)} + \left[x_n + \left(x_{s(k+1)} + \cdots + x_{s(n)}\right)\right], \quad \text{soit}$$
$$x_{s(1)} + \cdots + x_{s(k-1)} + \left[\left(x_{s(k+1)} + \cdots + x_{s(n)}\right) + x_n\right],$$

ce qui nous ramène au cas précédent. $\blacksquare$

Gardons les mêmes hypothèses. Soient I un ensemble fini de cardinal $n \geqslant 1$ et $(x_i)_{i \in I}$ une famille d'éléments de E. Par anticipation, utilisons la notation $\mathfrak{S}_n$ qui sera introduite dans le théorème et définition 8 de la page 118. Si f est une *bijection* de $[\![1, n]\!]$ sur I (*numérotation* des éléments de I), posons $T(f) := x_{f(1)} + \cdots + x_{f(n)}$. Si $g : [\![1, n]\!] \to I$ est une autre bijection, $s := f^{-1} \circ g \in \mathfrak{S}_n$, et $g = f \circ s$, donc $T(f) = T(g)$, car $x_{f(1)} + x_{f(2)} + \cdots + x_{f(n)} = x_{f(s(1))} + x_{f(s(2))} + \cdots + x_{f(s(n))}$, vu la proposition. Ainsi $T(f)$ *ne dépend pas de* f, elle est appelée *somme* de la famille $(x_i)_{i \in I}$ et notée $\sum\limits_{i \in I} x_i$.

L'indice i est *muet*, *i.e.* peut être changé $\left(\sum\limits_{i \in I} x_i = \sum\limits_{k \in I} x_k\right)$. En notation multiplicative, on écrit $\prod\limits_{i \in I} x_i$ au lieu de $\sum\limits_{i \in I} x_i$.

Les sommes ci-dessus obéissent aux règles de calcul suivantes :

Théorème 4. Soit E un ensemble muni d'une addition associative et commutative. Considérons deux ensembles finis non vides I, J.

1. Soient $(x_i)_{i\in I}$ une famille d'éléments de E et $I_1, \dots, I_m$ des parties finies non vides de I formant une *partition* de I. Si $h \in [\![1, m]\!]$, soit S_h la somme de la famille $(x_i)_{i\in I_h}$. On a la *formule d'associativité* :
$$\sum_{i\in I} x_i = S_1 + S_2 + \cdots + S_m\,, \quad \text{soit} \quad \sum_{i\in I} x_i = \sum_{h=1}^{m}\left(\sum_{i\in I_h} x_i\right). \tag{6}$$
2. Soit $(x_{i,j})_{(i,j)\in I\times J}$ une famille d'éléments de E. Alors :
$$\sum_{(i,j)\in I\times J} x_{i,j} = \sum_{i\in I}\left(\sum_{j\in J} x_{i,j}\right) = \sum_{j\in J}\left(\sum_{i\in I} x_{i,j}\right). \tag{7}$$
3. Soient $(x_i)_{i\in I}$ une famille d'éléments de E et t une bijection de J sur I. Alors :
$$\sum_{j\in J} x_{t(j)} = \sum_{i\in I} x_i \quad \text{(formule de changement d'indice).} \tag{8}$$

Démonstration. L'assertion 1 est laissée au lecteur à titre d'exercice. L'assertion 2 en résulte : lorsque i décrit I, les parties $\{i\}\times J$ forment une partition de $I\times J$, d'où la première égalité. La deuxième est analogue.
Pour l'assertion 3, soient $n := \operatorname{card}(J)$, f une bijection de $[\![1, n]\!]$ sur J, et $g := t \circ f$. Pour tout $j \in J$, posons $y_j := x_{t(j)}$. Dans la formule (8), la somme de droite vaut $x_{g(1)} + \cdots + x_{g(n)} = x_{t(f(1))} + \cdots + x_{t(f(n))}$, soit $y_{f(1)} + \cdots + y_{f(n)}$, et la somme de gauche vaut $\sum_{j\in J} y_j = y_1 + \cdots + y_n$. ∎

Partons de $(E, \top)$. Rappelons qu'un élément e de E est *élément neutre* pour $\top$ si l'on a $x\top e = x = e\top x$ pour tout $x \in E$. Un tel élément est unique *lorsqu'il existe*. On le note 0 pour une addition, 1 pour une multiplication. Ainsi, dans chacun des ensembles de nombres usuels, 0 est neutre pour l'addition et 1 est neutre pour la multiplication. Dans $\mathbb{N}^*$, 1 est neutre pour la loi « ppcm ». Vérifier qu'il n'y a pas de neutre pour la loi « pgcd ».
Soit E un ensemble. L'application identique Id_E de E est élément neutre pour la loi « composition des applications » sur $\mathcal{F}(E, E)$. Sur l'ensemble $\mathcal{P}(E)$ des parties de E, E est neutre pour la loi $\cap$ et la partie vide $\varnothing$ est neutre pour la loi $\cup$.
Soit E un ensemble muni d'une loi associative $\star$ et ayant un élément neutre e. Pour une famille finie $(x_1, \dots, x_n)$ d'éléments de E ($n \geqslant 1$), l'élément $x_1 \star \cdots \star x_n$ a été défini. Si $n = 0$, *i.e.* si la famille est vide, nous *convenons* que $x_1 \star \cdots \star x_n := e$ (soit $x_1 + \cdots + x_n := 0$ en notation additive, $x_1 \cdots x_n := 1$ en notation multiplicative). Ainsi, soit $x \in E$. Si la loi est multiplicative (resp. additive), on a $x^0 := 1$ (resp. $0x := 0$) (le 0 à gauche est un entier, le second membre est le neutre de E). Avec ces conventions, les formules (1) à (4) restent vraies pour deux entiers $m, n \geqslant 0$ (le vérifier).
Supposons de plus la loi additive et *commutative*. Soit $(x_i)_{i\in I}$ une famille finie d'éléments de E. La somme $\sum_{i\in I} x_i$ est définie si $I \neq \varnothing$. Si $I = \varnothing$, on *convient* que

cette somme vaut 0. Les formules (6), (7), (8) restent vraies si certains des ensembles d'indices intervenant sont vides. Généralisons : au lieu de supposer I fini, supposons que $I' = \{i \in I \mid x_i \neq 0\}$ soit fini ; I' est appelé *support* de la famille. *Par définition*, $\sum_{i\in I} x_i$ est alors égale à $\sum_{i\in I'} x_i$. Le théorème 4 de la page précédente reste vrai, en supposant par exemple dans la formule (7) que la famille $(x_{i,j})_{(i,j)\in I\times J}$ est *à support fini*.

Partons de $(E, \top)$, E possédant un élément neutre e pour $\top$, et soit $x \in E$. Rappelons qu'un élément x' de E est appelé *symétrique* de x si :

$$x \top x' = e = x' \top x. \tag{9}$$

Si un tel élément x' existe, (auquel cas il est unique lorsque la loi est associative), on dit que x est *symétrisable*. Par exemple, tout $x \in E$ tel que $x \top x = e$ est symétrisable : il est son propre symétrique. Un tel élément est dit *involutif*. Des exemples importants seront les « symétries ».

En notation additive, le symétrique x' d'un élément x est appelé *opposé* de x et noté $-x$. Ainsi $x + (-x) = 0 = (-x) + x$. En notation multiplicative, x' est appelé *inverse* de x et noté x^{-1}. Ainsi $x \times x^{-1} = 1 = x^{-1} \times x$; un élément possédant un inverse est dit *inversible*, au lieu de « symétrisable ».

Pour le premier des deux résultats suivants, voir le module I.1 :

Théorème 5 (Régularité). Soient E un ensemble admettant un élément neutre e pour une loi *associative* $\top$ et x un élément *symétrisable* de E. Alors x est *régulier* (ou *simplifiable*), ce qui signifie que, pour tous $y, z \in E$, on a les implications :

$$(x \top y = x \top z) \Longrightarrow (y = z) \quad \text{et} \quad (y \top x = z \top x) \Longrightarrow (y = z). \tag{10}$$

Proposition 6. Soient E un ensemble possédant un élément neutre e pour une multiplication *associative* et $x_1, \ldots, x_n$ des éléments inversibles de E ($n \geqslant 0$).
Alors $x_1 x_2 \cdots x_n$ est inversible, et son inverse est donné par la formule :

$$(x_1 x_2 \cdots x_n)^{-1} = x_n^{-1} \cdots x_2^{-1} x_1^{-1} \quad (\textit{remarquer l'ordre des facteurs}).$$

Démonstration. Par récurrence sur n (les cas $n = 0$ et $n = 1$ sont évidents), on se ramène au cas de deux éléments inversibles x, y. Alors :

$$(xy)(y^{-1}x^{-1}) = \big(x(yy^{-1})\big)x^{-1} = (xe)x^{-1} = xx^{-1} = e,$$

et l'on vérifie de même que $(y^{-1}x^{-1})(xy) = e$, d'où la conclusion. ∎

Ainsi, dans l'un des ensembles $\mathbb{Z}, \mathbb{Q}$, muni de son addition, tout élément x a un *opposé* $-x$. Pour l'addition de $\mathbb{N}$, seul 0 possède un opposé. Tout $x \in \mathbb{Q}^*$ a pour *inverse* multiplicatif $1/x$ (noté aussi x^{-1}). Dans $\mathbb{Z}\backslash\{0\}$, seuls -1 et 1 sont inversibles pour la multiplication.

Exemple. Soit E un ensemble. Si $f \in \mathcal{F}(E, E)$ possède un symétrique f' pour la loi $\circ$, on a $f' \circ f = \mathrm{Id}_E = f \circ f'$. La première (resp. deuxième) égalité montre que f est injective (resp. surjective). Ainsi f est bijective.

Soient inversement f une bijection de E sur E et f^{-1} la *bijection réciproque* de f. Puisque $f^{-1} \circ f = \mathrm{Id}_E = f \circ f^{-1}$, f^{-1} est symétrique de f. Les éléments symétrisables (on dit plutôt *inversibles*) de $\mathcal{F}(E, E)$ pour la loi $\circ$ sont donc exactement les *permutations* de E.

À partir d'une loi, on peut en définir d'autres, *cf.* le module I.1. Ainsi, partant de $(E, \top)$, soit F une partie de E *stable pour la loi* $\top$: pour tous $x, y \in F$, $x \top y$ appartient aussi à F, *i.e.* $F \top F \subset F$ (voir l'exemple 5 de la page 112). Alors $(x, y) \mapsto x \top y$ est une loi *sur* F, appelée *loi induite par* $\top$. Si F est stable pour une loi associative (resp. commutative), la loi induite a la même propriété. La partie vide et E sont stables pour $\top$. De même :

Proposition 7. Soit $\top$ une loi sur l'ensemble E. Si $(F_i)_{i \in I}$ est une famille de parties de E stables pour $\top$, l'*intersection* des F_i pour $i \in I$ est stable pour $\top$.

Dans $\mathbb{Z}$, $\mathbb{N}$ est *additivement et multiplicativement stable*, de même pour $n\mathbb{Z}$ si $n \in \mathbb{Z}$. Par contre, la *réunion* $F = 2\mathbb{Z} \cup 3\mathbb{Z}$ *n'est pas* additivement stable : $2 \in F$, $3 \in F$ mais $5 \notin F$ (5 n'est multiple ni de 2 ni de 3).
L'ensemble $\mathfrak{S}(E)$ des permutations d'un ensemble E est stable pour la composition des applications sur $\mathcal{F}(E, E)$: la composée de deux bijections de E sur E est une bijection de E sur E, *cf.* le module I.1.
Rappelons aussi ce qu'est une *loi produit*. Soient $E_1, \ldots, E_n$ ($n \geqslant 2$) des ensembles, chacun muni d'une loi $\top_i$. Soit $F := E_1 \times E_2 \times \cdots E_n$. On appelle *produit* des lois $\top_1, \ldots, \top_n$ la loi $\top$ définie sur F par la formule suivante, dans laquelle, pour $i = 1, \ldots, n$, x_i et y_i appartiennent à E_i :

$$(x_1, \ldots, x_n) \top (y_1, \ldots, y_n) := (x_1 \top_1 y_1, \ldots, x_n \top_n y_n). \tag{11}$$

Si chaque $\top_i$ est notée additivement (resp. multiplicativement) la loi $\top$ sera notée additivement (resp. multiplicativement). Si chaque $\top_i$ est associative (resp. commutative), $\top$ l'est aussi. Si chaque E_i possède un élément neutre e_i pour la loi $\top_i$, $e := (e_1, \ldots, e_n) \in F$ est élément neutre pour la loi $\top$. Soit de plus $x := (x_1, \ldots, x_n) \in F$. Si chaque x_i possède un symétrique x'_i pour la loi $\top_i$, $x' := (x'_1, \ldots, x'_n)$ est symétrique de x pour la loi $\top$.

Le nom de « loi produit » vient du *produit* $E_1 \times E_2 \times \cdots E_n$, il n'a rien à voir avec la notation, multiplicative ou additive ou autre.

Supposons les E_i tous égaux à un même ensemble E et les lois $\top_i$ toutes égales à une même loi $\top$. La loi produit sur $F := E^n$ sera encore notée $\top$. Ainsi l'addition des vecteurs du plan est une loi produit.
Plus généralement, partons de $(E, \top)$, et soit X un ensemble quelconque. Sur l'ensemble $\mathcal{F}(X, E)$ des applications de X dans E, on a défini comme suit une loi, notée aussi $\top$. Si $f, g \in \mathcal{F}(X, E)$, $f \top g$ est donnée par :

$$(f \top g)(x) \ := \ f(x) \top g(x) \quad \text{pour tout} \quad x \in X. \tag{12}$$

D'où une addition et une multiplication « naturelles » sur l'ensemble $\mathcal{F}(\mathbb{R}, \mathbb{R})$ des fonctions définies sur $\mathbb{R}$. La loi $\top$ sur $\mathcal{F}(X, E)$ est associative (ou commutative) si la loi $\top$ sur E l'est ; si $e \in E$ est neutre pour la loi $\top$, l'application $x \mapsto e$ de X dans E est neutre pour la loi $\top$ sur $\mathcal{F}(X, E)$.

2 Groupes

2.1 Définitions, règles de calcul

Définition 2. On appelle *groupe* tout ensemble muni d'une loi *associative*, possédant un élément *neutre*, et dans lequel tout élément est *symétrisable*. Le groupe est dit *commutatif* ou *abélien* lorsque sa loi est commutative.

Un groupe n'est jamais vide : il a au moins un élément neutre. Un groupe est dit *multiplicatif* (resp. *additif*) si sa loi est $+$ (resp. $\times$). Si G, muni de la loi $\top$, est un groupe, on dit aussi que $(G, \top)$ est un groupe ou que $\top$ est une loi de groupe sur G. Si l'ensemble G est fini, on dit que G est un *groupe fini*, et le cardinal de G est appelé *ordre de* G.
Chacun des ensembles $\mathbb{Z}, \mathbb{Q}$ est un groupe additif abélien ; pour $(\mathbb{Z}, +)$, voir la proposition 45 de la page 92. La multiplication sur $\mathbb{Q}$ est associative, commutative, et 1 est élément neutre. Tout $x \in \mathbb{Q}^*$ a un inverse $1/x$, mais 0 n'a pas d'inverse, donc $(\mathbb{Q}, \times)$ n'est pas un groupe. Par contre, $\mathbb{Q}^*$ est multiplicativement stable, et $(\mathbb{Q}^*, \times)$ est un groupe abélien (*cf.* la proposition 60 de la page 100).

Exemples.

1. Soit $G := \{e\}$ un *singleton*. Il y a sur G une unique loi , c'est évidemment une loi de groupe. Un tel groupe est dit *trivial*.
2. Soient $G_1, \ldots, G_n$ des groupes ($n \geqslant 1$). Alors $G := G_1 \times \cdots \times G_n$, muni de la *loi produit* est un groupe, appelé *groupe produit* $G_1 \times \cdots \times G_n$. Il est abélien si, et seulement si, chaque G_i l'est (le vérifier).
3. Soient $(A, +)$ un groupe et X un ensemble. L'addition sur $\mathcal{F}(X, A)$ est associative, avec pour élément neutre l'application *nulle* (notée 0) : $x \mapsto 0$. Tout $f \in \mathcal{F}(X, A)$ a un opposé $-f : x \mapsto -f(x)$ (le vérifier), donc $\big(\mathcal{F}(X, A), +\big)$ *est un groupe*.

   Si $f, g \in \mathcal{F}(X, A)$ et $x \in X$, on a :
$$(f + g)(x) \;=\; f(x) + g(x) \quad \text{et} \quad (-f)(x) \;=\; -f(x). \tag{13}$$

Voici un exemple important de groupe, en général non commutatif (cet exemple sera étudié plus en détail dans la section 2.3) :

Théorème et définition 8. L'ensemble $\mathfrak{S}(E)$ des permutations d'un ensemble E, muni de la loi $\circ$ (composition des applications) est un groupe, appelé *groupe symétrique* de E. Si $n \geqslant 1$ est un entier et $E = [\![1, n]\!]$, ce groupe est noté $\mathfrak{S}_n$.

Démonstration. Nous avons vu que $\mathfrak{S}(E)$ est une partie de $\mathcal{F}(E, E)$ *stable* pour la loi $\circ$. D'où une loi associative *induite* sur $\mathfrak{S}(E)$, notée encore $\circ$, dont Id_E est élément neutre. Tout $s \in \mathfrak{S}(E)$ possède un inverse (ou symétrique) pour la loi $\circ$, à savoir la bijection réciproque s^{-1} de s. ∎

Voici maintenant des règles de calcul importantes dans un groupe.

Proposition 9. Soit $(G, \top)$ un groupe. Notons e l'élément neutre et, pour tout $x \in G$, notons x' le symétrique de G. On a les propriétés suivantes :

1. Associativité : pour tous $x, y, z \in G$, $(x \top y) \top z = x \top (y \top z)$.
2. Pour tout $x \in G$, $x \top e = x = e \top x$.
3. Pour tout $x \in G$, $x \top x' = e = x' \top x$ et $(x')' = x$.
4. Pour tous $x, y \in G$, $(x \top y)' = y' \top x'$.
5. *Régularité* : pour tous $x, y, z \in G$, on a les implications :
$$(x \top y = x \top z) \Longrightarrow (y = z) \quad \text{et} \quad (y \top x = z \top x) \Longrightarrow (y = z).$$
6. Soient $a, b \in G$. L'équation $x \top a = b$ (resp. $a \top x = b$) possède une seule solution dans G, à savoir $x = b \top a'$ (resp. $x = a' \top b$).

Démonstration. Les propriétés 1, 2, 3 résultent des définitions $\big((x')' = x$, vu la symétrie des égalités (9) en $x, x'\big)$. Tout élément de G est symétrisable, donc 4 et 5 résultent de la proposition 6 de la page 116 et du théorème 5 de la page 116. Pour 6, soient $a, b \in G$ et $x := b \top a'$. Alors $x \top a = (b \top a') \top a = b \top (a' \top a) = b \top e = b$, donc $x \top a = b$. Si $y \in G$ et $y \top a = b$, $x \top a = y \top a$, d'où $x = y$, par régularité.

De même pour l'équation $a \top x = b$. ∎

Notons que, si $x, y \in G$ et si $x \top y = e$, alors $x = y'$ et $y = x'$. En effet $x \top y = e = x \top x'$, donc $y = x'$ (régularité), puis $x = (x')' = y'$.

Proposition 10. Soient G un groupe multiplicatif et $a \in G$.
L'application $t_a : x \mapsto ax$ de G dans G est appelée *translation à gauche* définie par a. C'est une *permutation* de G. En notation additive, t_a est l'application $x \mapsto a + x$.

Démonstration. Soient $a, b \in G$. Il existe un unique $x \in G$ tel que $t_a(x) = b$ (propriété 6), *i.e.* un unique antécédent de b par t_a. ∎

Soient G un groupe multiplicatif et $x \in G$. Rappelons que nous avons défini x^n pour tout entier $n \geqslant 1$, et que nous avons posé $x^0 := 1$. Généralisons :

Définition 3. Soient G un groupe et $x \in G$. Pour tout entier $n < 0$, on pose :
$$x^n := (x^{-n})^{-1} \ \text{(inverse de } x^{-n}) \ \text{en notation multiplicative,} \tag{14}$$
$$nx := -\big((-n)x\big) \ \text{(opposé de } (-n)x) \ \text{en notation additive.} \tag{15}$$

Les deux sens de x^{-1} (lesquels ?) coïncident, heureusement.

Proposition 11. Soient G un groupe et $x \in G$.

1. Pour tous entiers $m, n \in \mathbb{Z}$, on a :
$$x^m x^n = x^{m+n},\ (x^n)^{-1} = x^{-n} \quad \text{et} \quad (x^m)^n = x^{mn}, \quad \text{ou} \tag{16}$$
$$mx + nx = (m+n)x,\ -(nx) = (-n)x \quad \text{et} \quad m(nx) = (mn)x, \tag{17}$$
suivant que la notation soit multiplicative ou additive.
2. Soit $y \in G$ un élément *commutant* avec x. Pour tout $n \in \mathbb{Z}$ on a :
$$(xy)^n \ = \ x^n y^n \quad \text{en notation multiplicative (ici } xy = yx), \tag{18}$$
$$n(x+y) \ = \ nx + ny \quad \text{en notation additive (ici } x + y = y + x). \tag{19}$$

Démonstration. Prouvons la première égalité (16). Le résultat est vrai lorsque $m \geqslant 0$ et $n \geqslant 0$. Si $m < 0$ et $n < 0$, on a $-m > 0$, $-n > 0$, d'où :
$$x^m x^n = (x^{-m})^{-1}(x^{-n})^{-1} = (x^{-n}x^{-m})^{-1} = (x^{-(m+n)})^{-1} = x^{m+n},$$
car $-(m+n) > 0$. Si $m \geqslant 0$ et $n < 0$ (ou l'inverse), la preuve est analogue, en distinguant les cas $m+n \geqslant 0$ et $m+n < 0$. Les autres égalités sont laissées au lecteur en exercice. ■

Soient G un groupe additif et $x, y \in G$. La somme $x + (-y)$ est notée $x - y$ (*différence* de x et y). Vérifier par exemple que $-(x-y) = (y-x)$.

2.2 Sous-groupes, morphismes de groupes

Définition 4. Soit $(G, \top)$ un groupe. Une partie H de G est appelée *sous-groupe* de G si elle vérifie les conditions suivantes :

1. La partie H n'est pas vide.
2. La partie H est stable pour la loi $\top$.
3. Pour tout $x \in H$, le symétrique de x appartient aussi à H.

Les conditions 1, 2, 3 sont équivalentes aux conditions 4 et 5 ci-après :

Proposition 12. Soit $(G, \top)$ un groupe, d'élément neutre e ; pour tout $x \in G$, notons x' le symétrique de x. Une partie H de G est un sous-groupe de G si, et seulement si, elle vérifie les conditions suivantes :

4. L'élément neutre e de G appartient à H.
5. Pour tous $x, y \in H$, l'élément $x \top y'$ appartient aussi à H.

Démonstration. Supposons que H soit un sous-groupe de G. Soit $a \in H$ (condition 1). Puisque $a' \in H$ (condition 3), $e = a \top a' \in H$ (condition 2). Ensuite, si $x, y \in H$, $y' \in H$ (condition 3), puis $x \top y' \in H$ (condition 2).
Supposons que H vérifie 4 et 5. Puisque $e \in H$, $H \neq \varnothing$. Si $x \in H$, $e \top x' = x' \in H$ (condition 5), donc 3 est vérifiée. Si $x, y \in H$, on a $y' \in H$, donc $x \top (y')' = x \top y \in H$ (condition 5), et ainsi H est stable pour $\top$. ■

Dans un groupe G d'élément neutre e, G et $\{e\}$ sont des sous-groupes de G ; $\{e\}$ est appelé sous-groupe *trivial* de G.

Proposition 13. Soient $(G, \top)$ un groupe et H un sous-groupe de G. Muni de la loi $\top'$ induite par $\top$, H est un groupe (d'où la terminologie de « sous-groupe »).

Démonstration. La loi $\top'$ est associative, car $\top$ l'est. Le neutre e de G appartient à H, c'est un élément neutre pour $\top'$. Si $x \in H$, le symétrique x' de x dans G appartient à H, c'est un symétrique de x pour la loi $\top'$. ∎

Exemples.

1. Les sous-groupes de $(\mathbb{Z}, +)$ ont été décrits dans le théorème 50 de la page 95 : ce sont les $n\mathbb{Z}$, où $n \in \mathbb{N}$ est arbitraire (noter que la définition de sous-groupe donnée ici coïncide, dans le cas de $(\mathbb{Z}, +)$, avec celle qui précède le théorème 50 de la page 95). Noter aussi que $\mathbb{Z}$ est un sous-groupe de $(\mathbb{Q}, +)$.
2. L'ensemble $\mathbb{Q}_+^*$ et $\{-1, 1\}$ sont des sous-groupes de $(\mathbb{Q}^*, \times)$.

Exercice 1.
Montrer que l'intersection H d'une famille $(H_i)_{i \in I}$ de sous-groupes d'un groupe G est un sous-groupe de G.
Solution. En effet la proposition 7 de la page 117 montre que H est une partie stable de G ; le neutre de G appartient à chaque H_i, donc à H. Si $x \in H$, le symétrique x' de x appartient de même à H.

Exercice 2.
Soit $(G, \times)$ un groupe. Montrer que l'ensemble $Z(G)$ des $a \in G$ commutant avec *tout* élément de G est un sous-groupe de G, appelé *centre* de G.
Solution. D'abord $1 \in Z(G)$. Soient $a, b \in Z(G)$. Pour tout $x \in G$, on a :

$$(ab)x = a(bx) = a(xb) = (ax)b = (xa)b = x(ab),$$

donc $ab \in Z(G)$. De même $ax = xa$ d'où, en multipliant à gauche et à droite par a^{-1} : $xa^{-1} = a^{-1}x$, ce qui montre que $a^{-1} \in Z(G)$.

Voici un résultat important concernant les groupes finis :

Théorème 14 (Lagrange). Soient G un groupe fini et H un sous-groupe de G. L'ordre de H divise l'ordre de G.

Démonstration. On suppose, par exemple, G multiplicatif. Soient N, m les ordres respectifs de G et H.
Pour tout $a \in G$, $aH := \{ax \mid x \in H\}$ est l'image de H par la translation $t_a : x \mapsto ax$, appliquant bijectivement H sur aH (proposition 10 de la page 119). D'où $\operatorname{card}(aH) = m$.
La relation $\sim$ définie sur G par $a \sim b$ si $aH = bH$ est clairement une relation d'équivalence. Soit $a \in G$. La classe de a modulo $\sim$ est aH. En effet, soit $b \in G$. Si $b \sim a$, on a $b = b \times 1 \in bH = aH$. Inversement, si $b \in aH$, il existe $h \in H$ tel que $b = ah$. Alors $bH = ahH = aH$, car H est un sous-groupe de G et donc $hH = H$ (le vérifier).
D'après l'alinéa précédent, la classe aH de a modulo $\sim$ est de cardinal m.
Les classes $C_1, \ldots, C_r$ modulo $\sim$ forment une partition de G, et chacune est de cardinal m, nous l'avons vu. Ainsi $N = rm$, donc m divise N. ∎

L'exercice 1 de la page précédente justifie la définition suivante :

Définition 5. Soient G un groupe et S une partie de G. On appelle sous-groupe de G *engendré par* S l'intersection de *tous* les sous-groupes de G contenant S (G lui-même est un tel sous-groupe), et ce sous-groupe est noté $\langle S\rangle$. Si $\langle S\rangle = G$, on dit que S *engendre* G ou est une *partie génératrice* de G.

Par exemple, $\langle\varnothing\rangle = \{e\}$ et $\langle G\rangle = G$. Si $S := \{a\}$ ou $S := \{a, b\}$, on écrit $\langle a\rangle$ ou $\langle a, b\rangle$ au lieu de $\langle\{a\}\rangle$ ou $\langle\{a, b\}\rangle$. Voici un résultat évident :

Proposition 15. Soit S une partie du groupe G. Le sous-groupe $\langle S\rangle$ est, pour l'inclusion, le *plus petit sous-groupe de G contenant S* : c'est un sous-groupe de G contenant S, et il contient tout sous-groupe de G contenant S.

Proposition 16. Soient G un groupe multiplicatif (resp. additif) et $x \in G$. Alors $\langle x\rangle$ est formé des puissances x^n (resp. des multiples nx) de x, n décrivant $\mathbb{Z}$.

Démonstration. Soit $A := \{x^n \mid n \in \mathbb{Z}\}$. C'est un sous-groupe de G (formules (16)) contenant x, donc $A \supset \langle x\rangle$. Or $\langle x\rangle$ est un sous-groupe de G contenant x, il contient x^n pour tout entier $n \geqslant 0$ (récurrence sur n). Si $n < 0$, $\langle x\rangle$ contient x^{-n}, donc son inverse x^n. Ainsi $\langle x\rangle \supset A$. ∎

Définition 6. Un groupe G est dit *monogène* s'il existe $x \in G$ tel que $\langle x\rangle = G$ (un tel x est dit *générateur* de G). Un groupe monogène *fini* est dit *cyclique*.

Les groupes monogènes seront entièrement décrits dans le théorème 30 de la page 132.
Tout groupe fini G d'ordre p *premier* est cyclique. Soit en effet $x \in G$ distinct du neutre. Alors $\langle x\rangle$ est un sous-groupe non trivial de G, son ordre $m > 1$ est un diviseur de p. Puisque p est premier, $m = p$, d'où $\langle x\rangle = G$.
Noter que la *réunion* de deux sous-groupes d'un groupe G n'est pas, en général, un sous-groupe de G (*cf.* l'exemple de la page 117). Précisons :

Théorème et définition 17. Soit G un groupe *commutatif*, noté additivement. Considérons des sous-groupes $C_1, \ldots, C_n$ ($n \geqslant 2$) de G.
Le sous-groupe de G engendré par la *réunion* des C_i est l'ensemble $C_1 + \cdots + C_n$ formé des éléments du type $x_1 + \cdots + x_n$, où $(x_1, \ldots, x_n)$ décrit $C_1 \times \cdots \times C_n$.

Démonstration. Par récurrence sur n, on se ramène aisément au cas $n = 2$. Posons $C := C_1$, $C' := C_2$ et $C + C' := \{x + x' \mid (x, x') \in C \times C'\}$.
Si $u, v \in C + C'$, il existe (x, x'), $(y, y') \in C \times C'$ tels que $u = x + x'$, $v = y + y'$; $x + y \in C$ car C est un sous-groupe de G, de même $x' + y' \in C'$. Alors :

$$u + v = x + \overbrace{x' + y} + y' = x + (\overbrace{y + x'}) + y' = (x + y) + (x' + y') \in C + C'.$$

De même $-u \in C + C'$ et $0 \in C + C'$. Ainsi $C + C'$ est un sous-groupe de G, contenant $C = C + \{0\}$ (et C'). D'où $\langle C \cup C'\rangle \subset C + C'$ (proposition 15).

Soit $u \in C + C'$: il existe $(x, x') \in C \times C'$ tel que $u = x + x'$. Alors $x \in C \subset \langle C \cup C' \rangle$, de même $x' \in \langle C \cup C' \rangle$, donc $u \in \langle C \cup C' \rangle$, car $\langle C \cup C' \rangle$ est un sous-groupe de G.
Ainsi $C + C' \subset \langle C \cup C' \rangle$. ■

Par exemple, si $a, b \in \mathbb{Z}$, leur pgcd est le seul $d \in \mathbb{N}$ tel que $a\mathbb{Z} + b\mathbb{Z} = d\mathbb{Z}$.
Voici une définition importante :

Définition 7. Soient G, G' deux groupes multiplicatifs. On appelle *morphisme* de G dans G' toute application $f : G \to G'$ vérifiant la condition suivante :

$$f(xy) = f(x)f(y) \quad \text{pour tous} \quad x, y \in G. \qquad (*)$$

Pour des lois additives, la condition $(*)$ s'écrit ainsi : $f(x + y) = f(x) + f(y)$. Le lecteur vérifiera que le composé de deux morphismes est un morphisme. Soit e' le neutre de G'. L'application $x \mapsto e'$ est un morphisme, dit *trivial*.

Proposition 18. Soit $f : G \to G'$ un morphisme de groupes multiplicatifs.
Pour tout $x \in G$ et tout $n \in \mathbb{Z}$, $f(x^n) = f(x)^n$. En particulier, $f(x^{-1}) = f(x)^{-1}$.
Si e (resp. e') est l'élément neutre de G (resp. G'), $f(e) = e'$.

Démonstration. D'abord $ee = e$, donc $f(e) = f(ee) = f(e)f(e)$. Mais $f(e)e' = f(e)$, donc $f(e) = e'$ d'après la proposition 9 de la page 119 (propriété 5). Soit $x \in G$. L'égalité $xx^{-1} = e$ implique $e' = f(e) = f(x)f(x^{-1})$, ce qui entraîne $f(x^{-1}) = f(x)^{-1}$. Pour l'égalité $f(x^n) = f(x)^n$, le cas $n > 0$ est évident par récurrence sur n, à partir de l'égalité $(*)$, et le cas $n < 0$ s'en déduit. ■

Voici maintenant un résultat et une définition essentiels :

Théorème et définition 19. Soit $f : G \to G'$ un morphisme de groupes.

1. L'image $f(H)$ d'un sous-groupe H de G est un sous-groupe de G'. Ainsi $f(G)$ est un sous-groupe de G', noté $\mathrm{Im}(f)$ (*image* de f).
2. Si H' est un sous-groupe de G', son *image réciproque* $f^{-1}(H')$ est un sous-groupe de G. Rappelons que $f^{-1}(H') := \{x \in G \mid f(x) \in H'\}$.
3. En particulier, notons e' l'élément neutre de G'. Alors $f^{-1}(e')$ est un sous-groupe de G, appelé *noyau* de f et noté $\mathrm{Ker}(f)$.

Démonstration. Supposons, par exemple, les groupes G et G' multiplicatifs et notons e, e' leurs éléments neutres respectifs. Soit H un sous-groupe de G.
Puisque $e \in H$, $e' = f(e) \in f(H)$. Soient ensuite $x', y' \in f(H)$: il existe $x, y \in H$ tels que $x' = f(x)$, $y' = f(y)$. Alors :

$$x'y'^{-1} = f(x)f(y)^{-1} = f(x)f(y^{-1}) = f(xy^{-1}) \in f(H) \quad \text{car} \quad xy^{-1} \in H.$$

D'après la proposition 12 de la page 120, $f(H)$ est donc un sous-groupe de G'.

Soit H' un sous-groupe de G'. On a $e \in f^{-1}(H')$ car $f(e) = e' \in H'$. Si $x, y \in f^{-1}(H')$, $f(xy^{-1}) = f(x)f(y^{-1}) = f(x)f(y)^{-1} \in H'$ parce que $f(x), f(y) \in H'$, d'où $xy^{-1} \in f^{-1}(H')$. Ainsi $f^{-1}(H')$ est un sous-groupe de G.
L'assertion 3 résulte de l'assertion 2, car $\{e'\}$ est un sous-groupe de G'. ■
Le noyau d'un morphisme mesure la « non injectivité » de ce morphisme :

Théorème 20. Soit $f : G \to G'$ un morphisme de groupes.

1. Pour que f soit *injectif*, il faut, et il suffit, que son noyau soit *trivial*.
2. Si f est *bijectif*, la bijection *réciproque* f^{-1} *est aussi un morphisme*.
3. Supposons G *fini*. Alors $\mathrm{Im}(f)$ est finie, et l'on a :

$$\mathrm{ordre}(G) = \mathrm{ordre}\big(\mathrm{Ker}(f)\big) \times \mathrm{ordre}\big(\mathrm{Im}(f)\big). \tag{20}$$

Démonstration. Soit e, e' les éléments neutres respectifs de G et G' que l'on suppose multiplicatifs. Si $x, y \in G$, on a $f(x^{-1}y) = f(x)^{-1}f(y)$ (*cf.* la proposition 18 de la page précédente), d'où les équivalences :

$$\big(f(x) = f(y)\big) \iff \big(x^{-1}y \in \mathrm{Ker}(f)\big) \iff \big(y \in x\ \mathrm{Ker}(f)\big).$$

Si $\mathrm{Ker}(f)$ est trivial, on en déduit que f est injectif. Si f est injectif et $x \in \mathrm{Ker}(f)$, on a $f(x) = e' = f(e)$, donc $x = e$. Ainsi $\mathrm{Ker}(f) = \{e\}$.
Supposons f bijectif ; soient $x', y' \in G'$. Posant $x := f^{-1}(x')$ et $y := f^{-1}(y')$, on a $f(x) = x'$, $f(y) = y'$. Alors $f(xy) = f(x)f(y) = x'y'$, et par suite $f^{-1}(x'y') = xy = f^{-1}(x')f^{-1}(y')$. Cela prouve que f^{-1} est un morphisme.
Pour 3, on peut supposer f surjectif $\big($remplacer G' par $f(G)\big)$. Soient $b \in G'$ et x un antécédent de b par f. D'après les équivalences ci-dessus, on a $f^{-1}(b) = x\ \mathrm{Ker}(f)$, donc le cardinal de $f^{-1}(b)$ est égal à l'ordre de $\mathrm{Ker}(f)$ (*cf.* la proposition 10 de la page 119). La conclusion vient alors de la formule (1) de la page 72. ■

Définition 8. On appelle *isomorphisme* d'un groupe G sur un groupe G' tout morphisme *bijectif* de G sur G'. Un isomorphisme d'un groupe G sur lui-même est appelé *automorphisme* de G.

Comme exemple, soient $(G, \times)$ un groupe et $a \in G$. L'application $x \mapsto axa^{-1}$ est un morphisme f de G dans G (*conjugaison par* a) : si $x, y \in G$, on a : $f(x)f(y) = (axa^{-1})(aya^{-1}) = (ax)(a^{-1}a)(ya^{-1}) = a(xy)a^{-1} = f(xy)$. On vérifie que f est bijectif et que f^{-1} est l'application $x \mapsto a^{-1}xa$.
Deux groupes sont dits *isomorphes* s'il existe un isomorphisme de l'un sur l'autre. Ils possèdent alors les mêmes propriétés : si l'un est abélien (resp. fini, resp. monogène,...), l'autre l'est aussi.
Voici un exemple, anticipant sur le module IV.3. L'application $x \mapsto \mathrm{e}^x$ est une *bijection* f de $\mathbb{R}$ sur $\mathbb{R}_+^*$. C'est un morphisme de $(\mathbb{R}, +)$ dans $(\mathbb{R}_+^*, \times)$: $\mathrm{e}^{x+y} = \mathrm{e}^x\mathrm{e}^y$ pour tous $x, y \in \mathbb{R}$. Ainsi f est un *isomorphisme* de $\mathbb{R}$ sur $\mathbb{R}_+^*$. La bijection réciproque f^{-1} de f est l'application $t \mapsto \ln(t)$, c'est donc un *morphisme* de $(\mathbb{R}_+^*, \times)$ dans $(\mathbb{R}, +)$. D'où la propriété connue : $\ln(tu) = \ln(t) + \ln(u)$ pour tous $t, u \in \mathbb{R}_+^*$.

Théorème et définition 21.
Soient G un groupe, par exemple multiplicatif, et $x \in G$.

1. Il existe un unique morphisme f de $(\mathbb{Z}, +)$ dans G tel que $f(1) = x$.
   De plus $f(n) = x^n$ pour tout $n \in \mathbb{Z}$, et l'image de f est $\langle x \rangle$.
2. Si f est injectif, c'est un isomorphisme de $(\mathbb{Z}, +)$ sur $\langle x \rangle$.
   En particulier $\langle x \rangle$ est *infini*, et l'on dit alors que x *est d'ordre infini*.
3. Si f n'est pas injectif, il existe un unique $n \in \mathbb{N}^*$ tel que le noyau de f soit égal à $n\mathbb{Z}$. Alors $1, x, x^2, \ldots, x^{n-1}$ sont deux à deux distincts, $\langle x \rangle = \{1, x, x^2, \ldots, x^{n-1}\}$, et $\langle x \rangle$ est *cyclique d'ordre* n ; n est appelé *ordre de* x, c'est le plus petit entier $k > 0$ tel que $x^k = 1$.

Démonstration. Pour l'assertion 1, notons f l'application $n \mapsto x^n$ de $\mathbb{Z}$ dans G. Par définition, $f(\mathbb{Z}) = \langle x \rangle$. La première formule (16) montre que f est un morphisme de $(\mathbb{Z}, +)$ dans G ; de plus $f(1) = x$. L'unicité de f vient de l'exercice II.2.7. L'assertion 2 résulte évidemment de l'assertion 1.
Pour l'assertion 3, le théorème 20 de la page ci-contre montre que $\mathrm{Ker}(f)$ est un sous-groupe non trivial de $(\mathbb{Z}, +)$. D'après l'exemple 1 de la page 121, il existe un unique entier $n \geqslant 1$ tel que $\mathrm{Ker}(f) = n\mathbb{Z}$. Pour tout $k \in \mathbb{Z}$, l'égalité $x^k = 1$ signifie que $k \in n\mathbb{Z}$, *i.e.* que n divise k. D'où la dernière assertion de l'énoncé.
Soient $h, k \in [\![0, n-1]\!]$ tels que $x^h = x^k$ et par exemple $h \leqslant k$. Alors $k - h \in [\![0, n-1]\!]$, et $x^{k-h} = 1$. Ainsi n divise $k-h$, d'où $k-h = 0$. Posons enfin $A := \{1, x, \ldots, x^{n-1}\}$; A est donc de cardinal n, et $A \subset \langle x \rangle$. Soit inversement $k \in \mathbb{Z}$. Effectuons la division euclidienne de k par n : $k = nq+r$, où $r \in [\![0, n-1]\!]$. Alors $x^k = x^{qn+r} = (x^n)^q x^r = x^r \in A$, ce qui prouve l'inclusion $\langle x \rangle \subset A$. D'où $\langle x \rangle = A$. ■

Voici une conséquence de ce théorème :

Théorème 22 (Lagrange). Soient G un groupe fini et $x \in G$. L'ordre de x divise l'ordre de G.

Démonstration. Puisque G est fini, $\langle x \rangle$ l'est aussi. Ainsi x est d'ordre fini, égal à l'ordre de $\langle x \rangle$. Il suffit alors d'appliquer le théorème 14 de la page 121. ■

2.3 Groupe symétrique

L'étude des groupes symétriques servira surtout en algèbre linéaire, notamment pour les *déterminants*. Une étude plus poussée sera faite en L2.
Soit E un ensemble. Si $f : E \to E$ est une application, tout $x \in E$ tel que $f(x) = x$ est appelé *point fixe de* f, on dit aussi que f laisse fixe x.
Par définition, le *groupe symétrique de* E est l'ensemble $\mathfrak{S}(E)$ des *permutations* de E, muni de la loi $\circ$ (composition des applications). Le théorème et définition 8 de la page 118 montre que $\big(\mathfrak{S}(E), \circ\big)$ *est un groupe*. Si $E = [\![1, n]\!]$, $n \geqslant 1$, ce groupe est noté $\mathfrak{S}_n$.

Rappelons quelques points importants.
Si $s, t \in \mathfrak{S}(E)$ et $x \in E$, $(s \circ t)(x) = s\big(t(x)\big)$. L'élément neutre de $\mathfrak{S}(E)$ est l'application identique Id_E de E, notée 1 si cela ne prête pas à confusion. Soit $s \in \mathfrak{S}(E)$. L'*inverse* de s dans $\mathfrak{S}(E)$ est la *bijection réciproque* s^{-1} de s : $s^{-1} \circ s = 1 = s \circ s^{-1}$. Si E est fini de cardinal $n \geqslant 1$, l'*ordre* (ou cardinal) de $\mathfrak{S}(E)$ est $n!$ (*cf.* le théorème 28 de la page 77). Ainsi l'ordre de $\mathfrak{S}_n$ est $n!$.
Voici, sur un exemple, une notation « matricielle » pour définir une permutation s de E. Soit $E := \{a, b, c, d, e\}$. On écrit $s = \begin{pmatrix} a & b & c & d & e \\ c & b & d & e & a \end{pmatrix}$ pour désigner la permutation s de E appliquant a, b, c, d, e dans l'ordre sur c, b, d, e, a. Le fait que s soit bijective se traduit ainsi : tout élément de E apparaît une fois et une seule dans la seconde ligne de la « matrice » ci-dessus. Introduisons une notion importante, celle de *cycle* :

Définition 9. Soient E un ensemble, $n \geqslant 2$ un entier et $a_1, \ldots, a_n \in E$ des éléments deux à deux distincts. Considérons la permutation c de E définie ainsi :

$$\begin{cases} c(a_1) := a_2,\ c(a_2) := a_3, \ldots,\ c(a_{n-1}) := a_n \quad \text{et} \quad c(a_n) := a_1, \\ c(x) := x \ \text{ pour tout } x \in E \text{ distinct de } a_1, \ldots, a_n\,; \end{cases}$$

c est notée $(a_1\ a_2\ \ldots\ a_n)$. Une permutation de ce type est appelée n*-cycle* (ou *cycle de longueur* n) de E. Un 2-cycle est appelé *transposition*.

Vérifier que c est une permutation de E. Dans $\mathfrak{S}_5$, le 3-cycle $(2\ 3\ 5)$ a pour écriture matricielle $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 5 & 4 & 2 \end{pmatrix}$.
Dans $\mathfrak{S}_{100}$, on imagine la notation matricielle du 3-cycle $(2\ 3\ 5)$!

Exercice 3.
Le groupe $\mathfrak{S}_3$ est d'ordre $3! = 6$. Vérifier que

$$\mathfrak{S}_3 = \{1,\ (1\ 2),\ (1\ 3),\ (2\ 3),\ (1\ 2\ 3),\ (1\ 3\ 2)\},$$

puis écrire la table de multiplication de $\mathfrak{S}_3$.
Solution. On peut classer les permutations de trois éléments selon le nombre d'éléments qu'elles laissent fixes (ce nombre est compris entre 0 et 3). Trois éléments fixes : l'identité, bien entendu ; deux éléments fixes : impossible car, par bijectivité, le troisième le serait nécessairement ; un élement fixe : on trouve les trois transpositions (12), (13) et (23) ; aucun élément fixe : on trouve les deux permutations cycliques (123) et (132).
Pour dresser la table de multiplication, qui comporte $6 \times 6 = 36$ cases à remplir, on procède comme suit. La ligne et la colonne de 1 (élément neutre) reproduisent la liste des permutations. Sur la diagonale, le carré de chaque transposition est 1 et le carré de chacun des deux 3-cycles est l'autre. Le produit des deux 3-cycles est 1. On finit alors de remplir la table en notant que dans chaque ligne, resp. dans chaque colonne, chaque permutation apparaît une fois et une seule. C'est plus facile qu'un sudoku !

Ne pas confondre les notations : $(1\ 2)$ représente une transposition alors que $(1, 2)$ est un couple.

Proposition 23.
Soit $c := (a_1\ a_2\ \dots\ a_n)$ un n-cycle de E, comme dans la définition 9 de la page précédente.

1. La permutation c est la composée de $n-1$ transpositions :
$$(a_1\ a_2\ \dots\ a_n) = (a_1\ a_2) \circ (a_2\ a_3) \circ \cdots \circ (a_{n-1}\ a_n). \tag{21}$$
2. Pour tout $s \in \mathfrak{S}(E)$, on a la *formule de conjugaison* suivante :
$$s \circ (a_1\ a_2\ \dots\ a_n) \circ s^{-1} \ = \ \big(s(a_1)\ s(a_2)\ \dots\ s(a_n)\big). \tag{22}$$

Démonstration. Pour l'assertion 1, raisonnant par récurrence sur n (le cas $n = 2$ est évident), on se ramène à vérifier l'égalité suivante, pour $n \geqslant 3$:
$$(a_1\ a_2\ \dots\ a_n) \ = \ (a_1\ a_2\ \dots\ a_{n-1}) \circ (a_{n-1}\ a_n).$$

Soit c' le second membre de cette égalité. Tout $x \in E$ distinct des a_i est fixé par c et c'.

Ensuite $\qquad c'(a_{n-1}) = (a_1\ a_2\ \dots\ a_{n-1})(a_n) = a_n = c(a_{n-1})$

et $\qquad c'(a_n) = (a_1\ a_2\ \dots\ a_{n-1})(a_{n-1}) = a_1 = c(a_n).$

Si $i \in [\![1, n-2]\!]$, $c'(a_i) = (a_1\ a_2\ \dots\ a_{n-1})(a_i) = a_{i+1} = c(a_i)$. Ainsi $c = c'$.

Soit c'' le second membre de l'égalité (22). Si $x \in E$ est distinct des a_i, $s(x)$ est distinct des $s(a_i)$. Ainsi $c(x) = x$ et $c''\big(s(x)\big) = s(x)$, d'où $(s \circ c)(x) = s(x) = (c'' \circ s)(x)$. Pour tout $i \in [\![1, n-1]\!]$, $c''\big(s(a_i)\big) = s(a_{i+1})$ et $c(a_i) = a_{i+1}$, donc :
$$(s \circ c)(a_i) = s(a_{i+1}) = (c'' \circ s)(a_i).$$

Enfin $c(a_n) = a_1$ et $c''\big(s(a_n)\big) = s(a_1)$, donc $(s \circ c)(a_n) = s(a_1) = (c'' \circ s)(a_n)$.

D'où $s \circ c = c'' \circ s$, ce qui prouve l'assertion 2 (composer à droite avec s^{-1}). ■

Exemple. Si E est un ensemble ayant au moins trois éléments, le *centre* de $\mathfrak{S}(E)$ (exercice 2 de la page 121) est trivial (en particulier, $\mathfrak{S}(E)$ *n'est alors pas commutatif*). Soit en effet $s \in \mathfrak{S}(E)$, $s \neq 1$. Il existe $a \in E$ tel que $s(a) \neq a$. Posons $b := s(a)$; soit $x \in E \setminus \{a, b\}$. La formule (22) donne :
$$s \circ (a\ x) \circ s^{-1} = \big(b\ s(x)\big) \neq (a\ x) \qquad \text{car } b \neq a, x.$$

Ainsi $s \circ (a\ x) \neq (a\ x) \circ s$: s ne commute déjà pas avec $(a\ x)$. D'où la conclusion.

Toute transposition t est d'ordre 2 (le vérifier) : $t \neq 1$ mais $t^2 = t \circ t = 1$, soit $t^{-1} = t$. Voici un résultat important concernant les transpositions :

Théorème 24.

1. Soit $n \geqslant 2$ un entier. Chacune des parties suivantes engendre $\mathfrak{S}_n$:
   $S := \{(1\ 2),\ (2\ 3),\ \ldots, (n-1\ n)\}$, $T := \{(1\ 2),\ (1\ 3),\ \ldots, (1\ n)\}$, et
   $U := \{(1\ 2),\ (1\ 2\ \cdots\ n)\}$.
2. Toute permutation d'un ensemble *fini* E s'écrit comme composée de transpositions. L'ensemble des transpositions engendre donc $\mathfrak{S}(E)$.

Démonstration. Posons $c := (1\ 2\ \cdots\ n)$, $s_k := (k\ k+1)$ pour $k \in [\![1, n-1]\!]$ et $t_k := (1\ k)$ pour $k \in [\![2, n]\!]$. Appliquons de manière répétée la formule de conjugaison (22). Si $k \in [\![2, n-1]\!]$, $s_k = t_k \circ t_{k+1} \circ t_k^{-1} \in \langle T \rangle$, d'où $S \subset \langle T \rangle$. De même, si $k \in [\![1, n-2]\!]$, $s_{k+1} = c \circ s_k \circ c^{-1}$; comme $s_1 \in U$, ces égalités montrent que $s_k \in \langle U \rangle$ pour tout k (récurrence sur k), d'où $S \subset \langle U \rangle$. Pour l'assertion 1, il suffit donc de prouver que $\langle S \rangle = \mathfrak{S}_n$.
Soit $s \in \mathfrak{S}_n$. Montrons que s s'écrit comme composée d'éléments de S, de sorte que $s \in \langle S \rangle$. C'est vrai si $s = 1$. Si $s \neq 1$, raisonnons par récurrence sur le plus grand $k \in [\![1, n]\!]$ tel que $s(k) \neq k$ ($k \geqslant 2$ et, si $k = 2$, $s = (1\ 2) = s_1$). Soit $h := s(k)$. D'abord $s(h) \neq h$, car sinon $h = s^{-1}(h) = k$. D'où $h < k$. Alors $t := (h\ k) \circ s$ laisse fixe k, et aussi tout j tel que $k < j \leqslant n$. Vu l'hypothèse de récurrence, t est un composé d'éléments de S. On conclut grâce aux égalités suivantes, à vérifier :

$$s = (h\ k) \circ t \quad \text{et} \quad (h\ k) = (s_h \circ s_{h+1} \circ \cdots \circ s_{k-2}) \circ s_{k-1} \circ (s_{k-2} \circ \cdots \circ s_{h+1} \circ s_h).$$

Pour l'assertion 2, soit $n := \operatorname{card}(E)$. Si $n = 1$, $\mathfrak{S}(E)$ est trivial, et l'ensemble des transpositions de E est vide, d'où la conclusion. Si $n \geqslant 2$, on se ramène au cas où $E := [\![1, n]\!]$ (*cf.* l'exercice II.2.12). La conclusion vient alors de l'alinéa précédent. ■

Exemple. Comment ranger des réels $x_1, \ldots, x_n$ par ordre croissant ? Si les x_i sont des mots, comment les classer par ordre alphabétique ? Ce sont là deux problèmes de *tri*. Parmi de nombreux algorithmes de tri, importants dans la vie courante, en voici un, le *tri à bulle*, lent mais simple à décrire.

1. Prendre $k := 1$.
2. S'arrêter si $k = n$ (le tri est alors terminé).
3. Si $x_k > x_{k+1}$, échanger x_k et x_{k+1}, puis retourner à l'étape 1.
4. Remplacer k par $k+1$ et retourner à l'étape 2.

Le rapport avec $\mathfrak{S}_n$ est le suivant : trier les x_i, c'est trouver une permutation $s \in \mathfrak{S}_n$ telle que la suite $(x_{s(1)}, \ldots, x_{s(n)})$ soit croissante. L'algorithme proposé fournit une telle permutation s, comme composée de transpositions appartenant à la partie S du théorème 24 (le vérifier).

Lemme 25. Soient $n \geqslant 0$ un entier et $s_1, \ldots, s_n$ des transpositions d'un ensemble E telles que $s_1 \circ s_2 \circ \cdots \circ s_n = 1$. $\quad (*)$
Alors n est *pair*.

Démonstration. En préliminaire, soient $s := (a\ b)$ et t deux transpositions, supposons que $t(a) = a$. Vu la formule (22), $t \circ s \circ t^{-1} = s'$, où $s' := \big(a\ t(b)\big)$. Ainsi $t \circ s = s' \circ t$, et a intervient dans s', mais pas dans t.

Prouvons le lemme par récurrence sur n. Le cas $n = 0$ est trivial, le cas $n = 1$ est exclu. Supposons $n \geqslant 2$, le résultat étant vrai jusqu'au rang $n - 1$. Écrivons $s_h := (a_h\ b_h)$ $(h = 1, \ldots, n)$, et posons $a := a_1$. Par une utilisation répétée du préliminaire, on remplace l'égalité $(*)$ par une égalité analogue, ayant la propriété suivante (par commodité, on ne change pas de notations) : il existe $m \in [\![1, n]\!]$ tel que $a = a_1 = a_2 = \cdots = a_m$ mais $s_h(a) = a$ pour $h = m + 1, \ldots, n$. Posant $t := s_{m+1} \circ \cdots \circ s_n$, on a $t(a) = a$, et $(*)$ s'écrit :

$$s_1 \circ s_2 \circ \cdots \circ s_m \circ t \ = \ (a\ b_1) \circ (a\ b_2) \circ \cdots \circ (a\ b_m) \circ t \ = \ 1.$$

Il existe $r \in [\![1, m-1]\!]$ tel que $b_r = b_m$ (donc $s_r = s_m$), sinon on aurait $a = (s_1 \circ \cdots \circ s_m \circ t)(a) = b_m$. Si $r = m-1$, s_{m-1} et s_m peuvent être ôtés de l'égalité $(*)$, d'où une égalité analogue comportant $n - 2$ transpositions. D'après l'hypothèse de récurrence, $n - 2$ est pair, donc n est pair.

Si $r \leqslant m-2$, soit $u := s_{r+1} \circ \cdots \circ s_{m-1}$. Par ailleurs, notons f l'automorphisme $s \mapsto s_m \circ s \circ s_m^{-1}$ de $\mathfrak{S}(E)$ (*cf.* l'exemple de la page 124). Comme $s_r = s_m = s_m^{-1}$, il vient, en posant $s'_h := f(s_h)$ pour $h = 1, \ldots, n$:

$$\overbrace{s_r \circ s_{r+1} \circ \cdots s_{m-1} \circ s_m}^{m-r+1 \text{ facteurs}} = s_m \circ u \circ s_m^{-1} = f(u) = \overbrace{s'_{r+1} \circ \cdots \circ s'_{m-1}}^{m-r-1 \text{ facteurs}}. \qquad (**)$$

Vu la formule (22) (toujours elle !), les s'_h sont des transpositions. Grâce à $(**)$, on transforme $(*)$ en une égalité analogue comportant $n - 2$ transpositions, ce qui montre comme ci-dessus que n est pair. ∎

La notion de *signature* d'une permutation, que nous allons définir maintenant, sera importante pour la définition et le calcul des *déterminants*. Dans ce qui suit, on désigne par E un ensemble fini, de cardinal $n \geqslant 1$.

Théorème et définition 26.

1. Si $n \geqslant 2$, il existe un *unique* morphisme ε *non trivial* de $\mathfrak{S}(E)$ dans le groupe multiplicatif $\{-1, 1\}$; si $n = 1$, on note $\varepsilon : \mathfrak{S}(E) \to \{-1, 1\}$ le morphisme trivial. Si $s \in \mathfrak{S}(E)$, $\varepsilon(s)$ est appelé *signature de* s ; s est dite *paire* si $\varepsilon(s) = 1$ et *impaire* sinon.
2. Soit $s \in \mathfrak{S}(E)$. Il existe un entier $k \geqslant 0$ et des transpositions $s_1, \ldots, s_k$ tels que $s = s_1 \circ s_2 \circ \cdots \circ s_k$. Alors $\varepsilon(s) = (-1)^k$.
3. Si c est un cycle de longueur $m \geqslant 2$, $\varepsilon(c) = (-1)^{m-1}$. En particulier, la signature d'une transposition est -1.

Démonstration. On peut supposer $n \geqslant 2$. Soit $s \in \mathfrak{S}(E)$. D'après le théorème 24 de la page précédente, il existe des transpositions $s_1, \ldots, s_k$ $(k \geqslant 0)$ telles que $s = s_1 \circ s_2 \circ \cdots \circ s_k$. Une telle écriture n'est pas unique. Si $s = t_1 \circ t_2 \circ \cdots \circ t_h$ est une écriture analogue,

$$1 = s \circ s^{-1} = (s_1 \circ s_2 \circ \cdots \circ s_k) \circ (t_1 \circ t_2 \circ \cdots \circ t_h)^{-1} = s_1 \circ s_2 \circ \cdots \circ s_k \circ t_h \circ \cdots \circ t_2 \circ t_1$$

Vu le lemme précédent, $h + k$ est pair, *i.e.* $(-1)^h = (-1)^k$. On peut donc poser $\varepsilon(s) := (-1)^k$, ceci pour *toute* écriture $s = s_1 \circ s_2 \circ \cdots \circ s_k$ du type ci-dessus. D'où une *application* $\varepsilon : \mathfrak{S}(E) \to \{-1, 1\}$, pour laquelle l'assertion 2 est vraie. En particulier $\varepsilon(t) = -1$ pour toute transposition t. La définition de ε et la formule (1) entraînent que ε est un morphisme. Si c est un m-cycle, la formule (21) montre que $\varepsilon(c) = (-1)^{m-1}$.

Soit $f : \mathfrak{S}(E) \to \{-1, 1\}$ un morphisme. Montrons que f est soit trivial, soit égal à ε. On peut

supposer que $E := [\![1,n]\!]$. Soit $\alpha = f\big((1\ 2)\big)$. Si $k \in [\![3,n]\!]$ il vient, en notant que $\{-1,1\}$ est abélien et en utilisant la formule (22) :

$$f\big((1\ k)\big) = f\left((2\ k)\circ(1\ 2)\circ(2\ k)^{-1}\right) = f\big((2\ k)\big)f\big((1\ 2)\big)f\big((2\ k)\big)^{-1} = \alpha.$$

Ainsi $f(t) = \alpha$ pour tout $t \in T$ (*cf.* le théorème 24 de la page 128). Si $\alpha = -1$, f et ε coïncident sur T, donc $f = \varepsilon$ (*cf.* le théorème 24 de la page 128 et l'exercice II.2.7 de la page 156). Si $\alpha = 1$, le même argument montre que f est trivial. ■

Le noyau de ε est appelé *groupe alterné* de E, on le note $\mathfrak{A}(E)$ $\big(\mathfrak{A}_n$ si $E = [\![1,n]\!]\big)$.
Si $n \geqslant 2$, l'ordre de $\mathfrak{A}(E)$ est $n!/2$, vu la formule (20).

Proposition 27. Soient $n \geqslant 2$ un entier et $s \in \mathfrak{S}_n$. Appelons *inversion de* s tout couple d'entiers (i,j) tel que $1 \leqslant i < j \leqslant n$ et $s(i) > s(j)$. Notons $N(s)$ le nombre d'inversions de s. Alors $\varepsilon(s) = (-1)^{N(s)}$.

Démonstration. C'est vrai si $N(s) = 0$, car alors $s = 1$ (pourquoi ?). Si $N(s) \geqslant 1$, soit (i,j) une inversion de s telle que $j - i$ soit minimum. Alors $j = i+1$: si $j \geqslant i+2$, $(i,j-1)$ et $(j-1,j)$ ne sont pas des inversions de s, donc $s(i) < s(j-1) < s(j)$, c'est absurde. Soit $t := s \circ (i\ i+1)$. Alors $t(i) = s(i+1)$, $t(i+1) = s(i)$, et $t(k) = s(k)$ pour tout $k \in [\![1,n]\!]$ distinct de $i, i+1$. On en déduit que $N(t) = N(s) - 1$ (le vérifier avec soin, à titre d'exercice). Or $\varepsilon(t) = -\varepsilon(s)$, d'où la conclusion, par récurrence sur $N(s)$. ■

2.4 Groupe additif des entiers modulo n

Dans cette section, nous définissons les groupes additifs $\mathbb{Z}/n\mathbb{Z}$ (« entiers modulo n »). Ce sont des cas particuliers de quotients, mais la notion générale de groupe quotient sera étudiée en L2.
Comme introduction, notons qu'un entier relatif a deux parités possibles : il est pair ou impair. La somme de deux entiers impairs est paire, la somme d'un entier pair et d'un entier impair est impaire, etc. D'où la « table d'addition » suivante, à gauche :

	pair	impair
pair	pair	impair
impair	impair	pair

	$\overline{0}$	$\overline{1}$
$\overline{0}$	$\overline{0}$	$\overline{1}$
$\overline{1}$	$\overline{1}$	$\overline{0}$

La parité d'un entier k est déterminée par le reste dans la division de k par 2 ; ce reste, noté $k \bmod 2$, vaut 0 ou 1. En remplaçant « pair » et « impair » par $\overline{0}$ et $\overline{1}$, on obtient la table à droite. On écrit $\overline{1}$ et non pas 1, car l'addition $1 + 1 = 0$ pourrait laisser perplexe !
La table ci-dessus définit une loi de groupe sur $\{\overline{0},\overline{1}\}$, le groupe obtenu est $\mathbb{Z}/2\mathbb{Z}$. Avant de remplacer 2 par un autre entier, donnons une définition :

Définition 10. Soit $n \geqslant 1$ un entier. On dit que $x, y \in \mathbb{Z}$ sont *congrus modulo* n, et l'on note $x \equiv y \mod n$, si n divise $x - y$, *i.e.* si $x \bmod n = y \bmod n$.

Divisons x et y par n : $x = an + r$, $y = bn + s$; ainsi $r = x \bmod n$ et $s = y \bmod n$. Comme $x - y = n(a-b) + (r-s)$, n divise $x - y$ si, et seulement s'il divise $r - s$, *i.e.*

si $r = s$, puisque $|r - s| \in [\![0, n-1]\!]$. Sur $\mathbb{Z}$, la relation $\mathcal{R}$ définie par $x\mathcal{R}y$ si, et seulement si, $x \equiv y \mod n$ est évidemment une relation d'équivalence, appelée *congruence modulo* n.

Soit $n \geqslant 1$ un entier *fixé*. Définissons l'ensemble $\mathbb{Z}/n\mathbb{Z}$ ainsi :

$$\mathbb{Z}/n\mathbb{Z} := \{\overline{0}, \overline{1}, \ldots, \overline{n-1}\}.$$

Les éléments de $\mathbb{Z}/n\mathbb{Z}$ sont simplement n symboles distincts $\overline{0}, \overline{1}, \ldots, \overline{n-1}$, peu importe leur « nature ». On définit une application $k \mapsto \overline{k}$ de $\mathbb{Z}$ dans $\mathbb{Z}/n\mathbb{Z}$ comme suit : si $k \in \mathbb{Z}$, $\overline{k}$ est par définition égal à $\overline{r}$, où $r := k \mod n$ est le reste dans la division de k par n. Pour tous $x, y \in \mathbb{Z}$, on a donc :

$$\left(\overline{x} = \overline{y}\right) \iff \left(x \mod n = y \mod n\right) \iff \left(x \equiv y \mod n\right). \tag{23}$$

Pour cette raison, les éléments de $\mathbb{Z}/n\mathbb{Z}$ sont appelés *entiers modulo* n. Bien entendu, la notation $\overline{k}$ ne peut être utilisée que si n est clair dans le contexte, et s'il n'y a pas confusion avec la conjugaison dans $\mathbb{C}$!

Pour définir une addition sur $\mathbb{Z}/n\mathbb{Z}$, partons du lemme évident suivant :

Lemme 28. Soit $n \geqslant 1$ un entier. Soient $x, x', y, y' \in \mathbb{Z}$ des entiers. Supposons que $x \equiv x' \mod n$ et $y \equiv y' \mod n$. Alors $x + y \equiv x' + y' \mod n$.

Démonstration. Elle est immédiate : n divise $x - x'$ et $y - y'$, donc il divise leur somme, qui vaut $(x - x') + (y - y') = (x + y) - (x' + y')$. ∎

Une conséquence de ce lemme est que, si $a, b \in \mathbb{Z}$, $\overline{a+b} \in \mathbb{Z}/n\mathbb{Z}$ ne dépend que de $\overline{a}, \overline{b} \in \mathbb{Z}/n\mathbb{Z}$, et pas de a et b eux-mêmes. Il est donc licite de définir une addition sur $\mathbb{Z}/n\mathbb{Z}$ en posant, pour tous $x, y \in \mathbb{Z}$:

$$\overline{x} + \overline{y} := \overline{x+y}. \tag{24}$$

Par exemple $2 + 3 = 5$ et $5 \mod 4 = 1$, donc $\overline{2} + \overline{3} = \overline{1}$ dans $\mathbb{Z}/4\mathbb{Z}$.

Exercice 4.

Il est midi. Quelle heure sera-t-il dans 41 heures ?

Solution. Calculons dans $\mathbb{Z}/24\mathbb{Z}$: $12 + 41 = 53 = 2 \times 24 + 5$, donc $\overline{53} = \overline{5}$.

Il sera donc 5 heures.

Proposition 29. Soit $n \geqslant 1$ un entier. Muni de l'addition définie ci-dessus, $\mathbb{Z}/n\mathbb{Z}$ est un *groupe*. Ce groupe est *cyclique* d'ordre n, et $\overline{1}$ en est un générateur.

Démonstration. L'associativité de l'addition sur $\mathbb{Z}/n\mathbb{Z}$ résulte de l'associativité de l'addition sur $\mathbb{Z}$ et de (24). De même $\overline{0}$ est élément neutre et, si $x \in \mathbb{Z}$, $\overline{-x}$ est opposé de $\overline{x}$. Ainsi $(\mathbb{Z}/n\mathbb{Z}, +)$ est un groupe. Si $k \in \mathbb{N}$, (24) montre, par récurrence sur k, que $\overline{k} = k \cdot \overline{1}$; en passant aux opposés, on en déduit que $\overline{k} = k \cdot \overline{1}$ pour tout $k \in \mathbb{Z}$, donc $\overline{1}$ engendre $\mathbb{Z}/n\mathbb{Z}$. ∎

L'application $\pi : k \mapsto \overline{k}$ de $\mathbb{Z}$ dans $\mathbb{Z}/n\mathbb{Z}$ est, vu la définition de l'addition de $\mathbb{Z}/n\mathbb{Z}$, un morphisme de groupes, appelé *morphisme canonique* ; il est surjectif

et $\mathrm{Ker}(\pi) = n\mathbb{Z}$, d'après les équivalences (23). Pour un ordinateur, l'ensemble $\mathbb{Z}/n\mathbb{Z}$ est simplement $[\![0, n-1]\!]$, l'addition étant définie par $(a, b) \mapsto (a + b) \bmod n$: il calcule la somme $a + b$ dans $\mathbb{Z}$, puis la remplace par son reste modulo n.
Voici une réciproque de la proposition précédente :

Théorème 30. Un groupe monogène est isomorphe soit à $\mathbb{Z}$ s'il est infini, soit à $\mathbb{Z}/n\mathbb{Z}$ s'il est fini (donc cyclique) d'ordre $n \geqslant 1$.

Démonstration. Soient $(G, \times)$ un groupe monogène et x un générateur de G. Appliquons le théorème et définition 21 de la page 125, en reprenant ses notations. Si f est injectif, c'est un isomorphisme de $\mathbb{Z}$ sur $\langle x \rangle = G$, donc G est infini. Sinon, $\mathrm{Ker}(f) = n\mathbb{Z}$ pour un certain $n \geqslant 1$. Alors $G = \{1, x, x^2, \ldots, x^{n-1}\}$ est cyclique d'ordre n. Soit $u : \mathbb{Z}/n\mathbb{Z} \to G$ l'application bijective $\overline{k} \mapsto x^k$, $k = 0, 1, \ldots, n-1$. Montrons que c'est un morphisme. Soient $a, b \in [\![0, n-1]\!]$, divisons $a + b$ par n : $a + b = nq + r$. Alors $\overline{a} + \overline{b} = \overline{r}$, d'où, puisque $x^n = 1$:

$$u\big(\overline{a} + \overline{b}\big) = u\big(\overline{r}\big) = x^r = x^r (x^n)^q = x^{a+b} = x^a x^b = u\big(\overline{a}\big) u\big(\overline{b}\big). \qquad \blacksquare$$

Exercice 5.
Soient G un groupe (multiplicatif) cyclique d'ordre $n \geqslant 1$ et a un générateur de G. Soient $k \in \mathbb{Z}$ et d le pgcd de k et n ; écrivons $k = dk'$, $n = dn'$ ($k' \in \mathbb{Z}$, $n' \in \mathbb{N}$). Montrer que l'ordre de a^k est n', et que a^k est un générateur de G si, et seulement si, k est premier avec n.
Solution. Soit $t \in \mathbb{Z}$. Par définition de l'ordre d'un élément (*cf.* le théorème et définition 21 de la page 125), on a les équivalences :

$$\big((a^k)^t = 1\big) \iff (a^{kt} = 1) \iff (n \mid kt) \iff (n' \mid k't) \iff (n' \mid t);$$

puisque $k' \wedge n' = 1$, la dernière équivalence résulte du théorème de Gauß (théorème 54 de la page 97). Ainsi $\{t \in \mathbb{Z} \mid (a^k)^t = 1\} = n'\mathbb{Z}$, donc l'ordre de a^k est n'. Enfin a^k est un générateur de G si, et seulement si, son ordre est n, *i.e.* si $n' = n$, autrement dit si $k \wedge n = 1$.

3 Anneaux

3.1 Définitions, règles de calcul

Définition 11. On appelle *anneau* tout ensemble A muni de deux lois appelées *addition* et *multiplication*, notées en général $+$ et $\times$, telles que :

1. L'addition est une loi de groupe *commutative* sur A.
2. La multiplication est *associative*.
3. La multiplication est *distributive par rapport à l'addition*.
   Cela veut dire que, pour tous $x, y, z \in A$, on a :
   $$x(y + z) = xy + xz \quad \text{et} \quad (y + z)x = yx + zx. \tag{25}$$
4. Pour la multiplication, A possède un élément neutre, noté 1.

L'anneau A est dit *commutatif* si la multiplication est commutative.

Munis de leur addition et leur multiplication usuelles, $\mathbb{Z}$ et $\mathbb{Q}$ sont des anneaux commutatifs. Les deux lois d'un anneau A ne jouent pas le même rôle : $+$ est commutative, mais $\times$ peut ne pas l'être ; de même $(A,+)$ est un groupe, alors que $(A,\times)$ n'est un groupe que si A est « nul ». La seconde loi s'appelle *multiplication* mais peut être notée par un autre symbole que $\times$ (par exemple $\circ$), les égalités (25) devant être modifiées en conséquence. Nous garderons les notations $+$ et $\times$ et parlerons de l'anneau $(A,+,\times)$.
Le neutre de $(A,+)$ est noté 0. Si l'on veut éviter toute confusion (avec les nombres $0,1$ par exemple), $0,1 \in A$ sont notés 0_A, 1_A. Soit $x \in A$. Puisque $(A,+)$ est un groupe, x a un *opposé* $-x$. Ensuite, $+$ et $\times$ étant associatives et A ayant pour chacune un élément neutre, les suites $(nx)_{n\geqslant 0}$ et $(x^n)_{n\geqslant 0}$ sont définies (revoir leurs propriétés à la proposition 1 de la page 113) ; nx est parfois noté $n \cdot x$, pour ne pas confondre avec les produits ab $(a,b \in A)$.
Parmi les règles de calcul dans l'anneau A, il y a celles qui ne font intervenir que $+$, *i.e.* les règles de calcul dans le groupe $(A,+)$. Il y a ensuite celles qui font partie de la définition d'un anneau (distributivité, associativité de $\times$). Nous ne réécrirons pas toutes ces règles. Il y a enfin les règles se déduisant des précédentes, *cf.* la proposition qui va suivre.
Disons que deux éléments a,b de A *commutent* (ou que a *commute avec* b) si $ab = ba$. Si c'est le cas, toute puissance de a commute avec toute puissance de b (*cf.* la remarque de la page 113).

Proposition 31. Soit $(A,+,\times)$ un anneau. Pour des éléments $x,y \in A$ et $n \in \mathbb{Z}$ quelconques, on a :

1. $0_A\, x \;=\; 0_A \;=\; x\, 0_A$.
2. $(-x)y \;=\; -(xy) \;=\; x(-y)$.
3. $(nx)y \;=\; n(xy) \;=\; x(ny)$. Ainsi $n(xy)$ pourra être noté nxy.
4. $(n \cdot 1_A)x \;=\; nx \;=\; x(n \cdot 1_A)$.

Démonstration. Par distributivité, $0_A\, x = (0_A + 0_A)x = 0_A\, x + 0_A\, x$.
Par régularité dans $(A,+)$, $0_A\, x = 0_A$. De même $x\, 0_A = 0_A$, d'où 1. Par distributivité, $0_A = 0_A\, y = \big(x + (-x)\big)y = xy + (-x)y$, d'où l'on déduit $(-x)y = -(xy)$, et de même pour la deuxième égalité en 2.
L'égalité $(nx)y = n(xy)$ est vraie si $n := 0$, vu 1 ; elle est vraie si $n > 0$ par distributivité : par exemple $nx = x+x+\cdots+x$ (n termes). Si $n < 0$, soit $k := -n$. Alors $nx = -(kx)$, donc $(nx)y = \big(-(kx)\big)y = -\big((kx)y\big)$, puis $(nx)y = -\big(k(xy)\big) = (-k)(xy) = n(xy)$.
De même pour $x(ny) = n(xy)$.
Prenons $y := 1_A$ dans 3 : $nx = n(x1_A) = x(n1_A) = x(n \cdot 1_A)$.
La première égalité en 4 se démontre de même. ∎

Exemples.

1. Dans un anneau A, $0_A = 1_A$ si, et seulement si, $A = \{0_A\}$: si $0_A = 1_A$ et si $x \in A$, $x = x\, 1_A = x\, 0_A = 0_A$, donc $A = \{0_A\}$; la réciproque est triviale. Un tel anneau possédant un seul élément est dit *nul*.

2. Soient $A_1, \ldots, A_n$ des anneaux ($n \geqslant 1$). Alors $A := A_1 \times \cdots \times A_n$, muni de l'addition (resp. multiplication) *produit* de celles des A_i, est un anneau (le vérifier), appelé *anneau produit* $A_1 \times \cdots \times A_n$.
3. Soient A un anneau et X un ensemble. Nous avons vu que $\big(\mathcal{F}(X, A), +\big)$ est un groupe abélien . On a aussi une multiplication associative sur $\mathcal{F}(X, A)$, avec pour élément neutre l'application constante : $x \mapsto 1$. Si $f, g \in \mathcal{F}(X, A)$, le produit fg est défini ainsi :
$$(fg)(x) := f(x)g(x) \quad \text{pour tout } x \in X. \tag{26}$$
Alors $\big(\mathcal{F}(X, A), +, \times\big)$ *est un anneau* (vérifier la distributivité).

Exercice 6.
Soit $(E, +)$ un groupe abélien. Sur le groupe abélien $\big(\mathcal{F}(E, E), +\big)$, prenons comme multiplication la loi $\circ$; obtient-on un anneau ?
Solution. Soient $f, g, h \in \mathcal{F}(E, E)$. D'abord $(f + g) \circ h = (f \circ h) + (g \circ h)$ (le vérifier), mais l'égalité $f \circ (g+h) = (f \circ g) + (f \circ h)$ est fausse en général : soient $a, b \in E$ et g, h les applications constantes $x \mapsto a$ et $x \mapsto b$. L'égalité visée équivaut à $f(a+b) = f(a) + f(b)$, ce qui peut être faux.

Voici maintenant une importante règle de calcul dans un anneau. Cette formule, dite de Newton (1642-1727), a de nombreuses applications.

Théorème 32 (Formule du binôme). Dans un anneau $(A, +, \times)$, soient a, b deux éléments qui *commutent*. Pour tout entier $n \geqslant 0$, on a alors :
$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k}\, b^k. \qquad \text{(formule du binôme)} \tag{27}$$

Démonstration. Par récurrence sur n. Le cas $n = 0$ est évident : $\binom{0}{0} = 1$ et $a^0 = b^0 = (a + b)^0 = 1_A$. Supposons (27) vraie pour un entier n. Alors :
$$\begin{aligned}(a + b)^{n+1} &= (a + b)(a + b)^n \\ &= a\left(\sum_{k=0}^{n} \binom{n}{k} a^{n-k}\, b^k\right) + b\left(\sum_{k=0}^{n} \binom{n}{k} a^{n-k}\, b^k\right) \\ &= \overbrace{\sum_{k=0}^{n} \binom{n}{k} a^{n-k+1}\, b^k}^{S} + \overbrace{\sum_{k=0}^{n} \binom{n}{k} a^{n-k}\, b^{k+1}}^{T}.\end{aligned}$$
Parmi ces égalités, la 2ème vient de la distributivité ; la 3ème vient des égalités 3 de la proposition précédente et du fait que b commute avec toute puissance de a. Transformons T via la formule de changement d'indice (8). Puisque $t : h \mapsto h-1$ est une bijection de $[\![1, n+1]\!]$ sur $[\![0, n]\!]$, on a :
$$T = \sum_{k=0}^{n} \binom{n}{k} a^{n-k}\, b^{k+1} = \sum_{h=1}^{n+1} \binom{n}{h-1} a^{n-(h-1)}\, b^h = \sum_{k=1}^{n+1} \binom{n}{k-1} a^{n+1-k}\, b^k.$$

C'est la 2ème de ces égalités qui vient de la formule (8) (on a « posé » $k = h - 1$) ; pour la 3ème égalité, on a simplement remplacé l'indice muet h par k. Ainsi :

$$\begin{aligned}
(a+b)^{n+1} &= S + T = \sum_{k=0}^{n} \binom{n}{k} a^{n-k+1}\, b^k + \sum_{k=1}^{n+1} \binom{n}{k-1} a^{n+1-k}\, b^k \\
&= \binom{n}{0} a^{n+1} b^0 + \sum_{k=1}^{n} \left[\overbrace{\binom{n}{k} + \binom{n}{k-1}} \right] a^{n+1-k}\, b^k + \binom{n}{n} a^0\, b^{n+1} \\
&= \binom{n+1}{0} a^{n+1} b^0 + \sum_{k=1}^{n} \overbrace{\binom{n+1}{k}}\, a^{(n+1)-k}\, b^k + \binom{n+1}{n+1} a^0\, b^{n+1} \\
&= \sum_{k=0}^{n+1} \binom{n+1}{k} a^{(n+1)-k}\, b^k.
\end{aligned}$$

Nous avons utilisé la relation de Pascal (5) de la page 79, ainsi que les égalités suivantes (où $m \in \mathbb{N}$) : $\binom{m}{0} = 1 = \binom{m}{m}$. L'égalité (27) est donc vraie au rang $n+1$. ■

Exemple. Dans la formule du binôme, prenons $A := \mathbb{Z}$, $a := 1$, $b := \pm 1$ et $n \geqslant 1$. Puisque $(1+1)^n = 2^n$ et $(1-1)^n = 0^n = 0$, on obtient ceci :

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n \quad \text{et} \quad \sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0.$$

La première égalité dit que la somme des termes de la ligne d'indice n du triangle de Pascal vaut 2^n ; on aurait pu la déduire du théorème 26 de la page 76. La deuxième égalité dit que, dans un ensemble fini non vide, il y a autant de parties de cardinal pair que de parties de cardinal impair.

Comme autre exemple de calcul dans un anneau commutatif A, soient I un ensemble fini et $(x_i)_{i \in I}$, $(y_i)_{i \in I}$ deux familles d'éléments de A. Si $J \in \mathcal{P}(I)$, soit X_J le produit des x_i pour $i \in J$, de même pour Y_J. Alors :

$$\prod_{i \in I} (x_i + y_i) = \sum_{J \in \mathcal{P}(I)} X_{I \setminus J}\, Y_J\,. \tag{28}$$

On peut supposer que $I = [\![1, n]\!]$. Le produit P à gauche s'écrit alors :

$$P = (x_1 + y_1)(x_2 + y_2) \cdots (x_n + y_n).$$

En développant P par distributivité, on voit que P est somme de 2^n termes, chaque terme étant le produit d'un élément (x_1 ou y_1) de la première parenthèse, d'un élément de la deuxième parenthèse, etc. La formule (28) s'en déduit aussitôt (nous laissons les détails au lecteur).

Exercice 7.

Soient X un ensemble fini et $A_1, \ldots, A_n$ $(n \geqslant 2)$ des parties de X ; notons A la réunion des A_i. Démontrer la *formule du crible* :

$$\operatorname{card}(A) = \sum_{1 \leqslant i \leqslant n} \operatorname{card}(A_i) - \sum_{1 \leqslant i < j \leqslant n} \operatorname{card}(A_i \cap A_j)$$
$$+ \sum_{1 \leqslant i < j < k \leqslant n} \operatorname{card}(A_i \cap A_j \cap A_k) + \cdots + (-1)^n \operatorname{card}(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Solution. Pour toute partie A de X, notons χ_A la fonction caractéristique de A (définition 4 de la page 76), que nous considérons comme élément de l'anneau commutatif $\mathcal{F}(X, \mathbb{Z})$. Dans cet anneau, voici des formules valables pour toutes parties A, B de X et résultant de la définition des fonctions caractéristiques :

$$\sum_{x \in X} \chi_A(x) \;=\; \operatorname{card}(A)\,, \;\; \chi_{X \setminus A} \;=\; 1 - \chi_A\,, \;\; \chi_{A \cap B} \;=\; \chi_A\, \chi_B. \tag{29}$$

Ainsi, soit $x \in X$. Puisque $\chi_A(x), \chi_B(x) \in \{0, 1\}$, $\chi_A(x)\chi_B(x) = 1$ si, et seulement si, $\chi_A(x) = \chi_B(x) = 1$, *i.e.* si $x \in A \cap B$. D'où la dernière formule.
Introduisons les notations suivantes : pour tout $i \in [\![1, n]\!]$ on pose $B_i := X \setminus A_i$, de sorte que $\chi_{B_i} = 1 - \chi_{A_i}$. On pose de même $B := X \setminus A$. Puisque A est la *réunion* des A_i, B est l'*intersection* des B_i (*cf.* le module I.1), et par suite χ_B est le produit des χ_{B_i}. D'où :

$$\chi_B \;=\; \prod_{i=1}^{n} \chi_{B_i} \;=\; \prod_{i=1}^{n} (1 - \chi_{A_i}).$$

Appliquons la formule (28) avec, pour tout i, $x_i := 1$ et $y_i := -\chi_{A_i}$. Pour toute partie J de $[\![1, n]\!]$, de cardinal k, $X_{I \setminus J} = 1$ et $Y_J = (-1)^k \chi_{A_J}$, en notant A_J l'intersection des A_i, $i \in J$; si $J = \varnothing$, $Y_J = 1$ puisque $A_J = X$.

$$\chi_B = 1 + \sum_{k=1}^{n} (-1)^k \left(\sum_{\operatorname{card}(J) = k} \chi_{A_J} \right), \text{ soit } \chi_A = \sum_{k=1}^{n} (-1)^{k-1} \left(\sum_{\operatorname{card}(J) = k} \chi_{A_J} \right).$$

Dans chaque parenthèse, J décrit l'ensemble des parties de cardinal k de $[\![1, n]\!]$. Faisons la somme des valeurs de chaque membre de la dernière égalité en tous les $x \in X$. On obtient la formule voulue, écrite de façon formelle :

$$\operatorname{card}(A) \;=\; \sum_{k=1}^{n} (-1)^{k-1} \left(\sum_{\operatorname{card}(J) = k} \operatorname{card}(A_J) \right). \tag{30}$$

Un élément d'un anneau peut avoir diverses propriétés. Voici trois exemples.

Définition 12. Soient $(A, +, \times)$ un anneau et $x \in A$.

1. On dit que x est *inversible* s'il possède un inverse x^{-1} pour la multiplication. L'ensemble des éléments inversibles de A sera noté $A^\times$.
2. On dit que x est *idempotent* si $x^2 = x$; on dit qu'il est *nilpotent* si l'une de ses puissances est nulle.

On notera $A^* := A \setminus \{0\}$ l'ensemble des éléments non nuls, qu'on prendra garde à ne pas confondre avec l'ensemble $A^\times$ des éléments inversibles.
Par exemple $\mathbb{Q}^\times = \mathbb{Q}^*$, alors que $\mathbb{Z}^\times = \{-1, 1\}$.
Dans un anneau A, 0 et 1 sont idempotents ; si $x \in A$ est idempotent, $1 - x$ l'est aussi, car $(1-x)^2 = 1 - 2x + x^2 = 1 - 2x + x = 1 - x$. Si $x, y \in A$ sont idempotents et *commutent*, xy est idempotent : $(xy)^2 = x^2y^2 = xy$.
Les idempotents seront importants en algèbre linéaire (« projecteurs » d'un espace vectoriel), les éléments nilpotents aussi.

Exemple. Soit X un ensemble. Sur l'ensemble $\mathcal{P}(X)$ des parties de X, on définit deux lois $+$ et $\times$ en posant, pour toutes parties S, T de X :

$$S + T := (S \setminus T) \cup (T \setminus S) \qquad \text{et} \qquad ST := S \cap T;$$

$S + T$ est appelée *différence symétrique* de S et T ; c'est l'ensemble des éléments de X appartenant à une et une seule des parties S, T.
Voici des propriétés évidentes : ces deux lois sont commutatives, $\times$ est associative, $\varnothing$ est élément neutre pour $+$ et X est élément neutre pour $\times$. À titre d'exercice, démontrer l'associativité de $+$. Nous verrons, dans l'exercice II.2.37 de la page 159 que $\big(\mathcal{P}(X), +, \times\big)$ est un anneau.
Cet anneau est un exemple d'anneau dans lequel tout élément x vérifie : $2x = x + x = 0$ et $x^2 = x$; *tout* élément est donc idempotent. En effet, si $S \in \mathcal{P}(X)$, $S + S = \varnothing$ et $S \cap S = S$. Un tel anneau est appelé *anneau de Boole*. Vérifier en exercice qu'un anneau de Boole est commutatif.

Dans un anneau A, l'ensemble $A^\times$ des éléments inversibles contient 1, il est multiplicativement stable (proposition 6 de la page 116). On munit donc $A^\times$ de la multiplication induite par celle de A. Alors $1 \in A^\times$ est élément neutre de $A^\times$ et, si $x \in A^\times$, son inverse x^{-1} dans A est inverse de x dans $A^\times$. Ainsi $(A^\times, \times)$ est un groupe, appelé *groupe des éléments inversibles de* A.
Dans $\mathbb{Z}$, le produit de deux entiers non nuls est non nul. Par contre, dans l'anneau produit $A := \mathbb{Z}^2$, $(1, 0)$ et $(0, 1)$ sont distincts de $0_A = (0, 0)$, mais $(1, 0) \times (0, 1) = 0_A$. Cela motive la définition suivante :

Définition 13. Un anneau $(A, +, \times)$ est dit *intègre* s'il est *commutatif*, si $0 \neq 1$, et si le produit de deux éléments non nuls quelconques de A est non nul.

La dernière condition de cette définition peut être exprimée ainsi :

PROPRIÉTÉ. *Pour tous* $a, b \in A$, $(ab = 0) \Longrightarrow (a = 0)$ ou $(b = 0)$.

Voici un résultat très simple, mais important :

Proposition 33. Soient $(A, +, \times)$ un anneau et $x \in A$. On suppose que l'une (au moins) des deux conditions suivantes est satisfaite :

1. L'élément x de A est inversible.
2. L'anneau A est intègre et $x \neq 0$.

Alors x est *régulier* : pour tous $y, z \in A$, on a les implications :

$$(xy = xz) \Longrightarrow (y = z) \quad \text{et} \quad (yx = zx) \Longrightarrow (y = z). \tag{31}$$

Démonstration. Si x est inversible, on applique le théorème 5 de la page 116 à $(A, \times)$. Si A est intègre, soient $y, z \in A$ tels que $xy = xz$. Les égalités 2 de la proposition 31 de la page 133 et la distributivité donnent : $xy - xz = xy + x(-z) = x\big[y + (-z)\big]$, d'où $x(y - z) = 0$. Ainsi $y - z = 0$ (*cf.* la propriété ci-dessus), *i.e.* $y = z$. ■

La propriété suivante est évidente, comme le lecteur le vérifiera :

PROPRIÉTÉ. *Dans un anneau* intègre*, les seuls éléments idempotents sont* 0 *et* 1*, et le seul élément nilpotent est* 0.

3.2 Sous-anneaux, idéaux, morphismes

Dans ce qui suit, sauf mention du contraire, on note $(x, y) \mapsto x + y$ *et* $(x, y) \mapsto xy$ *les lois d'un anneau et* $0, 1$ *les deux éléments neutres.*

Dans un anneau, il y a deux notions analogues à celle de sous-groupe d'un groupe : ce sont les notions de « sous-anneau » et d'« idéal ». Il importe de bien distinguer ces deux notions, en apparence voisines.

Un peu d'histoire

L'origine du terme d'« idéal » est historique. Le « dernier » théorème de Fermat dit que, pour des entiers $n \geqslant 3$ et $x, y, z \neq 0$, l'égalité $x^n + y^n = z^n$ est impossible. Fermat (1601-1665) a affirmé détenir la preuve de ce théorème (sans la communiquer), et depuis les mathématiciens ont cherché une telle preuve. Sans succès jusqu'en 1994, date à laquelle une preuve, extraordinairement difficile, a été donnée, essentiellement par Wiles (1953 -).
Kummer (1810-1893) s'est inspiré du cas $n = 2$ (*cf.* l'exercice II.1.14 de la page 107), remplaçant $\mathbb{Z}$ par un certain sous-anneau A de $\mathbb{C}$. Il a prouvé que, **si** l'analogue du « théorème de factorisation » (théorème 37 de la page 86) est vrai dans cet anneau A, dépendant de n, le théorème de Fermat est vrai pour l'exposant n. Il se trouve que l'hypothèse ci-dessus est satisfaite pour $n \leqslant 22$, mais pas pour 23, par exemple. Kummer a vu ensuite que le théorème de factorisation est vrai dans A (sans que le théorème de Fermat en résulte !), *si* l'on remplace les éléments de A par des « éléments idéaux », qui sont en substance des idéaux de A au sens de la définition ci-dessous.

Définition 14. On appelle *sous-anneau* d'un anneau A tout sous-groupe de $(A, +)$ contenant 1 et multiplicativement stable.

Définition 15. Soit A un anneau *commutatif*. On appelle *idéal* de A tout sous-groupe I de $(A,+)$ vérifiant la condition suivante : pour tout $x \in I$ et tout $a \in A$, l'élément ax appartient à I.

Lorsque nous parlerons d'idéal d'un anneau, il sera donc sous-entendu que cet anneau est commutatif. Voici deux résultats évidents.

PROPRIÉTÉ. *Dans un anneau commutatif A, tout idéal contenant 1 est égal à A. Le seul sous-anneau de A qui soit un idéal de A est A.*

Soit I un idéal de A contenant 1. Soit $a \in A$. Puisque $a \in A$ et $1 \in I$, $a = a \times 1 \in I$ (définition d'un idéal). D'où $I = A$. Si I est un idéal *et* un sous-anneau de A, $1 \in I$ (définition d'un sous-anneau), donc $I = A$.

Proposition 34. Le seul sous-anneau de $\mathbb{Z}$ est $\mathbb{Z}$ lui-même. Les idéaux de $\mathbb{Z}$ sont exactement les $n\mathbb{Z}$, n décrivant $\mathbb{N}$.

Démonstration. Soit T un idéal ou sous-anneau de $\mathbb{Z}$; c'est déjà un sous-groupe de $(\mathbb{Z},+)$. Il existe donc un unique $n \in \mathbb{N}$ tel que $T = n\mathbb{Z}$. Si T est un sous-anneau de $\mathbb{Z}$, $1 \in T = n\mathbb{Z}$, d'où $n = 1$ et donc $T = \mathbb{Z}$.

Il reste à voir que, si $n \in \mathbb{N}$, $n\mathbb{Z}$ est un idéal de $\mathbb{Z}$. Soient $a \in \mathbb{Z}$ et $x \in n\mathbb{Z}$; il existe $t \in \mathbb{Z}$ tel que $x = nt$. Alors $ax = a(nt) = n(at) \in n\mathbb{Z}$. ∎

Remarque. Les idéaux de $\mathbb{Z}$ sont exactement les sous-groupes de $(\mathbb{Z},+)$. Tout idéal d'un anneau A est un sous-groupe de $(A,+)$, l'inverse peut être faux : $\mathbb{Z}$ est un sous-anneau, mais pas un idéal, de $\mathbb{Q}$.

Soit T un sous-anneau ou idéal de l'anneau A. D'abord, T est additivement et multiplicativement stable, on peut donc le munir des lois $+$ et $\times$ induites par celles de A. Si T est un sous-anneau de A, $1 \in T$, et alors $(T,+,\times)$ est un anneau (le vérifier), d'où le terme de sous-anneau.

Exemples.

1. Dans tout anneau A, A est un sous-anneau et un idéal de A, alors que $\{0\}$ est un idéal de A, mais pas un sous-anneau, sauf si A est nul.
2. Toute intersection de sous-anneaux (ou idéaux) d'un anneau A est un sous-anneau (ou idéal) de A (proposition 7 de la page 117 et exercice 1 de la page 121).
3. Soient X un ensemble, A un anneau commutatif. On considère l'anneau $\mathcal{F}(X,A)$ (exemple 3 de la page 134). Soient Y une partie de X et J un idéal de A fixés. Soit enfin I l'ensemble des $f \in \mathcal{F}(X,A)$ telles que $\forall y \in Y,\ f(y) \in J$. Il est aisé de montrer que I est un idéal de $\mathcal{F}(X,A)$ (le faire). Traduction lorsque $X \subset A = \mathbb{R}$ et $J = \{0\}$?

L'exemple 2 justifie la définition suivante :

Définition 16. Soient A un anneau et S une partie de A.

1. On appelle sous-anneau de A *engendré par* S l'intersection de *tous* les sous-anneaux de A contenant S (A est un tel sous-anneau).
2. Supposons A commutatif. On appelle idéal de A *engendré par* S l'intersection de *tous* les idéaux de A contenant S (A est un tel idéal). Cet idéal sera noté (S).

Si $S := \{a\}$ est un singleton, (S) sera noté (a). Pour les sous-anneaux ou les idéaux, l'analogue de la proposition 15 de la page 122 est vrai (le vérifier) : (S) est, au sens de l'inclusion, le *plus petit* des idéaux de A contenant S.

Proposition 35. Soient A un anneau commutatif et $I_1, \ldots, I_n$ des idéaux de A ($n \geqslant 1$). Soit $I := I_1 + \cdots + I_n$ la somme des sous-groupes I_i (théorème et définition 17 de la page 122). Alors I est l'*idéal* de A engendré par la réunion des I_i.

Démonstration. Soit S la réunion des I_i. *A priori*, I est, au sens de l'inclusion, le plus petit *sous-groupe* de $(A, +)$ contenant S. Montrons que I est un *idéal* de A. Soient $x \in I$ et $a \in A$. Il s'agit de prouver que $ax \in I$. D'après le théorème et définition 17 de la page 122, x s'écrit $x = x_1 + \cdots + x_n$, où $x_i \in I_i$ pour $i = 1, \ldots, n$. Pour tout i, $ax_i \in I_i$, car I_i est un idéal de A et $x_i \in I_i$. Ainsi la somme des ax_i appartient à I, par définition de I. Par distributivité, cette somme vaut ax, d'où $ax \in I$.

Si J est un idéal de A contenant S, c'est un sous-groupe de $(A, +)$, il contient I d'après la proposition 15 de la page 122 et le théorème et définition 17 de la page 122. ■

Définition 17. Soient A, A' deux anneaux. Une application $f : A \to A'$ est appelée *morphisme d'anneaux* si elle vérifie les conditions suivantes :

1. Pour tous $x, y \in A$, $f(x + y) = f(x) + f(y)$.
2. Pour tous $x, y \in A$, $f(xy) = f(x)f(y)$.
3. On a : $f(1_A) = 1_{A'}$.

Si de plus f est bijective, on dit que c'est un *isomorphisme d'anneaux* de A sur A' (ou *automorphisme de l'anneau* A lorsque $A' = A$).

La condition 1 signifie que f est un *morphisme de groupes* de $(A, +)$ dans $(A', +)$. La condition 3 ne doit pas être oubliée, elle ne résulte pas de 1 et 2 : *cf.* l'application nulle (si A' n'est pas nul).

Remarquons que le mot morphisme peut maintenant devenir ambigu. Voici une règle de bon sens :

Lorsqu'on parle d'un morphisme $f : A \to A'$, il faut préciser si c'est un morphisme de groupes, d'anneaux, …, sauf lorsque le contexte l'indique.

Proposition 36. Soit $f : A \to A'$ un morphisme d'anneaux. Si x est un élément inversible de A, $f(x)$ est un élément inversible de A' et $f(x)^{-1} = f(x^{-1})$.
Ainsi $x \mapsto f(x)$ est un morphisme de groupes, de $A^\times$ dans $A'^\times$.

Démonstration. Soit $x \in A^\times$. Puisque $xx^{-1} = 1_A$, la définition 17 de la page précédente implique : $1_{A'} = f(xx^{-1}) = f(x)f(x^{-1})$, et de même $1_{A'} = f(x^{-1})f(x)$, d'où la première assertion de l'énoncé. La deuxième assertion est alors évidente. ∎

Théorème 37. Soit $f : A \to A'$ un morphisme d'anneaux.

1. L'image $f(B)$ d'un sous-anneau B de A est un sous-anneau de A'. En particulier, $\mathrm{Im}(f) := f(A)$ est un sous-anneau de A'.
2. Si B' est un sous-anneau de A', $f^{-1}(B')$ est un sous-anneau de A.
3. Supposons A commutatif. Le *noyau* $\mathrm{Ker}(f)$ de f, considéré comme morphisme de $(A, +)$ dans $(A', +)$, est un *idéal* de A.

Démonstration. Dans tous les cas, la partie dont on veut prouver que c'est un sous-anneau ou un idéal est un sous-groupe additif, en vertu du théorème et définition 19 de la page 123. Pour l'assertion 1, $1_A \in B$, donc $1_{A'} = f(1_A) \in f(B)$. Soient $x', y' \in f(B)$. Il existe $x, y \in B$ tels que $x' = f(x)$, $y' = f(y)$. Puisque f est un morphisme, $x'y' = f(x)f(y) = f(xy)$; or $xy \in B$ (B est un sous-anneau de A), donc $x'y' = f(xy) \in f(B)$. Ainsi $f(B)$ est un sous-anneau de A'. L'assertion 2 est analogue.
Pour l'assertion 3, soient $a \in A$ et $x \in \mathrm{Ker}(f)$. Il suffit de voir que $ax \in \mathrm{Ker}(f)$, ce qui est clair : $f(ax) = f(a)f(x) = f(a)\,0 = 0$. ∎

Théorème 38. Soit $f : A \to A'$ un morphisme d'anneaux.

1. Pour que f soit *injectif*, il faut, et il suffit, que son noyau soit $\{0_A\}$.
2. Si f est *bijectif*, f^{-1} *est aussi un morphisme d'anneaux*.

Démonstration. L'assertion 1 résulte du théorème 20 de la page 124, car seule l'addition intervient dans la définition de $\mathrm{Ker}(f)$. Supposons f bijectif. On sait déjà que f^{-1} est un morphisme de groupes additifs. Ensuite $f(1_A) = 1_{A'}$, d'où $f^{-1}(1_{A'}) = 1_A$. En raisonnant comme pour le théorème 20 de la page 124, le lecteur vérifiera que $f^{-1}(x'y') = f^{-1}(x')f^{-1}(y')$ pour tous $x', y' \in A'$. ∎

Exemples.

1. Soient A un anneau et X un ensemble. Considérons l'anneau $\mathcal{F}(X, A)$ (exemple 3 de la page 134). Soit $x \in X$ fixé. Notons $e_x : \mathcal{F}(X, A) \to A$ l'application $f \mapsto f(x)$ (évaluation en x). Alors e_x est un morphisme d'anneaux, son noyau est formé des $f : X \to A$ s'annulant en x.
2. Soient A un anneau et a un élément inversible de A fixé. Alors $x \mapsto axa^{-1}$ est un automorphisme de l'anneau A (le vérifier).

L'étude des morphismes de $\mathbb{Z}$ dans d'autres anneaux est simple :

Théorème et définition 39 (Caractéristique d'un anneau).
Soient A un anneau et 1_A son élément neutre multiplicatif.

1. Il existe un *unique* morphisme d'anneaux ε de $\mathbb{Z}$ dans A.
2. Explicitement, on a $\varepsilon(k) = k \cdot 1_A$ pour tout $k \in \mathbb{Z}$.
3. Il existe un unique entier $n \geqslant 0$, appelé *caractéristique de* A et noté $\mathrm{car}(A)$, tel que $\mathrm{Ker}(\varepsilon) = n\mathbb{Z}$. De plus $nx = 0$ pour tout $x \in A$.

Démonstration. Définissons $\varepsilon : \mathbb{Z} \to A$ par la formule donnée en 2. D'abord $\varepsilon(1) = 1_A$. Soient $h, k \in \mathbb{Z}$. L'égalité $\varepsilon(h+k) = \varepsilon(h) + \varepsilon(k)$ résulte de la première formule (3) de la page 113. Ensuite :

$$\varepsilon(hk) = (hk) \cdot 1_A = h \cdot (k \cdot 1_A) = (h \cdot 1_A)(k \cdot 1_A) = \varepsilon(h)\varepsilon(k);$$

la 2*ème* égalité vient de la 2*ème* formule (3), la 3*ème* égalité vient des formules 4 de la proposition 31 de la page 133. Ainsi ε est un morphisme d'anneaux.

Soit $f : \mathbb{Z} \to A$ un morphisme d'anneaux, montrons que $f = \varepsilon$. D'après l'exercice II.2.26 de la page 158, $\{k \in \mathbb{Z} \mid f(k) = \varepsilon(k)\}$ est un sous-anneau de $\mathbb{Z}$, donc un sous-groupe de $(\mathbb{Z}, +)$ contenant 1. Il est ainsi égal à $\mathbb{Z}$, *i.e.* $f = \varepsilon$.

Puisque $\mathrm{Ker}(\varepsilon)$ est un sous-groupe de $(\mathbb{Z}, +)$, la première partie de l'assertion 3 vient du théorème 50 de la page 95. Si $x \in A$, on a : $nx = (n \cdot 1_A)x = 0_A\, x = 0_A$. ∎

La plupart des anneaux intervenant dans cet ouvrage seront *de caractéristique* 0. Les anneaux $\mathbb{Z}/n\mathbb{Z}$ feront exception (*cf.* la section 3.4).

PROPRIÉTÉ. *Soit A un anneau* intègre *de caractéristique* 0. *Pour tous $h, k \in \mathbb{Z}$ et $x \in A$, on a :*

$$(hx = 0) \Longleftrightarrow \big((h = 0) \text{ ou } (x = 0)\big) \quad \textit{et} \quad (hx = kx) \Longleftrightarrow (h = k) \text{ ou } (x = 0).$$

La première équivalence vient de l'égalité $hx = (h \cdot 1_A)x$, et la deuxième s'en déduit aussitôt, car $hx - kx = (h-k)x$.
Voici deux résultats simples concernant la caractéristique d'un anneau.

Proposition 40.
La caractéristique d'un anneau *intègre* A est 0 ou un nombre premier.

Démonstration. Supposons $n := \mathrm{car}(A) \neq 0$. D'abord $1 \neq 0$ (A est intègre), donc $n \neq 1$. Supposons n non premier : $n = hk$, où $h, k \in \mathbb{N}^*$ sont distincts de 1. Alors $(h \cdot 1_A)(k \cdot 1_A) = (hk) \cdot 1_A = n \cdot 1_A = 0$. L'un des deux éléments $h \cdot 1_A, k \cdot 1_A$ est nul, car A est intègre. Si par exemple $h \cdot 1_A = 0$, h appartient à $\mathrm{Ker}(\varepsilon) = n\mathbb{Z}$. Ainsi n divise h, ce qui est absurde. ∎

Proposition 41. Soit A un anneau commutatif dont la caractéristique est un nombre premier p. L'application $x \mapsto x^p$ est un morphisme d'anneaux de A dans A. En particulier, si $x, y \in A$, on a $(x+y)^p = x^p + y^p$.

Démonstration. Appliquons la formule du binôme (A est commutatif) :

$$(x+y)^p = \sum_{k=0}^{p} \binom{p}{k} x^{p-k} y^k = x^p + y^p + \overbrace{\sum_{k=1}^{p-1} \binom{p}{k} x^{p-k} y^k}^{S}.$$

Soit $k \in [\![1, p-1]\!]$. L'exercice 8 de la page 85 montre que $\binom{p}{k}$ est multiple de p : $\binom{p}{k} = mp$, où $m \in \mathbb{N}$. Pour tout $z \in A$, $pz = 0$, d'où $(mp)z = 0$, soit $\binom{p}{k} z = 0$. Ainsi chaque terme de la somme S est nul, d'où $(x+y)^p = x^p + y^p$.

L'application $F : x \mapsto x^p$ est un morphisme de groupes additifs. Or $F(1) = 1$ et, si $x, y \in A$, $(xy)^p = x^p y^p$. Ainsi F est un morphisme d'anneaux. ∎

3.3 Divisibilité dans un anneau intègre

Tous les anneaux considérés sont supposés intègres (donc commutatifs).

Cette section est un « pont » entre l'étude arithmétique de $\mathbb{Z}$ et celle des anneaux $K[X]$ (module II.6). Les éléments neutres d'un anneau A sont notés 0 et 1 (ainsi $0 \neq 1$), et $A^\times$ est le groupe des inversibles de A.

Proposition 42. Soient A un anneau et $a \in A$. L'idéal (a) de A engendré par a est égal à $aA := \{at \mid t \in A\}$.

Démonstration. D'abord aA est un sous-groupe de $(A, +)$: c'est l'image du morphisme $t \mapsto at$ de $(A, +)$ dans lui-même. Si $x \in aA$ et $b \in A$, il existe $t \in A$ tel que $x = at$, alors $bx = b(at) = a(bt) \in aA$. Ainsi aA est un idéal de A, contenant $a = a \times 1$, d'où $(a) \subset aA$. D'autre part $a \in (a)$ donc, pour tout $t \in A$, $at = ta \in (a)$, car (a) est un idéal de A. D'où $aA \subset (a)$. ∎

Définition 18. Soient A un anneau et $a, b \in A$. On dit que a *divise* b (et l'on note $a \mid b$), ou que b *est multiple de* a, s'il existe un $t \in A$ tel que $b = ta$. On dit que a et b sont *associés* (ce que l'on note $a \sim b$) si $a \mid b$ et $b \mid a$.

Dans cette définition, *si* $a \neq 0$, un $t \in A$ tel que $b = ta$ est *unique* s'il existe, car A est intègre. Noter aussi que tout $a \in A$ divise 0, car $0 = 0 \times a$. Par contre le seul multiple de 0 est 0, et le seul élément associé à 0 est 0.

Proposition 43. Soient A un anneau et $a, b \in A$. Alors :

1. Pour que a divise b, il faut, et il suffit, que aA *contienne* bA.
2. Pour que a et b soient associés, il faut, et il suffit, que aA et bA soient égaux, ou encore qu'il existe un $u \in A^\times$ tel que $b = ua$.

Démonstration. La proposition précédente montre que $bA = (b)$. Puisque aA est un idéal, il contient donc bA si, et seulement si, $b \in aA$, ce qui veut dire, par définition, que $a \mid b$. D'où 1. La première assertion de 2 en résulte.

Si $u \in A^\times$ et $b = ua$, on a $a = u^{-1}b$, donc a et b sont associés. Inversement, si $a \sim b$, distinguons deux cas. Si a est nul, b l'est aussi, et $b = 1 \times a$. Sinon, il existe $u, v \in A$ tels que $b = ua$ et $a = vb$. Alors $a = uva$, donc $uv = 1$ car A est intègre et $a \neq 0$. Ainsi u est inversible. ∎

Dans un anneau A, tout $a \neq 0$ possède des diviseurs évidents : tout élément inversible de A divise a, et tout élément associé à a divise a. De tels diviseurs de a sont dits *triviaux*. Généralisons la notion de nombre premier.

Définition 19. Soit A un anneau. Un élément p de A est dit *irréductible* s'il est non nul, non inversible, et si *ses seuls diviseurs sont ses diviseurs triviaux*. La dernière condition peut être exprimée ainsi : si $p = xy$, où $x, y \in A$, l'un (au moins) des deux éléments x, y est inversible.

Si deux éléments d'un anneau sont associés, l'un est irréductible si, et seulement si, l'autre l'est. Puisque $\mathbb{Z}^\times = \{-1, 1\}$, les diviseurs triviaux d'un entier $a \neq 0$ sont $-1, 1, a$ et $-a$. Il en résulte que *les éléments irréductibles de $\mathbb{Z}$ sont les nombres premiers et leurs opposés*. Dans $K[X]$, anneau des polynômes sur un corps K (*cf.* le module II.6), les inversibles seront les éléments de K^* ; les *polynômes irréductibles* joueront un rôle important.
Terminons cette section en introduisant la notion d'*anneau principal*.

Définition 20. Soit A un anneau commutatif. Un idéal I de A est dit *principal* s'il existe un élément $a \in A$ tel que $I = aA$; un tel a est appelé *générateur* de I. On dit que A est un *anneau principal* s'il est *intègre* et si tout idéal de A est *principal*.

Ainsi $\mathbb{Z}$ est principal (proposition 34 de la page 139), comme le seront les anneaux de polynômes sur un corps aussi (*cf.* le module II.6). Les anneaux principaux ont des propriétés voisines de celles de $\mathbb{Z}$. Par exemple :

Proposition 44. Soient A un anneau principal et $a, b \in A$. Considérons un générateur d de l'idéal $(a) + (b)$. Alors d est un *pgcd* de a et b, au sens suivant : d divise a et b, et tout diviseur commun à a et b divise d.

Démonstration. L'existence de d vient du fait que l'idéal $(a) + (b)$ est principal. Rappelons que $(a) + (b)$ est, au sens de l'inclusion, le plus petit idéal de A contenant $\{a, b\}$. Soit $t \in A$. Pour que t divise a et b, il faut, et il suffit, que (t) contienne (a) et (b), *i.e.* que (t) contienne $(a)+(b) = (d)$, ce qui revient à dire que t divise d. D'où le résultat. ∎

Remarque. La notion de pgcd (comme celle de divisibilité) est définie *à association près* : si d est comme ci-dessus, tout élément associé à d est encore un générateur de $(a)+(b)$, donc est un pgcd de a et b.

Proposition 45 (Lemme d'Euclide). Dans un anneau principal A, si un élément *irréductible* divise le produit de deux éléments, il divise (au moins) l'un de ces deux éléments.

Démonstration. Soient $p, a, b \in A$, supposons que p soit irréductible et divise ab. Soit d un générateur de $(p)+(a)$, *i.e.* un pgcd de p et a ; en particulier $d \mid p$. Or p est irréductible, donc d est un diviseur trivial de p. Ainsi, ou bien $d \sim p$, et dans ce cas $p \mid a$ car $d \mid a$; ou bien d est inversible, donc 1 engendre $(p)+(a)$. Il existe alors $u, v \in A$ tels que $1 = up + va$, d'où $b = p(ub) + v(ab)$. Comme p divise ab et $p(ub)$, il divise b. ■

3.4 Anneau des entiers modulo n

Considérons un entier $n \geqslant 1$ fixé. Nous allons voir que l'ensemble des entiers modulo n a une structure naturelle d'anneau. Dans la section 2.4, nous avons construit le groupe abélien $(\mathbb{Z}/n\mathbb{Z}, +)$ et le morphisme canonique $\pi : k \mapsto \overline{k}$ de $\mathbb{Z}$ sur $\mathbb{Z}/n\mathbb{Z}$. Définissons maintenant une multiplication sur $\mathbb{Z}/n\mathbb{Z}$, puis vérifions les propriétés de la définition 11 de la page 132. Partons d'un lemme :

Lemme 46. Soit $n \geqslant 1$ un entier. Soient $x, x', y, y' \in \mathbb{Z}$ tels que $x \equiv x' \mod n$ et $y \equiv y' \mod n$. Alors $xy \equiv x'y' \mod n$.

Démonstration. Elle est immédiate : l'entier n divise $x - x'$ et $y - y'$, il divise donc également $xy - x'y' = (x - x')y + x'(y - y')$. ■

Une conséquence de ce lemme est que, si $a, b \in \mathbb{Z}$, l'élément $\overline{ab} \in \mathbb{Z}/n\mathbb{Z}$ ne dépend que de $\overline{a}, \overline{b} \in \mathbb{Z}/n\mathbb{Z}$, et pas de a et b eux-mêmes. Il est donc licite de définir une multiplication sur $\mathbb{Z}/n\mathbb{Z}$ en posant, pour tous $x, y \in \mathbb{Z}$:

$$\overline{x} \times \overline{y} := \overline{xy}. \tag{32}$$

Par exemple $3 \times 4 = 12$ et $12 \bmod 7 = 5$, donc $\overline{3} \times \overline{4} = \overline{5}$ dans $\mathbb{Z}/7\mathbb{Z}$.

Proposition 47. Si $n \geqslant 1$ est un entier, $(\mathbb{Z}/n\mathbb{Z}, +, \times)$ est un anneau commutatif. De plus le morphisme canonique $\pi : k \mapsto \overline{k}$ de $\mathbb{Z}$ dans $\mathbb{Z}/n\mathbb{Z}$ est un morphisme d'*anneaux* surjectif, de noyau $n\mathbb{Z}$.

Démonstration. C'est une conséquence immédiate des formules (24) et (32), vu les propriétés déjà connues de π comme morphisme de groupes. ■

Exercice 8.
Calculer le reste de la division de 3^{1000} par 17.

Solution. Observons que $3^4 = 81 = 5 \times 17 - 4$. On en déduit, mod. 17, les congruences $3^4 \equiv -4$, puis $3^8 \equiv 16 \equiv -1$, et enfin $3^{1000} = (3^8)^{125} \equiv (-1)^{125} = -1$.
Ainsi $3^{1000} \equiv -1 \equiv 16 \mod 17$, *i.e.* le reste cherché est 16.

Les congruences ci-après sont mod. 9. On a $10 \equiv 1$, d'où $10^k \equiv 1$ pour tout $k \in \mathbb{N}$. Soit $x \in \mathbb{N}$ un entier écrit en base 10 : $x = a_r a_{r-1} \cdots a_1 a_0$, soit $x = \sum_{k=0}^{r} a_k 10^k$. Notant $s(x)$ la somme des chiffres $a_0, \ldots, a_r$ de x, on a $x \equiv s(x)$. Par exemple la multiplication $1\,263 \times 551 = 696\,913$ est *fausse* :

$$1\,263 \equiv s(1\,263) = 12 \equiv 3\,,\ 551 \equiv s(551) = 11 \equiv 2\,,$$

alors que $696\,913 \equiv s(696\,913) = 34 \equiv 7$. C'est là une « preuve par 9 ». En fait $1\,263 \times 551 = 695\,913$, et $695\,913 \equiv s(695\,913) = 33 \equiv 6$. La preuve par 9 ne garantit pas que la multiplication soit juste, elle donne seulement la congruence $1\,263 \times 551 \equiv 695\,913$; il y a donc « une chance sur 9 » pour que la multiplication soit fausse malgré tout.

Montrons que le nombre de Fermat $F_5 := 2^{2^5} + 1$ (exercice 3.2 de la page 85) n'est pas premier. D'abord $F_5 = 2^{32} + 1$. Un ordinateur factorise immédiatement F_5 comme produit de deux nombres premiers :

$$F_5 \ = \ 4\,294\,967\,297 \ = \ 641 \times 6\,700\,417.$$

Voici la preuve astucieuse (et sans ordinateur !) d'Euler du fait que F_5 ne soit pas premier ; elle illustre la puissance des congruences. Les congruences ci-après sont modulo 641. Euler part des égalités :

$$\begin{cases} 641 = 625 + 16 = 5^4 + 2^4 \\ 641 = 640 + 1 = 5 \times 128 + 1 = 5 \times 2^7 + 1. \end{cases}$$

Les premières égalités donnent : $5^4 \equiv -2^4$. Les dernières donnent : $5 \times 2^7 \equiv -1$. Élevons à la puissance 4 : $5^4 \times 2^{28} \equiv 1$, d'où $(-2^4) \times 2^{28} \equiv 1$. Ainsi $2^{32} + 1 \equiv 0$, ce qui signifie que F_5 est multiple de 641.

Proposition 48. Soit $n \geqslant 2$ un entier.
L'anneau $\mathbb{Z}/n\mathbb{Z}$ est intègre si, et seulement si, n est premier.

Démonstration. Supposons d'abord n premier. L'anneau $\mathbb{Z}/n\mathbb{Z}$ n'est pas nul. Soient $x, y \in \mathbb{Z}/n\mathbb{Z}$ tels que $xy = 0$. Il existe des entiers a, b tels que $x = \overline{a}$, $y = \overline{b}$. Alors $\overline{ab} = xy = 0$, donc ab est multiple de n $\big($*cf.* les équivalences (23)$\big)$. Puisque n est premier, le lemme d'Euclide montre que n divise a ou b, *i.e.* $x = 0$ ou $y = 0$. L'anneau $\mathbb{Z}/n\mathbb{Z}$ est donc intègre.

Supposons ensuite n non premier : $n = ab$, où $a, b \in [\![2, n-1]\!]$. Alors $\overline{a} \neq 0$, $\overline{b} \neq 0$, mais $\overline{a}\,\overline{b} = \overline{n} = 0$, donc $\mathbb{Z}/n\mathbb{Z}$ n'est pas intègre. ∎

Rappelons la définition suivante (*cf.* l'exercice 9 de la page 89) :

Définition 21.
Pour tout entier $n \geqslant 1$, notons $\varphi(n)$ le nombre d'entiers $k \in [\![1, n]\!]$ premiers avec n.
La fonction $\varphi : \mathbb{N}^* \to \mathbb{N}$ ainsi définie est appelée *fonction indicatrice d'Euler*.

Si p est premier, $\varphi(p) = p - 1$, car tout $k \in [\![1, p-1]\!]$ est premier avec p.

Proposition 49. Soit $n \geqslant 2$ un entier.

1. Soit $k \in \mathbb{Z}$. Pour que $\overline{k} \in \mathbb{Z}/n\mathbb{Z}$ soit inversible, il faut, et il suffit, que k soit *premier avec* n, *i.e.* que $\overline{k}$ soit un générateur de $(\mathbb{Z}/n\mathbb{Z}, +)$.
2. Le groupe multiplicatif $(\mathbb{Z}/n\mathbb{Z})^\times$ des inversibles de l'anneau $\mathbb{Z}/n\mathbb{Z}$ est d'ordre $\varphi(n)$.

Démonstration. Soit $k \in \mathbb{Z}$. Supposons que $\overline{k}$ possède un inverse. Il existe donc $h \in \mathbb{Z}$ tel que $\overline{h}\,\overline{k} = \overline{1}$, soit $hk \equiv 1 \mod n$. D'où un $t \in \mathbb{Z}$ tel que $hk - 1 = nt$. Ainsi k et n sont premiers entre eux. Inversement, si k est premier avec n, il existe $u, v \in \mathbb{Z}$ tels que $ku + nv = 1$ (théorème de Bézout). Alors $ku \equiv 1 \mod n$, soit $\overline{k}\,\overline{u} = 1$, et donc $\overline{k}$ est inversible. D'où l'assertion 1, vu l'exercice 5 de la page 132.

L'assertion 2 est une conséquence évidente de 1 et de la définition de φ, car $k \mapsto \overline{k}$ est une bijection de $[\![1, n]\!]$ sur $\mathbb{Z}/n\mathbb{Z}$. ■

Remarque. Si $k \in \mathbb{Z}$ est premier avec n, cette preuve *donne un moyen* de calculer $\overline{k}^{-1} \in \mathbb{Z}/n\mathbb{Z}$: on détermine une relation de Bézout entre k et n : $ku + nv = 1$ (*cf.* l'exercice 12 de la page 98), alors $\overline{k}^{-1} = \overline{u}$.

Donnons deux conséquences importantes de la proposition précédente.

Théorème 50 (Euler, 1707-1783). Soit $n \geqslant 2$ un entier. Pour tout entier $a \in \mathbb{Z}$ premier avec n, on a :

$$a^{\varphi(n)} \equiv 1 \mod n. \tag{33}$$

Démonstration. Soient $a \in \mathbb{Z}$ un entier premier avec n et $\overline{a}$ son image dans $\mathbb{Z}/n\mathbb{Z}$. Vu la proposition précédente, $\overline{a} \in (\mathbb{Z}/n\mathbb{Z})^\times$, et le groupe multiplicatif $(\mathbb{Z}/n\mathbb{Z})^\times$ est d'ordre $\varphi(n)$. Le théorème de Lagrange (théorème 22 de la page 125) montre donc que $\overline{a}^{\varphi(n)} = 1$. D'où la congruence annoncée. ■

Théorème 51 (Petit théorème de Fermat).
Soit p un nombre premier. Si $a \in \mathbb{Z}$ n'est pas multiple de p, on a :

$$a^{p-1} \equiv 1 \mod p, \quad \text{donc} \quad a^p \equiv a \mod p. \tag{34}$$

La deuxième de ces congruences est vraie en fait pour *tout* entier a.

Démonstration. Elle résulte du théorème d'Euler. Si $a \in \mathbb{Z}$ n'est pas multiple de p, il est premier avec p (car p est premier). Puisque $\varphi(p) = p - 1$, la première congruence (34) n'est autre que (33) (en prenant $n := p$). Lorsque p ne divise pas a, la deuxième congruence (34) résulte de la première en multipliant par a ; si p divise a, cette congruence est évidente. ■

Remarques.

1. Lorsqu'on effectue de nombreux calculs (additions, multiplications) sur des entiers, il peut y avoir « explosion » : les résultats intermédiaires peuvent dépasser la taille permise (sur ordinateur). *Il n'y a pas d'explosion* lorsqu'on calcule dans $\mathbb{Z}/n\mathbb{Z}$: les résultats ne dépassent jamais la taille de n. Cette remarque a beaucoup d'applications pratiques.
2. Le théorème ci-dessus est dit « petit » par opposition au dernier (ou grand) théorème de Fermat » (*cf.* la remarque de la page 138). Ce théorème a une conséquence apparemment surprenante. Soit $n > 1$ un entier impair. Si 2^{n-1} n'est pas congru à 1 modulo n, ce théorème dit que n *n'est pas premier*. Si n est « grand », par exemple s'il a $100\,000$ chiffres, le calcul de $\overline{2}^{n-1} \in \mathbb{Z}/n\mathbb{Z}$ par ordinateur ne pose pas de problème ; vu l'exemple de la page 82, ce calcul nécessite moins de $700\,000$ multiplications. Bien entendu, tous les calculs se font dans $\mathbb{Z}/n\mathbb{Z}$, par application itérée de la formule (32), et non dans $\mathbb{N}$: même pour un ordinateur, le calcul de 2^{n-1} n'est pas faisable !

   Si donc $\overline{2}^{n-1} \neq 1$, nous pouvons affirmer que n n'est pas premier, *sans pour autant connaître un diviseur non trivial de* n.

Voici une application du théorème de Fermat. Il s'agit d'une méthode de codage très utilisée en cryptographie, la méthode RSA (due à Rivest, Shamir et Adleman). Soient p et q deux nombres *premiers* distincts, $N = pq$ et $c \geqslant 1$ un entier fixé *premier* avec $(p-1)(q-1)$. L'expéditeur du message ne connaît que N et c : la *« clé publique »*, mais ni p ni q. Le message à transmettre est initialement constitué de *lettres* ou, plus généralement, de *caractères*. On le transforme en une *liste de chiffres* en remplaçant, dans le premier cas, chaque lettre par son rang dans l'alphabet et, dans le second, chaque caractère par le nombre qui le représente dans un certain *encodage*, par exemple en utilisant le codage ASCII (version étendue 8 bits) par un entier entre 0 et $255 = 2^8 - 1$. On découpe ensuite le message ainsi transformé en *tranches* de même longueur représentant chacune un nombre entier $x < N$, puis l'on remplace chaque tranche par le $\bar{x} \in \mathbb{Z}/N\mathbb{Z}$ correspondant. On *code* chaque $\bar{x}$ en le remplaçant par $\bar{y} = \bar{x}^c$ et l'on transmet la liste des $y \in [\![0, N-1]\!]$ correspondants.
Comment le destinataire du message peut-il *décoder*, *i.e.* retrouver $\bar{x}$ à partir de $\bar{y}$ *en connaissant* p *et* q ? Voici la réponse. D'après le théorème de Bézout, il existe deux entiers d et k tels que $cd - k(p-1)(q-1) = 1$; on peut de plus supposer $k \geqslant 1$ (*cf.* la proposition 57 de la page 98), d'où $d \geqslant 1$. Alors $\bar{x} = \bar{y}^d$. En effet, il s'agit de vérifier que $\bar{x}^{cd} = \bar{x}$, soit $x^{cd} \equiv x \mod pq$, ou encore que pq divise $x^{cd} - x$. Comme pq est le ppcm de p et q, il suffit de voir que $x^{cd} - x$ est multiple de p et de q. Par symétrie, il suffit de montrer que $x^{cd} - x \equiv 0 \mod p$. C'est clair si p divise x, auquel cas p divise x^{cd}. Si p ne divise pas x, le théorème de Fermat donne $x^{p-1} \equiv 1 \mod p$. Alors :

$$x^{cd} = x^{1+k(p-1)(q-1)} = x\left[x^{p-1}\right]^{k(q-1)} \equiv x \mod p.$$

Ainsi, pour récupérer $\bar{x}$, il suffit d'élever $\bar{y}$ à la puissance d.
En pratique, on choisit pour p et q deux « grands » nombres premiers distincts (de quelques centaines de chiffres chacun). Quel est l'intérêt ?

Le point capital est que, depuis un peu plus d'une vingtaine d'années, grâce à des méthodes que nous ne pouvons pas détailler ici, on sait fabriquer à la demande des nombres premiers p et q de la taille requise. En revanche, il se trouve que, *dans l'état actuel*, on ne peut pas factoriser un entier de la taille de $N = pq$.

Les entiers N et c constituent la « clé publique », ils peuvent sans inconvénient être connus de tous, et permettent de coder les messages. Pour décoder, il faut avoir d. Or, pour avoir d, il faut une relation de Bézout entre c et $(p-1)(q-1)$; il faut donc connaître $(p-1)(q-1)$, *i.e.*, puisque pq est connu, il faut connaître p et q (penser à une équation du second degré). L'entier d, appelé « clé secrète », ne peut donc pas être trouvé, même si N et c sont connus, ainsi que le message initial $\bar{x}$ et le message codé $\bar{y}$. C'est l'avantage essentiel de cette méthode de codage, dite *à clé publique*.

Voici un exemple numérique « petit » : $N := 415\,439 = pq$, où $p := 233$ et $q := 1\,783$. Alors $(p-1)(q-1) = 232 \times 1\,782 = 2^4.3^4.11.29$. Choisissons $c := 113$ et $d := 336\,593$; il est facile de vérifier que $cd \equiv 1 \mod (p-1)(q-1)$. Le mot « eau » est codé par $\bar{x} := \overline{5\,01\,21}$ (les numéros des lettres e, a et u sont 5, 1 et 21 respectivement). On vérifie que $\bar{x}^c = \overline{22\,09\,14}$. Les 22ème, 9ème et 14ème lettres de l'alphabet étant v, i et n, le mot codé est : « vin ».

Théorème 52 (Théorème Chinois). Soient $m_1, \ldots, m_r \geqslant 2$ des entiers premiers entre eux *deux à deux* ($r \geqslant 2$) et M leur produit.

Étant donnés des entiers $a_1, \ldots, a_r$, considérons le système de congruences :

$$(S) : \ x \equiv a_i \mod m_i \ (i = 1, \ldots, r). \tag{35}$$

Ce système possède une solution $x \in \mathbb{Z}$, qui est unique modulo M.

Démonstration. Commençons par l'unicité. Soient $x, y \in \mathbb{Z}$ deux solutions de (S). Pour tout i, $x \equiv a_i \equiv y \mod m_i$, donc $x - y$ est multiple de m_i. Ainsi $x - y$ est multiple du ppcm des a_i, qui vaut M puisque les m_i sont premiers entre eux deux à deux (théorème 44 de la page 91). D'où $x \equiv y \mod M$.

Pour l'existence, supposons que $r = 2$. Puisque $m_1 \wedge m_2 = 1$, il existe $u, v \in \mathbb{Z}$ tels que $um_1 + vm_2 = 1$. Posons $x_1 := vm_2$ et $x_2 := um_1$. Alors x_1 est congru à $0 \mod m_2$ et à $1 \mod m_1$, car $x_1 = 1 - um_1$. De même $x_2 \equiv 1 \mod m_2$ et $x_2 \equiv 0 \mod m_1$. Posons $x := a_1x_1 + a_2x_2$. Alors :

$$x \equiv a_1 \times 1 + a_2 \times 0 = a_1 \mod m_1 \qquad \text{et} \qquad x \equiv a_1 \times 0 + a_2 \times 1 = a_2 \mod m_2,$$

donc x est solution de (S).

Raisonnons alors par récurrence sur r. Si $r \geqslant 3$, le passage de $r-1$ à r se fait comme suit. Soit $P := m_1m_2 \cdots m_{r-1}$. D'après l'hypothèse de récurrence, il existe un $y \in \mathbb{Z}$ tel que $y \equiv a_i \mod m_i$ pour $i = 1, \ldots, r-1$. Par ailleurs $P \wedge m_r = 1$. D'après le cas $r = 2$, il existe un $x \in \mathbb{Z}$ tel que $x \equiv y \mod P$ et $x \equiv a_r \mod m_r$. Il est clair que x est solution de (S). ∎

Noter que la preuve précédente fournit une méthode pratique de résolution du système (S).

Exercice 9.
Résoudre par exemple le système suivant :

$$(S) : x \equiv 5 \mod 17\,,\ x \equiv 3 \mod 23.$$

Solution. Les nombres 17 et 23 étant premiers entre eux, il existe $u, v \in \mathbb{Z}$ tels que $17u + 23v = 1$. Par exemple $(u, v) := (-4, 3)$ convient. Avec les notations de la preuve ci-dessus, $x_1 := 3 \times 23 = 69$ et $x_2 := 17 \times (-4) = -68$. D'où une solution x de (S), à savoir $x := 5x_1 + 3x_2 = 5 \times 69 + 3 \times (-68) = 141$. La solution générale de (S) est donc $x := 141 + 391k$, où $k \in \mathbb{Z}$ est arbitraire.

Voici une autre version du théorème chinois.

Théorème 53. Soient $m_1, \ldots, m_r \geqslant 2$ des entiers premiers entre eux *deux à deux* $(r \geqslant 2)$ et M leur produit.

1. Il existe un isomorphisme d'anneaux « naturel » de $\mathbb{Z}/M\mathbb{Z}$ sur l'anneau produit $(\mathbb{Z}/m_1\mathbb{Z}) \times \cdots \times (\mathbb{Z}/m_r\mathbb{Z})$.
2. Cet isomorphisme induit un isomorphisme de groupes multiplicatifs de $(\mathbb{Z}/M\mathbb{Z})^\times$ sur le groupe $(\mathbb{Z}/m_1\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/m_r\mathbb{Z})^\times$. Ainsi :
$$\varphi(M) \;=\; \varphi(m_1)\varphi(m_2)\cdots\varphi(m_r)\,. \tag{36}$$

Démonstration. Soit B l'anneau produit $(\mathbb{Z}/m_1\mathbb{Z}) \times \cdots \times (\mathbb{Z}/m_r\mathbb{Z})$.
Notons $\pi : \mathbb{Z} \to \mathbb{Z}/M\mathbb{Z}$ et $\pi_i : \mathbb{Z} \to \mathbb{Z}/m_i\mathbb{Z}$ $(i = 1, \ldots, r)$ les morphismes canoniques. Soient $k \in \mathbb{Z}$ et $\overline{k} = \pi(k) \in \mathbb{Z}/M\mathbb{Z}$. Nous avons envie de poser :

$$F(\overline{k}) \;:=\; \big(\pi_1(k), \ldots, \pi_r(k)\big) \in B. \tag{$*$}$$

Pour justifier cette formule, qui définit une application $F : \mathbb{Z}/M\mathbb{Z} \to B$, il suffit de voir que le second membre ne change pas si l'on remplace k par un entier l, $l \equiv k \mod M$. C'est évident : pour tout i, $k \equiv l \mod m_i$, soit $\pi_i(k) = \pi_i(l)$, car $k - l$ est multiple de M, *a fortiori* de m_i.

Le fait que F soit un morphisme d'anneaux résulte des formules (24) et (32). Montrons que F est injectif, *i.e.* que $\mathrm{Ker}(F) = (0)$. Soit $x \in \mathrm{Ker}(F)$, $x = \overline{k}$, pour un certain $k \in \mathbb{Z}$. Pour tout i, $\pi_i(k) = 0$, *i.e.* k est multiple de m_i. Ainsi k est multiple de M, ppcm des m_i, d'où $x = \overline{k} = 0$.

Il se trouve que $\mathbb{Z}/M\mathbb{Z}$ et B ont même ordre, à savoir $M = m_1 m_2 \cdots m_r$. Puisque $F : \mathbb{Z}/M\mathbb{Z} \to B$ est injectif, il est bijectif (théorème 15 de la page 71), c'est donc un isomorphisme d'anneaux. En fait, le lecteur pourra vérifier que la surjectivité de F est équivalente au théorème chinois.

La proposition 36 de la page 141 entraîne que F applique $(\mathbb{Z}/n\mathbb{Z})^\times$ sur $B^\times$. Or $B^\times$ est le groupe produit des $(\mathbb{Z}/n_i\mathbb{Z})^\times$ (*cf.* l'exercice II.2.21 de la page 157), d'où la première assertion en 2. La seconde assertion en résulte, en appliquant la proposition 49 de la page 147, puis en prenant les ordres des deux groupes considérés. ∎

4 Corps

Les corps sont des anneaux particuliers, dans lesquels les divisions (par des éléments non nuls) sont possibles. Comme nous l'avons dit dans l'introduction, les corps les plus utiles dans cet ouvrage seront $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, les autres corps n'intervenant que de façon ponctuelle. Il n'y a donc pas lieu d'étudier les corps de manière aussi détaillée que les groupes ou les anneaux.

Définition 22. Un *corps* est un anneau non nul ($0 \neq 1$) dont tout élément *non nul* est *inversible*. Un corps est dit *commutatif* si sa multiplication est commutative.

Le seul corps dont nous disposions pour l'instant est $\mathbb{Q}$, les corps $\mathbb{R}$ et $\mathbb{C}$ seront définis et étudiés dans des modules ultérieurs. Tous les corps envisagés dans cet ouvrage seront commutatifs (hormis le corps des *quaternions*, qui n'apparaîtra que comme exemple). Nous ferons donc la convention suivante :

Désormais, le mot corps signifiera corps commutatif.

Soit K un corps. Le neutre additif de K est noté 0 (ou 0_K si besoin est), et le neutre multiplicatif de K est noté 1 (ou 1_K). Tout $x \in K$ possède un *opposé* $-x$. Le groupe $(K^\times, \times)$ des inversibles de l'anneau K est égal à $K^* = K \setminus \{0\}$: tout $x \in K^*$ possède un *inverse* x^{-1}, noté aussi $1/x$. Si $a \in K$ et $b \in K^*$, ab^{-1} est aussi noté $\dfrac{a}{b}$ et appelé *quotient de a par b*.
Pour un corps K, les groupes $(K, +)$ et $(K^*, \times)$ sont assez différents : ainsi, le lecteur vérifiera que, dans $(\mathbb{Q}, +)$, le seul élément d'ordre fini est 0 alors que, dans $(\mathbb{Q}^*, \times)$, il y a deux éléments d'ordre fini : -1 et 1. Dans $(\mathbb{C}^*, \times)$, il y a une infinité d'éléments d'ordre fini (les « racines de l'unité », voir le module II.5).
Résumons les règles de calcul dans un corps K :

PROPRIÉTÉS.

1. *Associativité de $+$: pour tous $a, b, c \in K$, $(a + b) + c = a + (b + c)$.*
2. *Commutativité de $+$: pour tous $a, b \in K$, $a + b = b + a$.*
3. *Pour tout $a \in K$, $a + 0 = a$ et $a + (-a) = 0$.*
4. *Associativité de $\times$: pour tous $a, b, c \in K$, $(ab)c = a(bc)$.*
5. *Commutativité de $\times$: pour tous $a, b \in K$, $ab = ba$.*
6. *Pour tout $a \in K$, $a \times 1 = a$.*
7. *Pour tout $a \in K^*$, $a \times \dfrac{1}{a} = 1$.*
8. *Pour tout $a \in K$, $0 \times a = 0$.*
9. *Distributivité : pour tous $a, b, c \in K$, $a(b + c) = ab + ac$.*
10. *Pour tous $a, b \in K$, $(-a)b = -(ab) = a(-b)$.*
11. *Régularité de $+$: pour tous $a, b, c \in K$, $(a + b = a + c) \Longrightarrow (b = c)$.*

12. *Pour tous* $a, b \in K$, *la différence* $a - b$ *est définie, elle vaut* $a + (-b)$.
13. *Intégrité : pour tous* $a, b \in K$, $(ab = 0) \Longrightarrow \big((a = 0) \text{ ou } (b = 0)\big)$.
14. *Régularité de* $\times$ *: pour tous* $a \in K^*$ *et* $b, c \in K$, $(ab = ac) \Longrightarrow (b = c)$.
15. *Pour tous* $a \in K$ *et* $b \in K^*$, *le quotient* $\dfrac{a}{b}$ *est défini, il vaut* ab^{-1}.

Vérifions 13 : soient $a, b \in K$ tels que $ab = 0$. Si $a \neq 0$, a est inversible (définition d'un corps), et $ab = 0 = a0$, d'où $b = 0$ (proposition 33 de la page 138).
Si $x \in K$, on sait définir les multiples nx, $n \in \mathbb{Z}$ de x et, si $x \neq 0$, les puissances x^n, $n \in \mathbb{Z}$ de x (revoir les règles (16) à (19) de la page 120).

Proposition 54. Soit K un corps.

1. L'anneau K est *intègre*, de caractéristique 0 ou un nombre premier.
2. L'anneau K possède exactement deux idéaux : $\{0\}$ et K.
3. Tout morphisme f de K dans un anneau *non nul* B est injectif.

Démonstration. D'abord $K \neq \{0\}$, soit $0_K \neq 1_K$, par définition d'un corps, donc la propriété 13 montre que K est intègre. La proposition 40 de la page 142 montre alors que la caractéristique de K est 0 ou un nombre premier.
Soit I un idéal de K. Supposons $I \neq \{0\}$, et montrons que $I = K$. Soient a un élément non nul de I et a^{-1} son inverse. Puisque $a \in I$, $1 = a^{-1}a \in I$ (définition d'un idéal), et donc $I = A$, vu la propriété de la page 139.
Pour l'assertion 3, $\mathrm{Ker}(f)$ est un idéal de K (assertion 3 du théorème 37 de la page 141). Cet idéal n'est pas égal à K, *i.e.* ne contient pas 1, car $f(1) = 1_B \neq 0_B$ par hypothèse. D'après 2, $\mathrm{Ker}(f) = \{0\}$, donc f est injectif (théorème 38 de la page 141). ■

Remarque. Soient K un corps et $x \in K$, $n \in \mathbb{Z}$. Si la caractéristique de K est 0, $(nx = 0)$ équivaut à $\big((n = 0) \text{ ou } (x = 0)\big)$. Si la caractéristique de K est un nombre premier p, $(nx = 0)$ équivaut à $\big((p \mid n) \text{ ou } (x = 0)\big)$.

Théorème 55. Si p est un nombre premier, l'anneau $\mathbb{Z}/p\mathbb{Z}$ est un corps à p éléments.

Démonstration. D'abord $\bar{0} \neq \bar{1}$ dans $\mathbb{Z}/p\mathbb{Z}$, car $1 \not\equiv 0 \mod p$. Soit $x \in \mathbb{Z}/p\mathbb{Z}$, $x \neq 0$, il s'agit de voir que x est inversible. Soit $a \in \mathbb{Z}$ tel que $x = \overline{a}$. Par hypothèse $\overline{a} \neq 0$, *i.e.* p ne divise pas a. Puisque p est premier, $a \wedge p = 1$, et la proposition 49 de la page 147 montre alors que $x = \overline{a}$ est inversible. ■

Exercice 10.
Considérons le nombre premier 37 ? Calculer l'inverse de $x := \overline{7} \in \mathbb{Z}/37\mathbb{Z}$.
Solution. On cherche des entiers u, v tels que $7u + 37v = 1$, alors $x^{-1} = \overline{u}$.
Le couple $(u, v) := (16, -3)$ répond à la question, donc $x^{-1} = \overline{16}$.
Effectivement, $7 \times 16 = 112 = 1 + 3 \times 37 \equiv 1 \mod 37$.

Pour tout nombre premier p, $\mathbb{Z}/p\mathbb{Z}$ est donc un corps *fini*, de cardinal p. Il existe des corps finis plus généraux, dont le cardinal est une *puissance* d'un nombre premier (*cf.* par exemple l'exercice II.2.35). Les corps finis, introduits par Galois, sont importants de nos jours, car ils permettent de fabriquer des « codes correcteurs », servant à la transmission des données.

Exemple. Soit K un corps. L'équation $x^2 = 1$ est simple à résoudre dans K. Elle a deux solutions évidentes : $\pm 1_K$, distinctes si, et seulement si, la caractéristique de K est distincte de 2. Soit $x \in K$ tel que $x^2 = 1$. Alors $0 = x^2 - 1_K = (x - 1_K)(x + 1_K)$; or K est intègre, donc l'un des éléments $x - 1_K$, $x + 1_K$ est nul, *i.e.* $x = \pm 1_K$.
Soit ensuite $K = \mathbb{Z}/p\mathbb{Z}$, où $p \geqslant 3$ est un nombre premier. Si $x \in K^*$, les conditions $(x = 1/x)$, $(x^2 = 1)$ et $\big(x \in \{-1_K, 1_K\}\big)$ sont équivalentes. Posant $m := (p-3)/2$, on peut écrire ainsi les $2m$ éléments de $K^* \setminus \{-1, 1\}$:

$$K^* \setminus \{-1, 1\} \;=\; \big\{x_1, 1/x_1,\, x_2, 1/x_2, \ldots, x_m, 1/x_m\big\}\,,$$

où $x_j \neq x_i$, $1/x_i$ lorsque $j \neq i$. Sous cette forme, il est clair que le produit des éléments de $K^* \setminus \{-1, 1\}$ vaut 1, donc le produit des éléments de K^* vaut -1. Or, les éléments de K^* étant les $\overline{k}$, $k = 1, \ldots, p-1$, leur produit est $\overline{(p-1)!}$. D'où la congruence suivante :

$$(p-1)! \;\equiv\; -1 \mod p \quad (\textit{théorème de Wilson})\,, \tag{37}$$

aussi valable lorsque $p = 2$. Cette congruence *caractérise* les nombres premiers : si $p > 2$ n'est pas premier, il a un diviseur $q \in [\![2, p-1]\!]$; alors q divise $(p-1)!$, donc $(p-1)!$ n'est pas congru à -1 modulo q, *a fortiori* modulo p.

Soit A un anneau (commutatif) *intègre*. De même que le corps $\mathbb{Q}$ a été construit à partir de l'anneau $\mathbb{Z}$, on peut construire, à partir de A un certain corps K, appelé *corps des fractions de* A. Cette construction étant identique à celle de $\mathbb{Q}$ (*cf.* le module II.1), à condition de remplacer les éléments de $\mathbb{Z}$ par ceux de A, nous nous contenterons d'énoncer un théorème :

Théorème 56.
Soit A un anneau intègre. Il existe un corps K vérifiant les conditions :

a) K contient A, plus précisément A est un sous-anneau de K.

b) Tout $x \in K$ s'écrit $x = \dfrac{a}{b}$, où $a \in A$, $b \in A^*$, $\dfrac{a}{b}$ étant le quotient de a par b *dans* K.

Dans le module II.6, le « corps des fractions rationnelles » sera ainsi construit à partir de l'anneau des polynômes. Le corps K construit ci-dessus est unique au sens suivant (*cf.* l'exercice II.2.38) : si K' est un autre corps vérifiant les conditions a) et b) du théorème, il existe un unique isomorphisme d'anneaux f de K sur K' tel que $f(a) = a$ pour tout $a \in A$. Ainsi, pour travailler dans K, peu importe la façon dont on a construit K. L'essentiel est que les conditions a) et b) du théorème soient vérifiées. Le corps K est appelé *corps des fractions* de A.

Définition 23. Soient K un corps et K' une partie de K. On dit que K' est un *sous-corps* de K si c'est un sous-anneau de K et si, pour tout élément non nul x de K', l'inverse de x dans K appartient à K'.

Cela signifie (le vérifier) que K' est un sous-anneau de K et que, muni de l'addition et de la multiplication induites par celles de K, K' est un corps. Il s'avérera ainsi que $\mathbb{Q}$ est un sous-corps de $\mathbb{R}$, et $\mathbb{R}$ un sous-corps de $\mathbb{C}$.

Exemples.

1. Si K, L sont deux corps et $f, g : K \to L$ deux morphismes d'anneaux, $K' := \{x \in K \mid f(x) = g(x)\}$ est un sous-corps de K (le vérifier). Si $L = K$, l'ensemble des points fixes de f est donc un sous-corps de K.
2. Toute intersection de sous-corps d'un corps K est un sous-corps de K. L'intersection k de *tous* les sous-corps de K est donc, au sens de l'inclusion, le plus petit sous-corps de K.

Exercice 11.
Soit K l'un des corps $\mathbb{Q}$ ou $\mathbb{Z}/p\mathbb{Z}$, où p est premier.

1. Montrer que le seul sous-corps de K est K lui-même.
2. Montrer que Id_K est le seul morphisme d'anneaux de K dans K.

Solution.

1. Soit K' un sous-corps de K. Comme sous-anneau de K, il contient 1_K, donc aussi $k \cdot 1_K$ pour tout $k \in \mathbb{N}^*$ (récurrence sur k). Si $K = \mathbb{Z}/p\mathbb{Z} = \{\overline{0}, \overline{1}, \ldots, \overline{p-1}\}$, il en résulte que $K' = K$. Supposons que $K = \mathbb{Q}$. Pour tout entier $k \geqslant 1$, K' contient $1/k$, donc aussi tout rationnel a/b ($a \in \mathbb{Z}$, $b \in \mathbb{N}^*$), car K' est multiplicativement stable. Ainsi $K = \mathbb{Q}$.
2. Soit $f : K \to K$ un morphisme d'anneaux. D'après l'exemple 1 ci-dessus, $K' := \{x \in K \mid f(x) = x\}$ est un sous-corps de K ; ainsi $K' = K$ (d'après l'assertion 1), *i.e.* $f = \mathrm{Id}_K$.

Exercice type corrigé

II.2.0 Soient $(G, \times)$ un groupe *fini* et $s : G \to G$ un morphisme *involutif* : $s \circ s = \mathrm{Id}_G$. On suppose de plus que 1 est le seul *point fixe* de s.

1. Montrer que l'application $x \mapsto x^{-1}s(x)$ de G dans G est injective. En déduire que $s(x) = x^{-1}$ pour tout $x \in G$, puis que G est abélien.
2. Montrer que l'ordre de G est impair.
3. Établir une réciproque lorsque G est abélien d'ordre impair.

Solution.
1. L'application $f : x \mapsto x^{-1}s(x)$ de G dans G est injective.
Soient en effet $x, y \in G$ tels que $f(x) = f(y)$, montrons que $x = y$. Multiplions les deux membres de l'égalité $x^{-1}s(x) = y^{-1}s(y)$ à gauche par x et à droite par $s(y)^{-1} = s(y^{-1})$:

$$x\big[x^{-1}s(x)\big]s(y^{-1}) = x\big[y^{-1}s(y)\big]s(y^{-1}).$$

Comme $s(x)s(y^{-1}) = s(xy^{-1})$, l'égalité obtenue s'écrit : $s(xy^{-1}) = xy^{-1}$. Ainsi xy^{-1} est un point fixe de s. Vu l'hypothèse, $xy^{-1} = 1$, *i.e.* $x = y$. Comme G est fini, le théorème 15 de la page 71 montre que f est *surjective*.
Soit $x \in G$. Il existe donc $a \in G$ tel que $x = f(a) = a^{-1}s(a)$. Puisque $s \circ s = \mathrm{Id}_G$, il vient :

$$s(x) = s\big(a^{-1}s(a)\big) = s(a)^{-1}s\big(s(a)\big) = s(a)^{-1}a = \big(a^{-1}s(a)\big)^{-1} = x^{-1}\,.$$

Nous avons utilisé la définition d'un morphisme, la proposition 18 de la page 123 et la propriété 4 de la proposition 9 de la page 119. Ainsi $s(x) = x^{-1}$ pour tout $x \in G$.
Soient alors $a, b \in G$. On a $(ab)^{-1} = s(ab) = s(a)s(b) = a^{-1}b^{-1}$, soit $b^{-1}a^{-1} = a^{-1}b^{-1}$. En prenant les inverses, on en déduit que $ab = ba$, d'où la commutativité de G.

2. Soit $E := G \setminus \{1\}$. Pour tout $x \in E$, $x^{-1} = s(x)$ est distinct de x, puisque 1 est le seul point fixe de s. Les éléments de E sont donc « rangés par paires $\{x, x^{-1}\}$ », et il en résulte intuitivement que E a un nombre pair d'éléments. Dans ces conditions, $\mathrm{card}\,(G) = 1 + \mathrm{card}\,(E)$ est impair. Formellement, la relation $\sim$ sur E définie par $x \sim y$ si, et seulement si, $y = x$ ou $y = x^{-1}$ est une relation d'équivalence sur E $\big($à cause de la formule $(x^{-1})^{-1} = x\big)$. Vu ce qui précède, chaque classe de E pour $\sim$ a deux éléments. Le fait que $\mathrm{card}\,(E)$ soit pair résulte alors du théorème 16 de la page 72.

3. Supposons que G soit un groupe abélien fini d'ordre *impair* N. Définissons $s : G \to G$ par $s(x) := x^{-1}$ pour tout $x \in G$. Soient $a, b \in G$. Il vient $s(ab) = (ab)^{-1} = b^{-1}a^{-1} = a^{-1}b^{-1}$, la dernière égalité venant de la commutativité de G. Ainsi $s(ab) = s(a)s(b)$, donc s est un morphisme. Ensuite la formule $(x^{-1})^{-1} = x$ montre que $s \circ s$ est l'identité. Il reste à montrer que, si x est un point fixe de s, alors $x = 1$. Supposons le contraire : $x = s(x) = x^{-1}$, donc $x^2 = 1$, mais $x \neq 1$. Ainsi x est d'ordre 2. Cela contredit le théorème de Lagrange (théorème 22 de la page 125), puisque 2 ne divise pas N. En conclusion, s possède les propriétés requises : c'est un morphisme involutif de G dans G ayant un seul point fixe, à savoir 1.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

II.2.1 Sur un ensemble fini à n éléments, montrer qu'il y a n^{n^2} lois de composition. Parmi elles, combien sont commutatives ?

II.2.2 Soit S (resp. I) l'ensemble des fonctions paires (resp. impaires) de $\mathbb{R}$ dans $\mathbb{R}$. Montrer que S et I sont des parties additivement stables de $\mathcal{F}(\mathbb{R}, \mathbb{R})$, et que S est multiplicativement stable. Que dire de I ?

II.2.3 $*$ Soient G un groupe multiplicatif et H une partie multiplicativement stable de G, *finie* et non vide. Donner deux preuves du fait que H soit un sous-groupe de G (pour l'une d'elles, utiliser la proposition 10 de la page 119).

II.2.4 Soient G un groupe et $a \in G$. Montrer que les éléments de G commutant avec a forment un sous-groupe $C_G(a)$ de G. Quelle est l'intersection des $C_G(a)$ lorsque a décrit G ?

II.2.5 Soit G un groupe multiplicatif. On suppose que $x^2 = 1$ pour tout $x \in G$. Montrer que G est abélien.

II.2.6 Soit $a \in \mathbb{Z}$. Montrer que l'application $x \mapsto ax$ (multiplication par a) est un morphisme du groupe $(\mathbb{Z}, +)$ dans lui-même. À quelle condition (sur a) ce morphisme est-il injectif (resp. surjectif, resp. bijectif) ?

II.2.7 Soient $f, g : G \to G'$ deux morphismes de groupes. Montrer que, si f et g coïncident sur une partie génératrice de G, $f = g$. Vérifier d'abord que $\{x \in G \mid f(x) = g(x)\}$ est un sous-groupe de G.

II.2.8 Sur $G := \mathbb{Z} \times \{-1, 1\}$, on définit une loi $\top$ ainsi :

$$(a, \varepsilon)\top(b, \eta) := (a + \varepsilon b, \varepsilon\eta) \quad \text{pour tous} \quad a, b \in \mathbb{Z} \quad \text{et} \quad \varepsilon, \eta \in \{-1, 1\}.$$

1. Montrer que $(G, \top)$ est un groupe. Ce groupe est-il commutatif ?
2. Soient $a := (0, -1)$ et $b := (1, -1)$.
Montrer que a, b sont des éléments d'ordre 2 de G et que $\{a, b\}$ engendre G.

II.2.9 $*$ Soit G un groupe multiplicatif engendré par cinq éléments a, b, c, d, e. On suppose que les égalités suivantes sont vérifiées :

$$ab = c\,,\ bc = d\,,\ cd = e\,,\ de = a\,,\ ea = b. \qquad (*)$$

1. Montrer que G est engendré par a et b. Vérifier les égalités :

$$babab^2ab = a\,,\ ab^2aba = b\,,\ bab^2 = a^2 \quad \text{et} \quad ba^4b = a. \qquad (**)$$

2. En déduire les égalités suivantes : $cb^4c = b$, puis :

$$b^5 = a^{-2}\,,\ c^{5^2} = a^4\,, \ldots,\ a^{5^5} = a^{-32} \qquad (***)$$

3. Montrer que les ordres de a et b sont finis et premiers avec 10, puis que b appartient à $\langle a \rangle$. En déduire que $a^{11} = 1$ et $b = a^4$. Montrer enfin que G est soit trivial, soit cyclique d'ordre 11.
4. Soient G' un groupe multiplicatif cyclique d'ordre 11 et a un générateur de G'. Trouver des éléments b, c, d, e de G' vérifiant les égalités $(*)$.

II.2.10 $*$ Soient $(G, \times)$ un groupe *abélien* et $a, b \in G$. On suppose que a (resp. b) est d'ordre fini m (resp. n), et que m et n sont premiers entre eux. Montrer que ab est d'ordre mn.

II.2.11 Soient G un groupe *abélien* fini, noté additivement, et p un diviseur premier de l'ordre n de G. On se propose de montrer que G possède un élément d'ordre p (« Théorème de Cauchy »).
1. Soient $x_1, \ldots, x_n$ les éléments de G.
Pour tout i, soit $H_i := \langle x_i \rangle$, et soit $A := H_1 \times H_2 \times \cdots \times H_n$.
Montrer que $f : (y_1, \ldots, y_n) \mapsto y_1 + y_2 + \cdots + y_n$ est un morphisme surjectif de A sur G.
2. Prouver qu'il existe $i \in [\![1, n]\!]$ tel que p divise l'ordre de x_i. Conclure.

II.2.12 Soit f une bijection de E sur un ensemble F. Montrer que $s \mapsto f \circ s \circ f^{-1}$ est un *isomorphisme* de $\mathfrak{S}(E)$ sur $\mathfrak{S}(F)$. Ainsi les groupes $\mathfrak{S}(E)$ et $\mathfrak{S}(F)$ ont les mêmes propriétés.

II.2.13 Montrer que les applications $t : x \mapsto x + 1$ et $s : x \mapsto -x$ de $\mathbb{Z}$ dans $\mathbb{Z}$ appartiennent à $\mathfrak{S}(\mathbb{Z})$. Comparer $s \circ t$ et $t \circ s$. Conclusion ? Établir un isomorphisme entre le sous-groupe Γ de $\mathfrak{S}(\mathbb{Z})$ engendré par $\{s, t\}$ et le groupe G de l'exercice II.2.8.

II.2.14
Soit V_4 la partie $\{1,\ (1\ 2)(3\ 4),\ (1\ 3)(2\ 4),\ (1\ 4)(2\ 3)\}$ de $\mathfrak{S}_4$ (on supprime le symbole $\circ$).

1. Montrer que V_4 est un sous-groupe du groupe alterné $\mathfrak{A}_4$, puis que V_4 est isomorphe au groupe produit $(\mathbb{Z}/2\mathbb{Z})^2$.
2. Un groupe $(G, \times)$ d'ordre 4 dont tout élément est de carré 1 est isomorphe à V_4 (étudier la table de multiplication de G). À isomorphisme près, combien y a-t-il de groupes d'ordre 4 ?

II.2.15 Soient E un ensemble et c un n-cycle de E ($n \geqslant 2$). Montrer que c est un élément d'ordre n de $\mathfrak{S}(E)$.

II.2.16 Soient E un ensemble fini de cardinal N et n un entier, $2 \leqslant n \leqslant N$. Combien y a-t-il de cycles de longueur n dans $\mathfrak{S}(E)$?

II.2.17 Soit $s \in \mathfrak{S}_n$, $n \geqslant 2$. Prouver la formule :

$$\varepsilon(s) \;=\; \prod_{1 \leqslant i < j \leqslant n} \left(\frac{s(j) - s(i)}{j - i} \right).$$

On pourra utiliser la proposition 27 de la page 130.

II.2.18 $**$ Soient $n \geqslant 2$ un entier et V une partie *génératrice* de $\mathfrak{S}_n$ formée de *transpositions*. Montrer que $\operatorname{card}(V) \geqslant n - 1$.

Anneaux et corps

II.2.19 On reprend les notations de l'exercice 6 de la page 134. Vérifier que l'ensemble $\operatorname{End}(E)$ des morphismes de E dans lui-même est un sous-groupe de $\big(\mathcal{F}(E, E), +\big)$.
Montrer que $\big(\operatorname{End}(E), +, \circ\big)$ est un anneau. Quels sont ses éléments inversibles ?

II.2.20 Soient A, B deux anneaux commutatifs. Dans l'anneau produit $A \times B$, montrer que $A \times \{0\}$ et $\{0\} \times B$ sont des idéaux. Montrer que, si A et B ne sont pas nuls, l'anneau $A \times B$ n'est pas intègre.

II.2.21
1. Considérons un anneau produit $A := A_1 \times \cdots \times A_n$ (exemple 2 de la page 134).
Montrer que $A^\times$ est le groupe produit $A_1^\times \times \cdots A_n^\times$.
2. Soient X un ensemble et A un anneau. Montrer que les éléments inversibles de l'anneau $\mathcal{F}(X, A)$ (exemple 3 de la page 134) sont les applications $f : X \to A$ telles que $f(X) \subset A^\times$.

II.2.22 Soient A un anneau commutatif et $a \in A$, $n \in \mathbb{N}$. Prouver la formule :

$$\prod_{k=0}^{n} \big(a^{2^k} + 1\big) \;=\; \sum_{j=0}^{2^{n+1}-1} a^j.$$

II.2.23 $*$
1. Soient $r \geqslant 2$ un entier et $i_1, \ldots, i_r \in \mathbb{N}$.
Montrer que le nombre rationnel suivant est en fait un entier :

$$\frac{(i_1 + i_2 + \cdots + i_r)!}{i_1!\, i_2! \cdots i_r!}\,.$$

2. Soient A un anneau et $a_1, \ldots, a_r$ des éléments de A commutant deux à deux ($r \geqslant 2$).
Démontrer, pour tout $n \in \mathbb{N}$, la *formule du multinôme* :

$$(a_1 + a_2 + \cdots + a_r)^n \;=\; \sum_{(i_1,\ldots,i_r)\,\in\,E} \left(\frac{(i_1 + i_2 + \cdots + i_r)!}{i_1!\, i_2! \cdots i_r!}\right) a_1^{i_1} a_2^{i_2} \cdots a_r^{i_r}\,,$$

où E est l'ensemble des r-uplets $(i_1, \ldots, i_r) \in \mathbb{N}^r$ de somme n.

II.2.24 Soit A un anneau. Montrer que, si $x \in A$ est nilpotent, $1 - x$ est inversible.
La réciproque est-elle vraie ?

II.2.25 Soit e un idempotent de l'anneau A. Montrer que $(eAe, +, \times)$ est un anneau.

II.2.26 Montrer que, si $f, g : A \to A'$ sont deux morphismes d'anneaux, alors l'ensemble $\{x \in A \mid f(x) = g(x)\}$ est un sous-anneau de A.

II.2.27 $*$ 1. Soient p un nombre premier fixé et A l'ensemble des rationnels de la forme a/b, où $a, b \in \mathbb{Z}$ et b n'est pas multiple de p. Montrer que A est un sous-anneau de $\mathbb{Q}$. Quels sont les éléments inversibles de A ?
2. Montrer que le sous-anneau de $\mathbb{Q}$ engendré par l'ensemble des inverses des nombres premiers est $\mathbb{Q}$. Montrer qu'il n'y a pas de partie *finie* S de $\mathbb{Q}$ telle que le sous-anneau de $\mathbb{Q}$ engendré par S soit $\mathbb{Q}$.

II.2.28 Montrer que les éléments nilpotents d'un anneau commutatif A forment un idéal de A.
Soit $n \geqslant 2$ un entier, $k \in \mathbb{Z}$ et $\overline{k} \in \mathbb{Z}/n\mathbb{Z}$ son image. Montrer que $\overline{k}$ est nilpotent si, et seulement si, tout diviseur premier de n divise k.

II.2.29 Soit $X := \mathbb{N} \setminus \{0\}$. On définit sur $\mathcal{F}(X, \mathbb{Z})$ une loi $\star$ comme suit.
Si $f, g \in \mathcal{F}(X, \mathbb{Z})$, $f \star g : X \to \mathbb{Z}$ est donnée par :

$$(f \star g)(n) \;:=\; \sum_{(a,b)\in X^2,\; ab\,=\,n} f(a)\, g(b) \quad \text{pour tout } n \in X\,.$$

1. Vérifier que la fonction $\delta : X \mapsto \mathbb{Z}$ valant 1 en 1 et 0 ailleurs est élément neutre pour $\star$.
2. Montrer que la loi $\star$ est associative $\big($utiliser les formules (6) à (8)$\big)$, puis que $\big(\mathcal{F}(X, \mathbb{Z}), +, \star\big)$ est un anneau commutatif.
3. Soit $\mu : X \to \mathbb{Z}$ la *fonction de Möbius*, définie ainsi. Soit $n \in X$; $\mu(n) := (-1)^r$ si n est produit de r nombres premiers *distincts* ($r \geqslant 0$), et $\mu(n) := 0$ sinon. Soit $c : X \to \mathbb{Z}$ la fonction constante égale à 1. Montrer que c et μ sont inverses l'une de l'autre dans $\big(\mathcal{F}(X, \mathbb{Z}), +, \star\big)$.
4. Étant donnés un groupe abélien $(E, +)$ et deux fonctions $f, g : X \to E$, établir l'équivalence suivante (*formule d'inversion de Möbius*) :

$$\Big(\forall n \in X,\; g(n) \;=\; \sum_{d|n} f(d)\Big) \Longleftrightarrow \Big(\forall n \in X,\; f(n) \;=\; \sum_{d|n} \mu(n/d) g(d)\Big).$$

Dans ces sommes, d décrit l'ensemble des diviseurs positifs de n.

II.2.30 Soit $n := 561 = 3 \times 11 \times 17$. Montrer grâce au théorème chinois que n a la propriété suivante : il n'est pas premier mais, pour tout $a \in \mathbb{Z}$ premier avec n, on a : $a^{n-1} \equiv 1 \mod n$ (1729 a la même propriété).

II.2.31 Soit $n > 0$ un entier *impair*. Montrer que le nombre de solutions de l'équation $x^2 = 1$ dans l'anneau $\mathbb{Z}/n\mathbb{Z}$ est 2^m, où m est le nombre de facteurs premiers de n distincts (commencer par le cas $m = 1$).

II.2.32 Soient $n \in \mathbb{N}$ et p un facteur premier de $2^{2^n} + 1$. Montrer que $p - 1$ est multiple de 2^{n+1}. On pourra considérer l'ordre de $\bar{2} \in (\mathbb{Z}/p\mathbb{Z})^\times$.

II.2.33 Si $a, b \in \mathbb{Z}$, $ab(a^{60} - b^{60})$ est multiple de $56\,786\,730$.

II.2.34 Soit $N \geqslant 2$ un entier. Montrer que la suite $(\overline{n}^n)_{n \geqslant 1}$ d'éléments de $\mathbb{Z}/N\mathbb{Z}$ est périodique à partir d'un certain rang.

II.2.35 Montrer qu'un corps de cardinal 4 est de caractéristique 2. Soit K un ensemble à quatre éléments, notés $0, 1, a, b$. Montrer qu'il existe sur K une addition $+$ et une multiplication $\times$ uniques faisant de K un corps, ayant pour élément neutre additif 0 et pour élément neutre multiplicatif 1 (écrire les tables d'addition et de multiplication).

II.2.36 $*$ Un anneau *intègre fini* est un corps (donner deux preuves).

II.2.37 Soit X un ensemble. Notons $K := \{0, 1\}$ le corps $\mathbb{Z}/2\mathbb{Z}$, et considérons l'anneau $\mathcal{F}(X, K)$ (*cf.* l'exemple 3 de la page 134). Si A est une partie de X, sa fonction caractéristique χ_A (*cf.* la définition 4 de la page 76) est considérée ici comme élément de $\mathcal{F}\big(X, \{0, 1\}\big) = \mathcal{F}(X, K)$.
1. Montrer que $\chi : A \mapsto \chi_A$ est une bijection de $\mathcal{P}(X)$ sur $\mathcal{F}(X, K)$, et expliciter la bijection réciproque.
2. Reprenons les notations de l'exemple de la page 137. Si $A, B \in \mathcal{P}(X)$, calculer χ_{A+B} et χ_{AB}. En déduire que $\big(\mathcal{P}(X), +, \times\big)$ est un anneau.

II.2.38 Soient A, A' deux anneaux intègres et K, K' leurs corps des fractions respectifs. Montrer que tout morphisme d'anneaux *injectif* $f : A \to A'$ se prolonge *de manière unique* en un morphisme d'anneaux $F : K \to K'$ (F « prolonge » f si $F(a) = f(a)$ pour tout $a \in A$).

II.2.39 Soient K un corps et k le plus petit sous-corps de K, au sens de l'inclusion (*cf.* l'exemple 2 de la page 154). Si K est de caractéristique 0, montrer qu'il existe un unique isomorphisme d'anneaux f de $\mathbb{Q}$ sur k tel que $f(a/b) = (a \cdot 1_K)(b \cdot 1_K)^{-1}$ pour tous $a \in \mathbb{Z}$ et $b \in \mathbb{Z}^*$. On suppose que $\operatorname{card}(K)$ est un nombre premier p. Montrer qu'il existe un unique isomorphisme d'anneaux f de $\mathbb{Z}/p\mathbb{Z}$ sur k tel que $f(\overline{a}) = a \cdot 1_K$ pour tout $a \in \mathbb{Z}$.

II.2.40 Résoudre les équations $x^3 = 1$ et $x^5 = 1$ dans chacun des corps suivants : $\mathbb{R}$, $\mathbb{Z}/11\mathbb{Z}$, $\mathbb{Z}/19\mathbb{Z}$, $\mathbb{C}$. Dans le cas de $\mathbb{C}$, il faut utiliser l'exponentielle complexe, étudiée dans le module II.5.

II.2.41 $*$ Cet exercice utilise les réels.
Soit A (resp. K) l'ensemble des nombres réels de la forme $a + b\sqrt{10}$, où $a, b \in \mathbb{Z}$ (resp. $a, b \in \mathbb{Q}$).
1. Montrer que $\sqrt{10}$ est irrationnel (*cf.* la proposition 66 de la page 104). En déduire que tout $x \in K$ s'écrit *de manière unique* $x = a + b\sqrt{10}$, où $(a, b) \in \mathbb{Q}^2$. 2. Montrer que A est un sous-anneau de $\mathbb{R}$ et que K est un sous-corps de $\mathbb{R}$. Montrer que K est isomorphe au corps des fractions de A.

3. Si $x := a + b\sqrt{10} \in K$ $\big((a,b) \in \mathbb{Q}^2\big)$, on pose $N(x) := a^2 - 10b^2$.
Montrer que $N : (K^*, \times) \to (\mathbb{Q}^*, \times)$ est un morphisme, et que $N(A^*) \subset \mathbb{Z}^*$, où $A^* := A \setminus \{0\}$ et $\mathbb{Z}^* := \mathbb{Z} \setminus \{0\}$.
4. Montrer que $A^\times = \big\{x \in A \bigm| N(x) = \pm 1\big\}$, puis que $A^\times$ est infini.
5. Soient $a, b \in \mathbb{Z}$. Montrer que $a^2 - 10b^2 \neq \pm 2$ (considérer le chiffre des unités de a). En déduire que 2 est un élément *irréductible* de A.
6. À partir de l'égalité $(4 + \sqrt{10})(4 - \sqrt{10}) = 2 \times 3$, montrer que l'anneau A n'est pas principal (penser au lemme d'Euclide).

Espaces vectoriels et applications linéaires

II.3

L'algèbre linéaire a deux sources. La première est la géométrie d'Euclide. Les élégants raisonnements de la géométrie plane (ainsi que de la géométrie dite « dans l'espace ») se structurent autour de deux notions clés : celle d'addition des vecteurs, pour laquelle la physique fournit d'ailleurs également des exemples ; et celle de transformation, du plan et de l'espace. Ces deux notions à elles seules justifient l'introduction des espaces vectoriels et des applications linéaires dans l'enseignement. L'autre source, plus diffuse, est la découverte, tout au long de la première moitié du vingtième siècle, que *toutes* les grandes théories mathématiques mettent en jeu des phénomènes *linéaires* et que la structure d'espace vectoriel y est omniprésente. On en trouvera de nombreuses illustrations dans les cours d'algèbre et de géométrie de cette année, mais aussi dans le cours d'analyse des années suivantes. Par exemple, on peut comprendre le calcul différentiel comme un effort pour approcher des fonctions quelconques par des fonctions linéaires.

Cette ubiquité de l'algèbre linéaire nous conduira à définir nos espaces vectoriels sur un corps commutatif K quelconque. Le présent module vise à introduire le vocabulaire et les idées générales qui simplifieront, dès les modules suivants, l'exposé de la théorie des matrices, des nombres complexes et des polynômes. Réciproquement, ces modules illustreront et motiveront ces généralités. La théorie deviendra plus substantielle au module II.7, avec l'étude des espaces vectoriels de dimension finie.

1 Vocabulaire et propriétés élémentaires

1.1 La structure d'espace vectoriel

Dans tout ce module, la lettre K désigne un corps commutatif.

Définition 1. On dit qu'un groupe abélien $(V,+)$ est muni d'une *structure d'espace vectoriel sur le corps* K si l'on a une loi de composition (ou multiplication) externe $(\lambda, v) \mapsto \lambda.v$ de $K \times V$ dans V vérifiant les axiomes suivants :

$$\forall \lambda, \mu \in K \ , \ \forall x \in V \ , \ (\lambda + \mu).x = \lambda.x + \mu.x,$$

$$\forall \lambda \in K \ , \ \forall x, y \in V \ , \ \lambda.(x + y) = \lambda.x + \lambda.y,$$

$$\forall \lambda, \mu \in K \ , \ \forall x \in V \ , \ \lambda.(\mu.x) = (\lambda\mu).x,$$

$$\forall x \in V \ , \ 1.x = x.$$

Dans ces axiomes, le symbole 1 désigne l'unité de K et le symbole $+$ désigne aussi bien l'addition du groupe K que celle du groupe V.
Le point de la notation $\lambda.x$ est en général omis dans l'écriture : $\lambda.x = \lambda x$.

On dit alors que $(V,+,.)$ ou (abusivement) que V est un espace vectoriel sur le corps K ou un K-espace vectoriel. Les éléments de V sont parfois appelés *vecteurs* et ceux de K, *scalaires*. On écrit toujours les scalaires *à gauche* des vecteurs dans la multiplication externe. Le lecteur est invité à examiner les formules ci-dessus et à décider, pour chaque occurrence d'une addition ou d'une multiplication, s'il s'agit de celle de K ou de V.

Les règles de calcul suivantes sont des conséquences faciles de la définition. Nous notons ici 0_V l'élément neutre du groupe V.

$$\forall \lambda \in K \ , \ \lambda.0_V = 0_V,$$

$$\forall x \in V \ , \ 0.x = 0_V,$$

$$\forall \lambda \in K \ , \ \forall x \in V \ , \ \lambda.\left(-x\right) = -\left(\lambda.x\right),$$

$$\forall \lambda \in K \ , \ \forall x \in V \ , \ (-\lambda).x = -\left(\lambda.x\right)$$

$$\forall \lambda \in K \ , \ \forall x \in V \ , \ \lambda.x = 0_V \Longrightarrow \lambda = 0 \text{ ou } x = 0_V.$$

Démontrons la dernière règle. Si $\lambda.x = 0_V$ et si $\lambda \neq 0$, on a, par applications successives des axiomes : $x = 1.x = (\lambda^{-1}.\lambda).x = \lambda^{-1}.(\lambda.x) = \lambda^{-1}.0_V = 0_V$ (cette dernière égalité résultant de la première règle).

Les propriétés suivantes étendent celles connues pour les groupes abéliens. Pour tout entier $m \in \mathbb{Z}$, on sait définir l'élément mx du groupe abélien V et l'élément $m.1$ du groupe abélien K. On vérifie :

$$m.x = (m.1).x.$$

De même, pour toute famille finie $(x_1, \ldots, x_p)$ d'éléments de V, on a :

$$\forall \lambda \in K \ , \ \lambda . \sum_{i=1}^{p} x_i = \sum_{i=1}^{p} (\lambda . x_i).$$

Plus généralement, pour toute famille $(x_i)_{i \in I}$ d'éléments de V, supposée *à support fini* (c'est-à-dire telle que les x_i sont nuls sauf pour un nombre fini d'indices i), on sait définir l'élément $\sum_{i \in I} x_i$ de V. Avec ces notations :

$$\forall \lambda \in K \ , \ \lambda . \sum_{i \in I} x_i = \sum_{i \in I} (\lambda . x_i).$$

Ces règles sont à rapprocher des égalités :

$$\sum_{i=1}^{p} x_i + \sum_{i=1}^{p} y_i = \sum_{i=1}^{p} (x_i + y_i) \quad \text{et} \quad \sum_{i \in I} x_i + \sum_{i \in I} y_i = \sum_{i \in I} (x_i + y_i),$$

valides dans tout groupe abélien.

Exemples.

1. Avec les définitions usuelles de l'addition vectorielle et de la multiplication d'un vecteur par un réel, on munit l'ensemble des vecteurs du plan (ou de l'espace) d'une structure d'espace vectoriel sur le corps $\mathbb{R}$. De manière plus précise, notons Π le plan rapporté au repère $(O, \vec{\imath}, \vec{\jmath})$. Tout vecteur s'écrit de manière unique sous la forme $x\vec{\imath} + y\vec{\jmath}$ avec $x, y \in \mathbb{R}$ et la loi d'addition des vecteurs (règle du parallélogramme) et la dilatation (ou homothétie) de rapport $\lambda \in \mathbb{R}$ se traduisent par les formules :

$$(x\vec{\imath} + y\vec{\jmath}) + (x'\vec{\imath} + y'\vec{\jmath}) = (x + x')\vec{\imath} + (y + y')\vec{\jmath},$$
$$\lambda(x\vec{\imath} + y\vec{\jmath}) = (\lambda x)\vec{\imath} + (\lambda y)\vec{\jmath}.$$

   L'espace vectoriel des vecteurs du plan (ou plan vectoriel) s'identifie donc à l'espace vectoriel obtenu en munissant le groupe $(\mathbb{R}^2, +)$ de la loi externe définie par $\lambda.(x, y) = (\lambda x, \lambda y)$.

2. En posant $\lambda.(x_1, \ldots, x_n) := (\lambda x_1, \ldots, \lambda x_n)$, on fait du groupe abélien $(K^n, +)$ un K-espace vectoriel. *Le lecteur est vivement encouragé à vérifier en détail les axiomes de la définition sur cet exemple !* Si l'on prend $K = \mathbb{R}$ et $n = 2$, on retrouve le plan vectoriel. Si l'on prend $n = 1$, on trouve que K lui-même est un K-espace vectoriel.

3. L'ensemble $M_{n,1}(K)$ des *vecteurs colonnes* est également formé des n-uplets d'éléments de K, mais *écrits verticalement*. La structure d'espace vectoriel de $M_{n,1}(K)$ est définie par les opérations :

$$\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} := \begin{pmatrix} \lambda_1 + \mu_1 \\ \vdots \\ \lambda_n + \mu_n \end{pmatrix} \qquad \text{et} \qquad \lambda \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} := \begin{pmatrix} \lambda\lambda_1 \\ \vdots \\ \lambda\lambda_n \end{pmatrix}$$

   L'ensemble $M_{1,n}(K)$ des *vecteurs lignes* est semblablement muni d'une structure d'espace vectoriel qui s'identifie naturellement à $(K^n, +)$ étudié ci-dessus.

4. L'ensemble $K[X]$ des polynômes à coefficients dans le corps K sera muni, au module II.6, d'une structure d'espace vectoriel sur K.
5. L'addition des fonctions et la multiplication d'une fonction par un scalaire définissent le $\mathbb{R}$-espace vectoriel $\mathcal{F}(I,\mathbb{R})$ des fonctions numériques sur l'intervalle I de $\mathbb{R}$.
6. On appelle *espace vectoriel trivial* ou *nul* tout espace vectoriel réduit à un seul élément : cet élément est nécessairement l'élément neutre de l'addition et le groupe sous-jacent est le groupe trivial $\{0\}$.
7. Soient $K \subset L$ deux corps commutatifs et soit V un L-espace vectoriel. En restreignant la multiplication externe $L \times V \to V$ à $K \times V \to V$, on munit V d'une structure de K-espace vectoriel.
8. Par exemple, L lui-même est un L-espace vectoriel et donc également un K-espace vectoriel : ainsi, $\mathbb{R}$ est un $\mathbb{R}$-espace vectoriel et un $\mathbb{Q}$-espace vectoriel ; de même, $\mathbb{C}$ est un $\mathbb{C}$-espace vectoriel, un $\mathbb{R}$-espace vectoriel et un $\mathbb{Q}$-espace vectoriel.

Les deux derniers exemples montrent qu'il est parfois indispensable de préciser le corps de base lorsque l'on parle d'espaces vectoriels. Par exemple, le $\mathbb{C}$-espace vectoriel $\mathbb{C}$ est engendré par l'élément 1 (*cf* définition 6 de la page 170), il est donc de dimension 1 (*cf* paragraphe 3.4). Mais le $\mathbb{R}$-espace vectoriel $\mathbb{C}$ ne peut être engendré par une famille à un seul élément (nous verrons au module II.5 qu'il est de dimension 2). Le $\mathbb{C}$-espace vectoriel $\mathbb{C}$ et le $\mathbb{R}$-espace vectoriel $\mathbb{C}$ sont donc différents !

Nous allons traiter à part un exemple important, qui généralise d'ailleurs certains des exemples ci-dessus. Soient I un ensemble et V un K-espace vectoriel. L'ensemble $\mathcal{F}(I,V)$ des applications de I dans V a été muni, au module II.2 (exemple 3 de la page 118), d'une structure de groupe : la somme $f+g$ de $f,g \in \mathcal{F}(I,V)$ est l'application $i \mapsto f(i)+g(i)$. Définissons de même, pour tout $\lambda \in K$, l'application $\lambda f : i \mapsto \lambda f(i)$. On fait ainsi de $\mathcal{F}(I,V)$ un K-espace vectoriel. Rappelons d'autre part qu'une famille d'éléments de V indexée par I est, à la notation près, une application de I dans V : la famille $(x_i)_{i\in I} \in V^I$ d'éléments de V s'identifie à l'application $f : I \to V$ si $x_i = f(i)$. L'addition et la multiplication externe dans l'espace vectoriel V^I sont donc données par les formules $(x_i)_{i\in I} + (y_i)_{i\in I} := (x_i + y_i)_{i\in I}$ et $\lambda.(x_i)_{i\in I} := (\lambda x_i)_{i\in I}$. Le cas le plus utile de cette construction est celui où $V = K$: on obtient le K-espace vectoriel $K^I := \mathcal{F}(I,K)$. Lorsque de plus $I = [\![1,n]\!]$, l'espace vectoriel K^I s'identifie naturellement à K^n et l'on retrouve l'exemple 2.

Exercice 1.

Soit E un ensemble quelconque. On prend $V = \mathcal{P}(E)$, l'ensemble des parties de E. On le munit de la loi interne :

$$A \oplus B = (A \setminus B) \cup (B \setminus A) \quad \text{(différence symétrique).}$$

On sait que $(V,\oplus)$ est un groupe abélien dont l'élément neutre est la partie vide $\varnothing$, que nous noterons 0, et tel que tout élément de V est son propre opposé. On va définir sur V une loi externe $\mathbb{F}_2 \times V \to V$ sur le corps à deux éléments $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$; pour cela, on pose, pour tout $x \in V$: $0.x := 0$ et $1.x := x$. Démontrer que l'on obtient ainsi un $\mathbb{F}_2$-espace vectoriel.

Solution. On vérifie tous les axiomes en donnant aux scalaires λ, μ successivement les valeurs 0 et 1. La seule vérification qui n'est pas tautologique est celle de l'égalité : $(1+1).x = 1.x \oplus 1.x$. Le membre gauche vaut $0.x$ (car $1+1=0$ dans $\mathbb{F}_2$) donc 0, par définition de la multiplication ; le membre droit vaut $x \oplus x$. La validité de cet axiome tient donc à la propriété spéciale de notre groupe, à savoir que tout élément est son propre opposé.

1.2 Combinaisons linéaires

Soient $\lambda, \mu, \nu, \ldots$ des scalaires et $x, y, z, \ldots$ des vecteurs. On peut former les expressions $\lambda x + \mu y, \lambda x + \mu y + \nu z, \ldots$. Plus généralement, soit $(x_1, \ldots, x_p)$ une suite finie de vecteurs, donc une famille indexée par $[\![1, p]\!] \subset \mathbb{N}$. Soit de même une suite finie $(\lambda_1, \ldots, \lambda_p)$ d'éléments du corps K. Alors $\lambda_1 x_1, \ldots, \lambda_p x_p$ sont des vecteurs et l'on peut former leur somme :

$$\sum_{i=1}^{p} \lambda_i x_i = \lambda_1 x_1 + \cdots + \lambda_p x_p.$$

Celle-ci est appelée *combinaison linéaire des (vecteurs)* x_i et les (scalaires) λ_i sont appelés *coefficients* de cette combinaison linéaire.

Plus généralement encore, soit $(x_i)_{i \in I}$ une famille d'éléments du K-espace vectoriel V indexée par l'ensemble I. Soit par ailleurs $(\lambda_i)_{i \in I}$ une famille d'éléments de K indexée par I. On suppose la famille $(\lambda_i)_{i \in I}$ *à support fini* : l'ensemble des indices $i \in I$ tels que $\lambda_i \neq 0$ est un ensemble fini ; on dit aussi que « presque tous les λ_i sont nuls » (cette condition est bien sûr vérifiée si l'ensemble I est fini). Alors les $\lambda_i x_i \in V$ sont presque tous nuls, et la somme $\sum_{i \in I} \lambda_i x_i$ est un élément bien défini du groupe V.

Définition 2. Toute telle somme $\sum_{i \in I} \lambda_i x_i$ est appelée *combinaison linéaire* des (vecteurs) x_i. Les (scalaires) λ_i sont les *coefficients* de cette combinaison linéaire.

Dans le cas d'une famille finie, on prendra presque toujours l'ensemble d'indices I de la forme $[\![1, p]\!] \subset \mathbb{N}$, ce qui nous ramènera à la première forme de la définition. Nous allégerons parfois la notation des familles finies. Par exemple, quand on parlera de la famille (x, y, z) formée de trois éléments de V, il s'agira de la famille $(x_i)_{i \in I}$ indexée par $I = \{1, 2, 3\}$ et telle que $x_1 = x$, $x_2 = y$ et $x_3 = z$. Devant la notation $(x_1, \ldots, x_n)$, le lecteur devra donc se demander s'il a affaire à une famille de n éléments d'un espace vectoriel V ou bien à un élément de l'espace vectoriel K^n.

Plus généralement, étant donnés une famille $(x_i)_{i \in I}$ quelconque et un vecteur x qui est combinaison linéaire des x_i, il résulte de la définition que x est combinaison linéaire d'un nombre fini de ces vecteurs : il suffit de ne pas tenir compte de ceux qui sont affectés d'un coefficient nul.
Il existe donc $i_1, \ldots, i_p \in I$ distincts tels que x est combinaison linéaire de $x_{i_1}, \ldots, x_{i_p}$:

$$x = \sum_{i \in I} \lambda_i x_i = \lambda_{i_1} x_{i_1} + \cdots + \lambda_{i_p} x_{i_p}.$$

Pour simplifier les notations, on emploiera souvent la formulation (légèrement abusive) suivante : il existe des éléments $x_1, \ldots, x_p$ (extraits de notre famille infinie) et des scalaires $\lambda_1, \ldots, \lambda_p \in K$ tels que $x = \lambda_1 x_1 + \cdots + \lambda_p x_p$.

Exemples.

1. Les combinaisons linéaires de la famille à un seul élément (x) sont les λx, $\lambda \in K$, celles de la famille à deux éléments (x, y) sont les $\lambda x + \mu y$, $\lambda, \mu \in K$, etc.
2. L'élément neutre 0 est une combinaison linéaire des x_i (prendre tous les coefficients nuls). Chaque x_{i_0} est une combinaison linéaire des x_i (prendre son coefficient égal à 1 et tous les autres nuls).
3. La famille $(X^n)_{n\in\mathbb{N}}$ des monômes en X est une famille d'éléments du K-espace vectoriel $K[X]$. Tout polynôme est une somme finie $a_o X^0 + \cdots + a_d X^d$, donc une combinaison linéaire $\sum\limits_{n\in\mathbb{N}} a_n X^n$ dans laquelle on a pris tous les coefficients a_n d'indice $n > d$ nuls.
4. Soit $V := K^{\mathbb{N}}$ l'espace vectoriel des suites $(x_n)_{n\in\mathbb{N}}$ d'éléments de K. Pour tout entier $n \in \mathbb{N}$, soit $\delta^{(n)}$ la suite dont le $n^{ème}$ terme vaut 1 et les autres 0 ; avec la notation de Kronecker (voir exemple de la page 56) : $\forall i \in \mathbb{N}$, $\delta_i^{(n)} = \delta_{i,n}$. La combinaison linéaire $\sum\limits_{n\in\mathbb{N}} \lambda_n \delta^{(n)}$ n'est définie que si presque tous les λ_n sont nuls. Dans ce cas, c'est la suite $(\lambda_n)_{n\in\mathbb{N}}$. Les combinaisons linéaires des $\delta^{(n)}$ sont donc exactement les suites à support fini.

Dans le dernier exemple, on pourrait considérer une suite quelconque $(\lambda_n)_{n\in\mathbb{N}}$ d'éléments du corps K comme égale à l'expression $\sum\limits_{n\in\mathbb{N}} \lambda_n \delta^{(n)}$, alors même que celle-ci comporte une infinité de coefficients non nuls. Cependant, on ne considérera pas une telle expression comme une combinaison linéaire : elle n'a pas pu être obtenue par application répétée des opérations de l'espace vectoriel et ne relève donc pas du calcul algébrique. De la même manière, lorsque l'on écrit $\sum\limits_{n\geqslant 0} u_n$ la somme d'une série de réels ou de complexes, il ne s'agit pas d'une somme algébrique mais du résultat d'un passage à la limite.

On peut appliquer à une *partie* A de V ce qui a été dit pour une famille d'éléments de V, et l'on parlera donc de *combinaison linéaire d'éléments de* A. Si l'ensemble A est fini, il y a alors avantage à numéroter ses éléments : $A = \{x_1, \ldots, x_p\}$ et les combinaisons linéaires des éléments de A sont identiques aux combinaisons linéaires de la famille finie $(x_1, \ldots, x_p)$.

Un cas spécial mérite un commentaire. De même que la somme des éléments d'une famille vide ou d'une partie vide d'un groupe abélien est, par convention, l'élément neutre 0 de ce groupe, nous dirons que toute combinaison linéaire d'éléments d'une famille vide ou d'une partie vide vaut 0.

La règle de calcul suivante découle des précédentes :

$$\lambda \sum_{i\in I} \lambda_i x_i + \mu \sum_{i\in I} \mu_i x_i = \sum_{i\in I} (\lambda\lambda_i + \mu\mu_i)\, x_i. \tag{1}$$

Une combinaison linéaire de deux combinaisons linéaires des x_i est donc elle-même une combinaison linéaire des x_i. Plus généralement, toute combinaison linéaire de combinaisons linéaires des x_i est elle-même une combinaison linéaire des x_i, mais il faut des notations pour le formuler correctement.

Proposition 1. Soit $(x_i)_{i\in I}$ une famille d'éléments de V indexée par l'ensemble I. Soit un ensemble J d'indices et, pour tout indice $j \in J$, une famille de coefficients $(\lambda_i^{(j)})_{i\in I}$ (une famille d'éléments de K indexée par I). Soit enfin $(\mu_j)_{j\in J}$ une famille d'éléments de K indexée par J. On suppose que toutes ces familles de coefficients sont à support fini. Alors :

$$\sum_{j\in J} \mu_j \left(\sum_{i\in I} \lambda_i^{(j)} x_i \right) = \sum_{i\in I} \left(\sum_{j\in J} \mu_j \lambda_i^{(j)} \right) x_i. \tag{2}$$

Nous ne démontrerons pas cette formule générale. Dans tous les cas où nous l'appliquerons, elle sera évidente. Si par exemple J a deux éléments, mettons a et b, notant $\lambda := \mu_a$, $\mu := \mu_b$, $\lambda_i := \lambda_i^{(a)}$ et $\mu_i := \lambda_i^{(b)}$, on retrouve 1. Lorsque les ensembles d'indices sont finis, par exemple $I = [\![1, p]\!]$ et $J = [\![1, q]\!]$, la formule (2) devient :

$$\sum_{j=1}^{q} \mu_j \left(\sum_{i=1}^{p} \lambda_i^{(j)} x_i \right) = \sum_{i=1}^{p} \left(\sum_{j=1}^{q} \mu_j \lambda_i^{(j)} \right) x_i.$$

Une autre propriété élémentaire des combinaisons linéaires ne fait que reprendre une propriété analogue des sommes dans un groupe abélien. Nous la formulons d'abord sous une forme très concrète et dans un cas très simple :

$$\sum_{i=1}^{p+q} \lambda_i x_i = \sum_{i=1}^{p} \lambda_i x_i + \sum_{i=p+1}^{p+q} \lambda_i x_i. \tag{3}$$

La généralisation naturelle de cette égalité est la suivante. On se donne une famille $(x_i)_{i\in I}$ de vecteurs et une famille $(\lambda_i)_{i\in I}$ de scalaires toutes deux indexées par l'ensemble I. On se donne par ailleurs une partition $(I_j)_{j\in J}$ de l'ensemble I : autrement dit, I est la réunion de ses sous-ensembles I_j et ceux-ci sont deux à deux disjoints. On a alors :

$$\sum_{i\in I} \lambda_i x_i = \sum_{j\in J} \left(\sum_{i\in I_j} \lambda_i x_i \right).$$

La formule (3) correspond au cas où $I = [\![1, p+q]\!]$ est décomposé selon la partition formée de $I_1 = [\![1, p]\!]$ et de $I_2 = [\![p+1, p+q]\!]$.

Définition 3. Nous conservons les notations précédentes. Une *relation linéaire* entre les éléments de la famille $(x_i)_{i\in I}$ est une égalité de la forme :

$$\sum_{i\in I} \lambda_i x_i = 0.$$

C'est une *relation linéaire triviale* si tous les coefficients λ_i sont nuls, une *relation linéaire non triviale* autrement.

Exemples.

1. Toute relation linéaire non triviale concernant la famille à un seul élément x a la forme $\lambda x = 0$ avec $\lambda \neq 0$. Elle entraîne donc $x = 0$.

2. Toute relation linéaire non triviale concernant la famille à deux éléments (x, y) a la forme $\lambda x + \mu y = 0$ avec λ, μ non tous deux nuls. Si $\lambda \neq 0$, on en déduit $x = \dfrac{-\mu}{\lambda} y$; si $\mu \neq 0$, on en déduit $y = \dfrac{-\lambda}{\mu} x$. Si λ ou μ est nul, l'un au moins des éléments de la famille est nul. Si λ et μ sont tous deux non nuls, chacun des deux éléments de la famille est multiple de l'autre. Dans tous les cas, l'un au moins des deux éléments de la famille est multiple de l'autre. On dira qu'ils sont *proportionnels*.

1.3 Sous-espaces vectoriels

Définition 4. Soit $(V, +, .)$ un espace vectoriel sur K. On dit que $W \subset V$ en est un *sous-espace vectoriel* (ou simplement un *sous-espace*) si W est un sous-groupe et s'il est de plus stable par multiplication externe :

$$\forall \lambda \in K \ , \ \forall x \in W \ , \ \lambda x \in W.$$

Par exemple, les sous-groupes $\{0\}$ et V de V en sont des sous-espaces.

Exercice 2.
Démontrer que le K-espace vectoriel K n'admet pas d'autres sous-espaces vectoriels que $\{0\}$ et K.
Solution. Soit en effet $W \subset K$ un sous-espace vectoriel autre que $\{0\}$. Soit $x \in W \setminus \{0\}$. Alors, pour tout $y \in K$, l'élément $\lambda = yx^{-1} \in K$ vérifie $\lambda.x = y$.
Or, par définition, $\lambda.x \in W$. On a donc $W = K$.

La restriction de l'application $(\lambda, x) \mapsto \lambda.x$ définit alors une application $K \times W \to W$, qui munit celui-ci de la *structure d'espace vectoriel induite* par celle de V. Toutes les fois que l'on aura un sous-espace vectoriel W d'un espace vectoriel V, il sera implicitement muni de cette structure induite (et donc vu lui-même comme un espace vectoriel).

Proposition 2. Soit W une partie non vide de l'espace vectoriel V. Alors W est un sous-espace vectoriel de V si, et seulement si, l'une des trois conditions équivalentes suivantes est vérifiée :

(i) $\forall \lambda, \mu \in K \ , \ \forall x, y \in W \ , \ \lambda x + \mu y \in W$.

(ii) $\forall \lambda \in K \ , \ \forall x, y \in W \ , \ \lambda x + y \in W$.

(iii) $\forall x, y \in W \ , \ x + y \in W$ et $\forall \lambda \in K \ , \ \forall x \in W \ , \ \lambda x \in W$.

Démonstration. Si W est un sous-espace vectoriel et si $\lambda, \mu \in K$ et $x, y \in W$, alors $\lambda x, \mu y \in W$ (stabilité pour la loi externe), donc $\lambda x + \mu y \in W$ (c'est un sous-groupe) : la condition (i) est satisfaite. Celle-ci implique à son tour la condition (ii) (prendre $\mu = 1$).
Supposons (ii) vérifiée. En prenant $\lambda = 1$, on voit que W est stable par addition, en prenant $y = 0$, on voit qu'il est stable pour la loi externe, d'où la condition (iii).
Supposons (iii) vérifiée. En prenant $\lambda = -1$, on voit que W est stable par passage à l'opposé ; comme il est stable par par addition, c'est un sous-groupe, et, comme il est stable pour la loi externe, c'est un sous-espace vectoriel. ∎

Corollaire 3. L'ensemble des combinaisons linéaires d'éléments d'une partie A de V (resp. l'ensemble des combinaisons d'éléments d'une famille $(x_i)_{i\in I}$ d'éléments de V) est un sous-espace vectoriel de V qui contient A (resp. qui contient tous les x_i).

Démonstration. Pour la partie ou la famille vide, on trouve le sous-espace trivial $\{0\}$. En général, cet ensemble est non vide, et il suffit d'appliquer l'équation (1) de la page 166 et la condition (i) de la proposition pour voir que c'est un sous-espace vectoriel. Il est clair qu'il contient A (resp. tous les x_i). ■

Proposition 4. Tout sous-espace vectoriel est stable par combinaisons linéaires.

Démonstration. Autrement dit, toute combinaison linéaire d'éléments de W est élément de W. En effet, chaque $\lambda_i x_i$ est élément de W (stabilité pour la loi externe) et toute somme d'éléments d'un sous-groupe d'un groupe abélien est élément de de ce sous-groupe. ■

Exemples.

1. Soient $x, y \in V$. Alors $Kx + Ky := \{\lambda x + \mu y \mid \lambda, \mu \in K\}$ est un sous-espace vectoriel de V et l'ensemble $Kx := \{\lambda x \mid \lambda \in K\}$ des multiples de x est un sous-espace vectoriel de $Kx + Ky$, donc de V.
2. Soit $J \subset I$. Le sous-groupe de K^I formé des $(x_i)_{i\in I}$ tels que $\forall j \in J\ ,\ x_j = 0$ est un sous-espace vectoriel de K^I. Par exemple, les sous-groupes $\{0\} \times K$ et $K \times \{0\}$ de l'espace vectoriel K^2 en sont des sous-espaces vectoriels.

Exercice 3.
Les parties suivantes sont-elles des sous-espaces vectoriels de K^2 :

$$A := \{(x, y) \in K^2 \mid y = x\}, \quad B := \{(x, y) \in K^2 \mid y = x^2\}, \quad C := \{(x, y) \in K^2 \mid y \geqslant x\}\ ?$$

Solution. Le sous-ensemble A est un sous-espace vectoriel (appelé la diagonale). Le sous-ensemble B est l'ensemble des couples (x, x^2). C'est un sous-espace vectoriel si, et seulement si, pour tous $x, y \in K$, on a : $(\lambda x + \mu y)^2 = \lambda x^2 + \mu y^2$. Le lecteur vérifiera que cela est vrai pour le corps $K = \mathbb{F}_2$ (corps à deux éléments) et pour lui-seul. Enfin, le sous-ensemble C ne peut être défini que si K a été muni d'une relation d'ordre. Prenons donc le cas où K est un sous-corps de $\mathbb{R}$ (c'est le seul exemple à notre disposition). Il est clair que $(0, 1)$ est élément de C mais pas son opposé ; C n'est donc pas un sous-espace vectoriel.

Remarque. La notion de sous-espace vectoriel permet parfois de munir un ensemble de la structure d'espace vectoriel sans avoir à en vérifier tous les axiomes. Ainsi, on a déjà vu que l'ensemble $\mathcal{F}(I, \mathbb{R})$ des fonctions numériques sur l'intervalle I de $\mathbb{R}$ est un $\mathbb{R}$-espace vectoriel. Comme une combinaison linéaire de fonctions continues, est continue, le sous-ensemble $\mathcal{C}(I, \mathbb{R})$ des fonctions continues est un sous-espace vectoriel de $\mathcal{F}(I, \mathbb{R})$, ce qui en fait un espace vectoriel sur $\mathbb{R}$; il en va de même pour l'ensemble $\mathcal{C}^\infty(I, \mathbb{R})$ des fonctions indéfiniment dérivables sur un intervalle ouvert I.

Lemme 5. Soit $(W_i)_{i\in I}$ une famille de sous-espaces vectoriels de l'espace vectoriel V. Alors leur intersection $W := \bigcap_{i\in I} W_i$ est un sous-espace vectoriel de V.

Démonstration. L'ensemble W est non vide, car il contient 0. Soient $\lambda, \mu \in K$ et $x, y \in W$. Alors, pour tout $i \in I$, $x, y \in W_i$, donc $\lambda x + \mu y \in W_i$.
On en déduit que $\lambda x + \mu y \in \bigcap_{i\in I} W_i$. ∎

Théorème et définition 6. (i) Soit A une partie de l'espace vectoriel V. Il existe alors un plus petit sous-espace vectoriel de V contenant A. On l'appelle *sous-espace vectoriel engendré par* A et on le note $\mathrm{Vect}(A)$. Le sous-espace vectoriel engendré par l'ensemble $\{x_i \mid i \in I\}$ est appelé *sous-espace vectoriel engendré par les* x_i, $i \in I$ et noté $\mathrm{Vect}\big((x_i)_{i\in I}\big)$.
(ii) Le sous-espace vectoriel de V engendré par une partie A (resp. par une famille $(x_i)_{i\in I}$) est l'ensemble des combinaisons linéaires d'éléments de A (resp. l'ensemble des combinaisons linéaires des x_i).

Démonstration. (i) Il suffit en effet de prendre l'intersection de tous les sous-espaces vectoriels de V contenant A.
(ii) Pour la famille vide et pour l'ensemble vide, on trouve l'espace vectoriel trivial. Nous ferons la démonstration pour une partie A non vide, le cas d'une famille s'en déduit immédiatement. L'ensemble W des combinaisons linéaires d'éléments de A est, d'après le corollaire 3 de la page précédente, un sous-espace vectoriel de V qui contient A, donc $\mathrm{Vect}(A) \subset W$. Par ailleurs, toute combinaison linéaire d'éléments de A appartient à $\mathrm{Vect}(A)$ d'après la proposition 4 de la page précédente, donc $W \subset \mathrm{Vect}(A)$. ∎
Par exemple, le sous-espace vectoriel engendré par la partie vide (ou par la famille vide) est le sous-espace vectoriel trivial $\mathrm{Vect}(\varnothing) = \{0\}$. En effet, tous les sous-espaces vectoriels contiennent la partie vide (ou la famille vide) et $\{0\}$ est le plus petit d'entre eux.
Exemples.

1. Le sous-espace vectoriel de V engendré par $x \in V$ est $\mathrm{Vect}(x) = Kx$.
2. Le sous-espace vectoriel de V engendré par $x, y \in V$ est $\mathrm{Vect}(x, y) = Kx + Ky$.
3. Le sous-espace vectoriel de $K^{\mathbb{N}}$ engendré par la famille des $\delta^{(n)}$ (voir l'exemple 4 de la page 166) est l'ensemble $K^{(\mathbb{N})}$ des suites à support fini.
4. Plus généralement, appelons, pour tout $i \in I$, $\delta^{(i)}$ l'élément de K^I dont la composante d'indice i vaut 1 et dont toutes les autres composantes valent 0. Alors, le sous-espace vectoriel de K^I engendré par la famille des $\delta^{(i)}$ est l'ensemble $K^{(I)}$ des familles à support fini.

Exemple. Quel est le nombre minimum de vecteurs nécessaire pour engendrer $W = Kx \subset V$? Ce nombre est majoré par 1 puisque x convient ! Si ce nombre est 0, c'est que W est engendré par la partie vide, c'est-à-dire $W = \{0\}$. Cela équivaut évidemment à $x = 0$ (et dans ce cas, le nombre minimum est bien 0). Sinon, ce nombre est 1. De plus, tout vecteur non nul y engendre W. En effet, $y = \lambda x$ (puisque $W = Kx$), $\lambda \neq 0$ (puisque $y \neq 0$), donc tout élément μx de $W = Kx$ s'écrit $(\mu\lambda^{-1})y$.

Exercice 4.
Quel est le nombre minimum de vecteurs nécessaire pour engendrer $W = Kx+Ky \subset V$?
Solution. Ce nombre est majoré par 2 puisque x et y conviennent ! Si ce nombre est 0, c'est que W est engendré par la partie vide, c'est-à-dire $W = \{0\}$. Cela équivaut à $x = y = 0$, et, dans ce cas, le nombre minimum est 0. Supposons donc x, y non tous deux nuls. Si l'un des deux est multiple de l'autre (et en particulier si l'un des deux est nul), ce nombre est 1 : si par exemple $y = \lambda x$, on voit que $W = Kx + Ky = Kx$. Enfin, si aucun des deux n'est multiple de l'autre, ce nombre est 2. En effet, si ce n'était pas le cas, on aurait $W = Kz$ pour un certain $z \in V$, d'où $x = \lambda z$ et $y = \mu z$. Supposons par exemple $x \neq 0$: alors $\lambda \neq 0$ et $y = \dfrac{\mu}{\lambda} x$, contradiction.

Définition 5. On appelle *droite vectorielle* un espace vectoriel non trivial engendré par un seul élément. On appelle *plan vectoriel* un espace vectoriel engendré par deux éléments et qui n'est pas une droite vectorielle.

Tout élément non nul d'une droite vectorielle l'engendre. Deux éléments quelconques d'une droite vectorielle sont proportionnels. Enfin, on verra plus loin que deux éléments non proportionnels d'un plan vectoriel l'engendrent.

Proposition 7. Soient $W_1, \ldots, W_p$ des sous-espaces vectoriels de l'espace vectoriel V. Le plus petit sous-espace de V contenant tous les W_i est :

$$\mathrm{Vect}\Big(\bigcup_{i=1}^{p} W_i\Big) = \big\{x_1 + \cdots + x_p \mid x_1 \in W_1, \ldots, x_p \in W_p\big\}.$$

Démonstration. Les sous-espaces contenant tous les W_i sont ceux qui contiennent la réunion A de ceux-ci, donc le plus petit est bien $\mathrm{Vect}(A)$. Toute somme $\sum\limits_{i=1}^{p} x_i$ de la forme indiquée est dans $\mathrm{Vect}(A)$. L'ensemble de ces sommes est non vide, stable par addition et multiplication externe, donc un sous-espace vectoriel contenant A, donc égal à $\mathrm{Vect}(A)$. ∎

Définition 6. Le sous-espace $\mathrm{Vect}\big(\bigcup\limits_{i=1}^{p} W_i\big)$ est appelé *somme des sous-espaces* W_i et noté $W_1 + \cdots + W_p = \sum\limits_{i=1}^{p} W_i$.

Exemples.

1. La somme des sous-espaces Kx et Ky de V est le sous-espace $Kx + Ky$.
2. La somme des sous-espaces $Kx_i, 1 \leqslant i \leqslant p$ est $\mathrm{Vect}((x_i)_{1 \leqslant i \leqslant p})$, que l'on pourra donc écrire $\sum\limits_{i=1}^{p} Kx_i = Kx_1 + \cdots + Kx_p$.

Exercice 5.
Soient W et W' deux sous-espaces de V. Leur réunion est-elle un sous-espace de V ?
Solution. Si par exemple $W \subset W'$, leur réunion est W' qui est un sous-espace. Si aucun des deux n'est inclus dans l'autre, leur réunion ne peut être un sous-espace.
Soient en effet $x \in W \setminus W'$ et $x' \in W' \setminus W$. Alors x et x' sont éléments de $W \cup W'$. Si l'on avait $x + x' \in W$, on en tirerait $x' = (x + x') - x \in W$, ce qui n'est pas ; donc $x + x' \notin W$. De même, $x + x' \notin W'$. Finalement, $x + x' \notin W \cup W'$, qui n'est donc pas un sous-espace vectoriel. Cependant, $W \cup W'$ est stable par multiplication externe.

2 Applications linéaires

La notion d'application linéaire est au moins aussi importante que celle d'espace vectoriel. C'est le phénomène de linéarité qui est fondamental, et l'introduction des espaces vectoriels vise essentiellement à donner un cadre rigoureux pour l'étudier. *Tous* les modules d'algèbre qui suivent (et certains modules d'analyse) en illustreront l'importance.

2.1 Vocabulaire et exemples

Définition 7.
Soient V et V' deux espaces vectoriels sur K. Une *application linéaire* de V dans V' est une application $f : V \to V'$ qui est un morphisme de groupes et telle que :

$$\forall \lambda \in K \ , \ \forall x \in V \ , \ f(\lambda x) = \lambda f(x).$$

L'espace vectoriel V (resp. l'espace vectoriel V') est appelé *source* ou *espace de départ* (resp. *but* ou *espace d'arrivée*) de l'application f. Si l'on veut préciser le corps de base, on pourra dire que f est K-linéaire.

Proposition 8. Soit f une application de l'espace vectoriel V dans l'espace vectoriel V'. Alors f est linéaire si, et seulement si, l'une quelconque des deux conditions équivalentes suivantes est vérifiée :

(i) $\forall \lambda, \mu \in K \ , \ \forall x, y \in V \ , \ f(\lambda x + \mu y) = \lambda f(x) + \mu f(y)$.

(ii) $\forall \lambda \in K \ , \ \forall x, y \in V \ , \ f(\lambda x + y) = \lambda f(x) + f(y)$.

La preuve de cette proposition est mécanique et laissée au lecteur. Nous omettrons également la preuve du fait suivant : les applications linéaires préservent les combinaisons linéaires. Autrement dit, on a les égalités :

$$f\left(\sum_{i \in I} \lambda_i x_i\right) = \sum_{i \in I} \lambda_i f(x_i),$$
$$f\left(\lambda_1 x_1 + \cdots + \lambda_p x_p\right) = \lambda_1 f(x_1) + \cdots + \lambda_p f(x_p).$$

Définition 8. L'application $x \mapsto \lambda x$ du K-espace vectoriel V dans lui-même est appelée *homothétie de rapport* $\lambda \in K$ *de l'espace vectoriel* V et notée H_λ.

L'homothétie H_λ est une application linéaire de V dans lui-même. En prenant $\lambda = 1$, on trouve l'identité de V. En prenant $\lambda = 0$, on trouve l'application constante de V dans V' de valeur $0_{V'}$ (élément neutre de V'), que l'on l'appelle *application (linéaire) nulle* ou *triviale* de V dans V'. Pour certains usages on exige $\lambda \neq 0$.

Exemples.

1. Soient D et D' deux droites distinctes passant par l'origine du plan $\Pi = \mathbb{R}^2$. Alors la projection sur la droite D parallèlement à la droite D' est une application linéaire : c'est en effet le *théorème de Thalès.* Si D est l'axe des abscisses et D' celui des ordonnées, cette projection est l'application $(x, y) \mapsto (x, 0)$. En revanche, la restriction de cette application au cercle $\Gamma := \{(\cos t, \sin t) \mid t \in \mathbb{R}\}$ n'est pas linéaire, car son ensemble de départ Γ n'est pas un espace vectoriel.
2. Si W est un sous-espace vectoriel de V, l'application qui, à l'élément x de W, associe l'élément x de V est linéaire. C'est en effet équivalent à dire que les lois de composition interne et externe sur W ont été définies par restriction de celles de V.
3. Soit f une application linéaire de l'espace vectoriel V dans l'espace vectoriel V'. Soient W un sous-espace vectoriel de V et W' un sous-espace vectoriel de V', et supposons que $f(W) \subset W'$. Alors l'application de W dans W' induite par *restriction* de f à W (et *corestriction* à W') est linéaire.
4. L'application qui, à la fonction $f \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$, associe sa dérivée f' est linéaire de $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ dans lui-même. L'application qui, à $f \in \mathcal{C}([0, 1], \mathbb{R})$, associe son intégrale $\displaystyle\int_0^1 f(t)\,\mathrm{d}t$, est linéaire de $\mathcal{C}([0, 1], \mathbb{R})$ dans $\mathbb{R}$.

Il faut parfois préciser le corps de base. Par exemple, l'application $z \mapsto \operatorname{Re} z$ de $\mathbb{C}$ dans $\mathbb{C}$ est $\mathbb{R}$-linéaire, mais pas $\mathbb{C}$-linéaire.

Exercice 6.

Donner un exemple d'application linéaire du $\mathbb{F}_2$-espace vectoriel $\mathbb{F}_2^2$ dans le $\mathbb{Q}$-espace vectoriel $\mathbb{Q}^2$.

Solution. Il n'y en a aucune, car ce ne sont pas des espaces vectoriels sur le même corps, et la condition de compatibilité avec la loi externe n'aurait ici aucun sens. Remarquons cependant que l'application constante de valeur $(0, 0) \in \mathbb{Q}^2$ est un morphisme de groupes de $\mathbb{F}_2^2$ dans $\mathbb{Q}^2$ (c'est d'ailleurs le seul : sauriez-vous le démontrer ?), mais que l'on ne peut pas écrire la relation $f(\lambda x) = \lambda f(x)$ car on ne saurait pas où prendre λ.

Exercice 7.

Décrire toutes les applications linéaires de K dans V.

Solution. Pour tout $\lambda \in K$, on a $f(\lambda) = f(\lambda.1) = \lambda f(1)$. Notant $x_0 = f(1) \in V$, on voit que f est de la forme $\lambda \mapsto \lambda x_0$. Réciproquement, pour tout $x_0 \in V$, l'application $\lambda \mapsto \lambda x_0$ convient. Il y a donc une correspondance bijective entre les applications linéaires de K dans V et les éléments de V.

Exercice 8.

Décrire toutes les applications linéaires de K^2 dans un espace vectoriel V. En déduire les applications linéaires de K^2 lui-même.

Solution. Notons $x_0 := f((1,0)) \in V$ et $y_0 := f((0,1)) \in V$.
De l'égalité $(\lambda, \mu) = \lambda.(1,0) + \mu.(0,1)$, on tire $f((\lambda,\mu)) = \lambda x_0 + \mu y_0$. Réciproquement, pour tout $(x_0, y_0) \in V \times V = V^2$, l'application $(\lambda,\mu) \mapsto \lambda x_0 + \mu y_0$ convient. Il y a donc une correspondance bijective entre les applications linéaires de K^2 dans V et les éléments de V^2. Dans le cas où $V = K^2$, l'application est déterminée par $x_0 = (a,c)$ et $y_0 = (b,d)$, donc de la forme $(\lambda,\mu) \mapsto (a\lambda + b\mu, c\lambda + d\mu)$.

Soient $f : V \to V'$ et $f' : V' \to V''$ des applications linéaires : cela sous-entend implicitement que V, V' et V'' sont des espaces vectoriels sur un même corps K. L'application composée $f' \circ f$ va de V dans V''.

Proposition 9. L'application composée $f' \circ f$ est linéaire.

Démonstration. On sait déjà (module II.2, p. 123, après la définition 7) que c'est un morphisme de groupes. Soient $x \in V$ et $\lambda \in K$. Alors, d'après les hypothèses :

$$(f' \circ f)(\lambda x) = f'\big(f(\lambda x)\big) = f'\big(\lambda f(x)\big) = \lambda f'\big(f(x)\big) = \lambda (f' \circ f)(x). \qquad \blacksquare$$

Exemple. Soit $\lambda \in K$. Notons H_λ et H'_λ les homothéties de rapport λ de V et de V'. Alors, si $f : V \to V'$ est linéaire, $H'_\lambda \circ f = f \circ H_\lambda$. Cette application (qui sera notée plus loin λf) associe à tout x de V l'élément $\lambda f(x)$ de V'. Lorsque $\lambda = 0$, notant temporairement $0_{V,V'}$ l'application nulle de V dans V', etc. on obtient la relation : $0_{V',V'} \circ f = f \circ 0_{V,V} = 0_{V,V'}$. Lorsque $\lambda = 1$, notant Id_V et $\mathrm{Id}_{V'}$ les applications identités de V et V', on obtient la relation : $\mathrm{Id}_{V'} \circ f = f \circ \mathrm{Id}_V = f$. Lorsque f est elle-même une homothétie, on trouve : $H_\lambda \circ H_{\lambda'} = H_{\lambda\lambda'}$.

Définition 9. Un *isomorphisme* de l'espace vectoriel V sur l'espace vectoriel V' est une application linéaire bijective de V sur V'. On dit que V et V' sont *isomorphes* s'il existe un isomorphisme de V sur V'.

L'application identité Id_V est donc un isomorphisme de V sur V.

Proposition 10. Si f est un isomorphisme de V sur V', alors sa réciproque f^{-1} est un isomorphisme de V' sur V. Si f est un isomorphisme de V sur V' et si f' est un isomorphisme de V' sur V'', alors $f' \circ f$ est un isomorphisme de V sur V''.
La relation d'isomorphie entre espaces vectoriels sur un même corps est donc réflexive, symétrique et transitive.

Démonstration. Pour prouver la première affirmation, déjà connue pour les isomorphismes de groupes (module II.2, théorème 20 de la page 124), prenons $x' \in V'$ et $\lambda \in K$. Soit $x := f^{-1}(x')$ l'antécédent de x' par f : donc $x' = f(x)$. Comme f est linéaire, $f(\lambda x) = \lambda x'$, d'où $f^{-1}(\lambda x') = \lambda x = \lambda f^{-1}(x')$: l'application f^{-1} est bien linéaire bijective, donc un isomorphisme. Enfin, $f' \circ f$ est linéaire bijective donc un isomorphisme. $\blacksquare$

Exemples.

1. Deux espaces vectoriels triviaux *sur un même corps* K sont isomorphes entre eux. En effet, l'unique application de l'un dans l'autre est l'application nulle, qui est linéaire et bijective. Réciproquement, si un espace vectoriel est isomorphe à un espace vectoriel trivial, il est lui même trivial. On parlera donc de *l'*espace vectoriel trivial, ou nul (avec l'article défini) sur un corps K fixé.
2. Soit V une droite vectorielle sur K et soit $x \in V$ un élément non nul. L'application $\lambda \mapsto \lambda x$ est un isomorphisme de K sur V. Toutes les droites vectorielles sont donc isomorphes à K, donc entre elles. De plus, tout espace vectoriel isomorphe à une droite vectorielle est une droite vectorielle.
3. Si le plan vectoriel V est engendré par x et y, l'application $(\lambda, \mu) \mapsto \lambda x + \mu y$ est un isomorphisme de K^2 sur V : il est évident qu'elle est linéaire, elle est surjective parce que x, y engendrent V et elle est injective parce que x, y sont non multiples l'un de l'autre. Les plans vectoriels sont donc exactement les espaces vectoriels isomorphes à K^2.
4. L'application $(\lambda_1, \ldots, \lambda_n) \mapsto \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}$ est un isomorphisme $K^n \simeq M_{n,1}(K)$.
5. On démontrera à la section 3 que les espaces vectoriels K et K^n ne sont isomorphes que si $n = 1$. On démontrera plus généralement au module II.7 que les espaces vectoriels K^m et K^n ne sont isomorphes que si $m = n$.

Définition 10. On appelle *endomorphisme* de l'espace vectoriel V une application linéaire de V dans lui-même. On appelle *automorphisme* de V un endomorphisme bijectif (autrement dit, un isomorphisme de V sur lui-même).

Le composé de deux endomorphismes de V est donc un endomorphisme de V (proposition 9 de la page ci-contre) et l'application réciproque d'un automorphisme de V est un automorphisme de V (proposition 10 de la page précédente).

Exemples.

1. Toute homothétie H_λ est un endomorphisme de V. Si $\lambda \neq 0$, c'est un automorphisme d'inverse $H_{\lambda^{-1}}$.
2. Tous les endomorphismes d'une droite vectorielle sont des homothéties.
3. Tout endomorphisme de K^2 est de la forme $f : (x, y) \mapsto (ax + by, cx + dy)$, avec $a, b, c, d \in K$. On dit que f est *associé à la matrice* $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

Exemple. Soient f, f' les endomorphismes de K^2 respectivement associés aux matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ et $\begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$. Un calcul immédiat donne :

$$(f' \circ f)(x, y) = \big((a'a + b'c)x + (a'b + b'd)y, (c'a + d'c)x + (c'b + d'd)y\big) .$$

Donc l'endomorphisme $f' \circ f$ est associé à la matrice $\begin{pmatrix} a'a+b'c & a'b+b'd \\ c'a+d'c & c'b+d'd \end{pmatrix}$.

Pour savoir si f est bijective, il faut chercher si l'on peut inverser la relation $(x', y') = f(x, y)$, c'est-à-dire trouver un unique antécédent (x, y) à tout élément (x', y').

Cela revient à résoudre le système : $\begin{cases} ax+by = x' \\ cx+dy = y' \end{cases}$

On sait que ce système admet une unique solution pour tout choix des seconds membres $x', y' \in K$ si, et seulement si, $ad - bc \neq 0$. On a alors :

$$\begin{cases} x = \dfrac{d}{ad-bc}x' + \dfrac{-b}{ad-bc}y' \\ y = \dfrac{-c}{ad-bc}x' + \dfrac{a}{ad-bc}y' \end{cases}$$

Ainsi, f est un automorphisme si, et seulement si, $ad - bc \neq 0$, et f^{-1} est alors associé à la matrice : $\begin{pmatrix} \dfrac{d}{ad-bc} & \dfrac{-b}{ad-bc} \\ \dfrac{-c}{ad-bc} & \dfrac{a}{ad-bc} \end{pmatrix} = \dfrac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Il y a des opérations qui ont un sens avec les endomorphismes mais pas avec les autres applications linéaires. Par exemple, on ne peut calculer $f \circ f$ que si source et but de f sont identiques, donc pour un endomorphisme. De même, on ne peut comparer $f \circ g$ à $g \circ f$ que si f et g sont des endomorphismes d'un même espace vectoriel V. Dans ce cas, on a déjà vu que, quel que soit le scalaire $\lambda \in K$, on a $H_\lambda \circ f = f \circ H_\lambda$: on dit que les homothéties H_λ commutent avec tout endomorphisme f. Si V est un espace vectoriel trivial ou une droite vectorielle, tout endomorphisme est une homothétie. Dans ce cas, la composition des endomorphismes est commutative, mais ce n'est pas généralement le cas.

Exemple. Les endomorphismes f et f' de K^2 respectivement associés à $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ et à $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ ne commutent pas. En effet, $f' \circ f$ est associé à $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ (donc égal à f), alors que $f \circ f'$ est associé à $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ (donc nul).

Exercice 9.

Quels endomorphismes de K^2 commutent avec tous les endomorphismes ?

Solution. Il s'agit de choisir $a, b, c, d \in K$ tels que, pour tous $a', b', c', d' \in K$:

$$\begin{pmatrix} a'a+b'c & a'b+b'd \\ c'a+d'c & c'b+d'd \end{pmatrix} = \begin{pmatrix} aa'+bc' & ab'+bd' \\ ca'+dc' & cb'+dd' \end{pmatrix}.$$

En prenant successivement $(a', b', c', d') = (1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$ et $(0, 0, 0, 1)$, on trouve : $a = d$ et $b = c = 0$, autrement dit, $f = H_a$.

C'est un cas particulier du théorème 42 de la page 200.

2.2 Noyau et image

Soit $f : V \to V'$ une application linéaire entre K-espaces vectoriels. C'est donc en particulier un morphisme de groupes, et l'on a défini son noyau : $\operatorname{Ker} f := \{x \in V \mid f(x) = 0_{V'}\}$. Ce dernier est un sous-groupe de V. On a de plus démontré que l'application f est injective si, et seulement si, son noyau est le sous-groupe trivial : f est injective $\Leftrightarrow \operatorname{Ker} f = \{0_V\}$. À l'opposé, l'application f est triviale (nulle) si, et seulement si, le noyau est l'espace source tout entier : f est nulle $\Leftrightarrow \operatorname{Ker} f = V$. Ainsi, pour que f soit à la fois nulle et injective, il faut, et il suffit, que la source de f soit l'espace trivial $V = \{0\}$.
Remarquons enfin que, pour tout sous-espace vectoriel $W \subset V$, la restriction de f à W a pour noyau $\operatorname{Ker} f \cap W$.

Proposition 11. Le noyau $\operatorname{Ker} f$ de l'application linéaire f est un sous-espace vectoriel de l'espace vectoriel V.

Démonstration. Il suffit de vérifier que $\operatorname{Ker} f$ est stable pour la loi externe. Mais, si $\lambda \in K$ et $x \in \operatorname{Ker} f$: $f(\lambda x) = \lambda f(x) = \lambda.0_{V'} = 0_{V'}$, et l'on a bien $\lambda x \in \operatorname{Ker} f$. ■

Exemples.

1. Reprenons l'exemple du théorème de Thalès : soient D et D' deux droites passant par l'origine du plan $\Pi = \mathbb{R}^2$; ce sont donc des sous-espaces vectoriels de $\mathbb{R}^2$. Soit $p : \Pi \to D$ la projection sur D parallèlement à D'. Son noyau est D', elle n'est donc pas injective. Soit Δ une droite passant par l'origine (donc un sous-espace vectoriel) et soit $q = p_{|\Delta} : \Delta \to D$ la restriction à Δ de la projection p. Le noyau de q est $D' \cap \Delta$. Si Δ est parallèle à la direction de projection, autrement dit, si $\Delta = D'$, alors $\operatorname{Ker} q = \Delta \cap D' = \Delta$ et l'application q est nulle. Si Δ n'est pas parallèle à la direction de projection, autrement dit, si $\Delta \neq D'$, $\operatorname{Ker} q = \Delta \cap D' = \{0\}$ et l'application q est injective.
2. Soit $f : V \to V'$ une application linéaire dont la source V est une droite vectorielle. Les seuls sous-espaces de V sont V lui-même et le sous-espace trivial $\{0_V\}$. L'application f est donc nulle ou injective.
3. Soit f un endomorphisme de K^2 et soit $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ la matrice associée. Si $ad - bc \neq 0$, l'application f est bijective et son noyau est nul. Si a, b, c, d sont tous nuls, l'application f est nulle et son noyau est K^2. Le cas « intermédiaire » est celui où $ad - bc = 0$ et a, b, c, d ne sont pas tous nuls. Dans ce cas, les vecteurs $(d, -c)$ et $(b, -a)$ sont proportionnels, ils ne sont pas tous deux nuls et ils engendrent le noyau, qui est donc une droite vectorielle.
4. L'endomorphisme $D : f \mapsto f'$ du $\mathbb{R}$-espace vectoriel $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ admet pour noyau le sous-espace vectoriel des fonctions constantes. Celui-ci est engendré par la fonction constante 1, c'est donc une droite vectorielle. L'endomorphisme M qui, à la fonction f de $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ associe la fonction xf, c'est-à-dire la fonction $x \mapsto xf(x)$, est injectif. Soit en effet un élément f du noyau : on a alors $xf(x) = 0$ pour tout réel x, donc $f(x) = 0$ pour tout réel non nul x. Par continuité, f est l'application constante nulle : $f = 0$. Le noyau de l'endomorphisme M est donc $\{0\}$ et M est injectif.

Exercice 10.

Soient a, b, c des éléments du corps K.

Quel est le noyau de l'application $f : K^2 \to K$ définie par $(x, y) \mapsto ax + by + c$?

Solution. Si $c \neq 0$, l'application f n'est pas linéaire et la question n'a aucun sens. Nous supposerons donc que c est nul. Si $a = b = 0$, l'application f est nulle et son noyau est K^2. Si a et b ne sont pas tous deux nuls, on vérifie sans peine que $\operatorname{Ker} f$ est engendré par le vecteur non nul $(b, -a)$: c'est donc une droite vectorielle.

Exercice 11.

Y a-t-il des applications linéaires injectives $K^n \to K$?

Solution. Si $n = 0$, l'unique application de $K^n = \{0\}$ dans K est linéaire et elle est à la fois nulle et injective. Si $n = 1$, les applications linéaires de K^n dans K sont les homothéties $H_\lambda : x \mapsto \lambda x$; elles sont injectives si, et seulement si, $\lambda \neq 0$. Supposons $n \geqslant 2$. Toute application linéaire f de K^n dans K est de la forme $(x_1, \ldots, x_n) \mapsto a_1x_1 + \cdots + a_nx_n$, les a_i étant des éléments de K fixés. Si les a_i sont tous nuls, f est nulle et n'est donc pas injective. Si, par exemple, a_1 n'est pas nul, le vecteur $(a_2, -a_1, 0, \ldots, 0)$ n'est pas nul et appartient au noyau ; si c'est une autre composante a_i qui n'est pas nulle, le lecteur construira de même un élément non nul du noyau. On conclut que, pour $n \geqslant 2$, il n'y a pas d'application linéaire injective de K^n dans K. En particulier, les espaces K^n et K ne sont pas isomorphes.

Soit $f : V \to V'$ une application linéaire entre K-espaces vectoriels. C'est donc en particulier un morphisme de groupes, et l'on a défini son image :

$$\operatorname{Im} f = f(V) := \{x' \in V' \mid \exists x \in V \ : \ x' = f(x)\}.$$

C'est un sous-groupe de V'. L'application f est surjective si, et seulement si, son image est le groupe V' tout entier : f est surjective $\Leftrightarrow \operatorname{Im} f = V'$.

À l'opposé, l'application f est triviale si, et seulement si, son image est le sous-groupe trivial de V' : f est nulle $\Leftrightarrow \operatorname{Im} f = \{0_{V'}\}$. Ainsi, pour que l'application f soit à la fois triviale et surjective, il faut, et il suffit, que son but soit le groupe trivial $V' = \{0\}$.

Proposition 12. L'image $\operatorname{Im} f$ de l'application linéaire f est un sous-espace vectoriel de l'espace vectoriel V'.

Démonstration. Il suffit de vérifier que $\operatorname{Im} f$ est stable pour la loi externe. Mais, si $\lambda \in K$ et $x' \in \operatorname{Im} f$, on peut écrire $x' = f(x)$ avec $x \in V$ et $\lambda x' = \lambda f(x) = f(\lambda x)$; on a donc bien $\lambda x' \in \operatorname{Im} f$. ∎

Exemples.

1. Dans l'exemple du théorème de Thalès, avec les notations précédentes, l'image de la projection $p : \Pi \to D$ est D. L'application p est donc surjective. Si $\Delta = D'$, la restriction q est l'application linéaire nulle et $\operatorname{Im} q = \{0\}$. Si $\Delta \neq D'$, l'image de q est $\operatorname{Im} q = D$ et l'application q est surjective.

2. Soit $f : V \to V'$ une application linéaire dont le but V' est une droite vectorielle. Les seuls sous-espaces de V' sont V' lui-même et le sous-espace trivial $\{0_{V'}\}$.
   L'application f est donc nulle ou surjective.
3. Soit f un endomorphisme de K^2 de matrice associée $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$.
   Si $ad - bc \neq 0$, l'application f est bijective et son image est K^2. Si a, b, c, d sont tous nuls, l'application f est nulle et son image est $\{0\}$. Si $ad - bc = 0$ et a, b, c, d ne sont pas tous nuls, les vecteurs (a, c) et (b, d) sont proportionnels, ils ne sont pas tous deux nuls et ils engendrent l'image, qui est donc une droite vectorielle.
4. L'endomorphisme $D : f \mapsto f'$ du $\mathbb{R}$-espace vectoriel $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ admet pour image l'espace vectoriel tout entier, car il est surjectif ; par exemple, un antécédent de $f \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ est la primitive de f qui s'annule en 0, c'est-à-dire la fonction $x \mapsto \displaystyle\int_0^x f(t)\,\mathrm{d}t$.
   L'endomorphisme $M : f \mapsto xf$ de $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ a pour image le sous-espace vectoriel de $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ formé des fonctions qui s'annulent en 0. En effet, il est clair que la fonction xf s'annule en 0. Réciproquement, si la fonction $g \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ s'annule en 0, la fonction $x \mapsto g(x)/x$ qui est *a priori* définie sur $\mathbb{R}^*$ et de classe $\mathcal{C}^\infty$ se prolonge par continuité en une fonction f telle que $f(0) = \lim\limits_{x \to 0} g(x)/x = g'(0)$ et l'on a $xf = g$. On peut alors démontrer (mais ce n'est nullement évident !) que $f \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$, et f est un antécédent de g par M.

Exercice 12.

Y a-t-il des applications linéaires surjectives $K \to K^n$?

Solution. Si $n = 0$, l'unique application de K dans $K^n = \{0\}$ est linéaire et elle est à la fois nulle et surjective. Si $n = 1$, les applications linéaires de K dans K^n sont les homothéties $H_\lambda : x \mapsto \lambda x$; elles sont surjectives si, et seulement si, $\lambda \neq 0$. Supposons $n \geqslant 2$. Toute application linéaire f de K dans K^n est de la forme $x \mapsto (xa_1, \ldots, xa_n)$, où $f(1) = (a_1, \ldots, a_n) \in K^n$ est fixé. Si les a_i sont tous nuls, f est nulle et n'est donc pas surjective. Si, par exemple, a_1 n'est pas nul, le vecteur $(0, 1, \ldots, 1)$ n'appartient pas à l'image ; si c'est une autre composante a_i qui n'est pas nulle, le lecteur construira de même un élément de K^n sans antécédent. Dans tous les cas, on conclut que, pour $n \geqslant 2$, il n'y a pas d'application linéaire surjective de K sur K^n. En particulier, on a redémontré que les espaces K^n et K ne sont pas isomorphes.

Remarque. Deux faits significatifs se déduisent des exemples précédents et seront généralisés au module II.7 : un endomorphisme de K ou de K^2 est injectif si, et seulement s'il est bijectif (on démontrera que ce fait est valable pour tout endomorphisme de K^n, $n \in \mathbb{N}$) ; l'espace K^n est isomorphe à $\{0\} = K^0$ ou à $K = K^1$ si, et seulement si, $n = 0$ ou 1 (on démontrera que K^n et K^p sont isomorphes si, et seulement si, $n = p$).

Soit $f : V \to V'$ une application linéaire. Pour tout sous–espace vectoriel $W' \subset V'$, on vérifie facilement que l'image réciproque par f de W' :

$$f^{-1}(W') := \{x \in V \mid f(x) \in W'\}$$

est un sous–espace vectoriel de V. De même, pour tout sous–espace vectoriel $W \subset V$, son image par f : $f(W) := \{x' \in V' \mid \exists x \in W \ : \ f(x) = x'\}$ est un sous–espace vectoriel de V'.
Remarquons qu'en prenant $W' = \{0_{V'}\}$ et $W = V$, on retrouve le noyau $\operatorname{Ker} f = f^{-1}\big(\{0\}\big)$ et l'image $\operatorname{Im} f = f(W)$.

2.3 Quelques applications linéaires particulières

Définition 11.
Un endomorphisme de V idempotent pour la loi $\circ$ est appelé *projecteur*.

Rappelons qu'un élément a est dit idempotent pour une loi de composition interne $\star$ si $a \star a = a$. Un projecteur est donc un endomorphisme p tel que $p \circ p = p$, c'est-à-dire que, pour tout x de V : $p\big(p(x)\big) = p(x)$.

Exemples. Les endomorphismes $p : (x, y) \mapsto (x, 0)$ et $q : (x, y) \mapsto (0, y)$ de $\mathbb{R}^2$ sont des projecteurs. Pour que l'homothétie H_λ soit un projecteur, il faut, et il suffit, que $H_{\lambda^2} = H_\lambda$, donc $\lambda^2 = \lambda$. Les seules possibilités sont l'endomorphisme nul H_0 et l'endomorphisme identité H_1.

Proposition 13. L'endomorphisme p de V est un projecteur d'image $W \subset V$ si, et seulement si, $\operatorname{Im} p \subset W$ et $p_{|W} = \mathrm{Id}_W$.

Démonstration. Si p est un projecteur d'image W, il est évident que $\operatorname{Im} p \subset W$. Tout élément x de W s'écrit $x = p(y)$ et l'on a alors : $p(x) = p\big(p(y)\big) = p(y) = x$, de sorte que la restriction de p à W est bien l'identité de W.
Réciproquement, soit p un endomorphisme de V d'image $\operatorname{Im} p \subset W$ et tel que $p_{|W} = \mathrm{Id}_W$. Alors, pour tout x de V, l'élément $y = p(x)$ appartient à $\operatorname{Im} p$, donc à W.
Ainsi $p(y) = p_{|W}(y) = \mathrm{Id}_W(y) = y$, c'est-à-dire que $p\big(p(x)\big) = p(x)$: l'endomorphisme p est bien un projecteur. Par hypothèse, son image est contenue dans W ; mais tout élément y de W est égal à $\mathrm{Id}_W(y) = p_{|W}(y) = p(y) \in \operatorname{Im} p$, et l'image de p est $\operatorname{Im} p = W$. ∎

Pratiquement, pour prouver que p est un projecteur d'image W, on vérifiera les deux conditions suivantes :

$$\forall x \in V \ , \ p(x) \in W \qquad \text{et} \qquad \forall y \in W \ , \ p(y) = y. \tag{4}$$

Proposition 14. Soit p un projecteur de V. L'application $q : x \mapsto x - p(x)$ est un projecteur de V d'image $\operatorname{Im} q = \operatorname{Ker} p$.

Démonstration. On vérifie aisément que c'est une application linéaire, donc un endomorphisme de V. En fait, on définira en 2.4 l'addition des applications linéaires ; selon cette

définition, on a $q = \mathrm{Id}_V - p$.
Nous allons vérifier les conditions (4). Notons $W = \operatorname{Ker} p$. Soit $x \in V$. Alors :

$$p\big(q(x)\big) = p\big(x - p(x)\big) = p(x) - p\big(p(x)\big) = 0,$$

et l'on a bien $q(x) \in \operatorname{Ker} p = W$. Soit maintenant $y \in W$, donc $p(y) = 0$. Alors $q(y) = y - p(y) = y$, d'où la deuxième condition, ce qui permet de conclure. ■

Puisque $x - q(x) = x - \big(x - p(x)\big) = p(x)$, on peut d'ailleurs appliquer la deuxième proposition et en tirer que $\operatorname{Im} p = \operatorname{Ker} q$.

Proposition 15. Si p est un projecteur de V, alors :

$$V = \operatorname{Im} p + \operatorname{Ker} p \qquad \text{et} \qquad \operatorname{Im} p \cap \operatorname{Ker} p = \{0\}.$$

Démonstration. La première affirmation vient de l'égalité évidente : $x = p(x) + q(x)$ et de la proposition précédente.
Soit $x \in \operatorname{Im} p \cap \operatorname{Ker} p$. Alors $x = p(x)$, car $x \in \operatorname{Im} p$ et $p(x) = 0$, car $x \in \operatorname{Ker} p$. ■

Comme nous le verrons à la section 4, on dit alors que V est la somme directe de $\operatorname{Im} p$ et $\operatorname{Ker} p$. Nous montrerons également que toute écriture de V comme somme directe de deux sous-espaces vectoriels s'obtient de cette manière.

Exemple.
Les projecteurs $p : (x, y) \mapsto (x, 0)$ et $q : (x, y) \mapsto (0, y)$ de $\mathbb{R}^2$ vérifient $q = \mathrm{Id}_{\mathbb{R}^2} - p$.
On a $\operatorname{Im} p = \operatorname{Ker} q = \mathbb{R} \times \{0\}$ (la droite horizontale) et $\operatorname{Im} q = \operatorname{Ker} p = \{0\} \times \mathbb{R}$ (la droite verticale).

Définition 12. Une *forme linéaire* sur le K-espace vectoriel V est une application linéaire de V dans K. Un *hyperplan* (ou *hyperplan vectoriel*) de V est le noyau d'une forme linéaire non nulle sur V.

On a déjà vu qu'une forme linéaire f sur V est soit nulle soit surjective. Dans le premier cas, le noyau $\operatorname{Ker} p$ est bien sûr l'espace V tout entier.

Exemples.

1. L'espace trivial ne contient aucun hyperplan. Pour $n \geqslant 1$, les hyperplans de K^n sont les sous-espaces de la forme :

$$H := \{(x_1, \ldots, x_n) \in K^n \mid a_1 x_1 + \cdots + a_n x_n = 0\},$$

   les coefficients a_i étant non tous nuls. Le seul hyperplan d'une droite vectorielle est le sous-espace trivial. Dans K^2, les hyperplans sont les droites vectorielles. Un hyperplan de l'espace $\mathbb{R}^3$ est un plan vectoriel (*i.e.* passant par l'origine). Il est défini par une équation : $ax + by + cz = 0$, dont les coefficients a, b, c ne sont pas tous nuls.
2. L'ensemble des fonctions numériques continues sur le segment $[0, 1]$ et nulles en 0 est un hyperplan de $\mathcal{C}([0, 1], \mathbb{R})$: c'est en effet le noyau de la forme linéaire non nulle $f \mapsto f(0)$.

Nous allons illustrer sur une même construction les notions de projecteur et de forme linéaire. On se donne un hyperplan H de V et l'on veut construire un projecteur d'image H. On introduit d'abord une forme linéaire non nulle f sur V dont le noyau est $\operatorname{Ker} f = H$: il en existe, par définition des hyperplans. Puisque f est non nulle, on peut choisir un élément $x_0 \in V$ tel que $f(x_0) \neq 0$. Notons D le sous-espace vectoriel engendré par x_0 : c'est donc une droite vectorielle de V. On pose enfin $p(x) = \dfrac{f(x)}{f(x_0)} x_0$. Le lecteur consciencieux vérifiera (en tout cas, il le devrait) que p est une application linéaire, donc un endomorphisme de V. Montrons que c'est un projecteur d'image D, en utilisant les deux conditions pratiques indiquées plus haut. Tout d'abord, pour tout x de V, on a $p(x) \in D$ puisque c'est un multiple de x_0. Soit ensuite $x \in D$; on peut donc écrire $x = \lambda x_0$ et calculer :

$$p(x) = p(\lambda x_0) = \lambda p(x_0) = \lambda \frac{f(x_0)}{f(x_0)} x_0 = \lambda x_0 = x.$$

Cela démontre notre assertion. Le noyau de p est H, car $\dfrac{f(x)}{f(x_0)} x_0$ s'annule si, et seulement si, $f(x) = 0$. Si l'on pose : $q(x) = x - \dfrac{f(x)}{f(x_0)} x_0$, il découle de l'étude précédente que q est un projecteur d'image H et de noyau D.

2.3.1 Produit d'espaces vectoriels

Soient $V_1, \ldots, V_k$ des groupes. Le groupe produit : $\prod_{i=1}^{k} V_i = V_1 \times \cdots \times V_k$ a été défini en munissant l'ensemble produit $V_1 \times \cdots \times V_k$ de la loi interne :

$$(x_1, \ldots, x_k) + (y_1, \ldots, y_k) := (x_1 + y_1, \ldots, x_k + y_k).$$

Définition 13. Soient $V_1, \ldots, V_k$ des K-espaces vectoriels. L'*espace vectoriel produit* est défini en munissant le groupe produit $V_1 \times \cdots \times V_k$ de la loi externe :

$$\lambda(x_1, \ldots, x_k) := (\lambda x_1, \ldots, \lambda x_k).$$

On définit, pour chaque indice j, une application $j^{ème}$ projection p_j de $\prod_{i=1}^{k} V_i$ dans V_j, qui associe à l'élément $(x_1, \ldots, x_k)$ sa $j^{ème}$ composante x_j. On définit également, pour chaque indice j, une application $j^{ème}$ section $s_j : V_j \to \prod_{i=1}^{k} V_i$, qui associe à l'élément $a \in V_j$ l'élément $(0, \ldots, a, \ldots, 0)$ dont la $j^{ème}$ composante est a et dont toutes les autres composantes sont nulles. Le lecteur vérifiera que les applications projections et sections sont linéaires, que les p_j sont surjectives et que les s_j sont injectives. En particulier, le produit de deux K-espaces vectoriels V et V' est l'ensemble produit $V \times V'$, muni des opérations :

$$\forall (x, x'), (y, y') \in V \times V' \ , \ (x, x') + (y, y') := (x + y, x' + y'),$$
$$\forall \lambda \in K \ , \ \forall (x, x') \in V \times V' \ , \ \lambda(x, x') := (\lambda x, \lambda x').$$

Les projections sur V et V' sont respectivement les applications linéaires $(x, x') \mapsto x$ et $(x, x') \mapsto x'$. Les sections sont l'application $a \mapsto (a, 0_{V'})$ de V dans $V \times V'$ et l'application $a' \mapsto (0_V, a')$ de V' dans $V \times V'$. Si par exemple on prend pour $V_1, \ldots, V_k$ le K-espace vectoriel K, l'espace vectoriel produit $V_1 \times \cdots \times V_k$ n'est autre que l'espace K^k. L'application $j^{ème}$ projection p_j associe à l'élément $(x_1, \ldots, x_k) \in K^k$ sa $j^{ème}$ composante $x_j \in K$.

2.4 Espaces d'applications linéaires

Définition 14. Soient V et V' deux K-espaces vectoriels. On note $\mathcal{L}_K(V,V')$ (ou $\mathcal{L}(V,V')$, s'il n'y a pas de risque de confusion) l'ensemble des applications K-linéaires de V dans V'.
L'ensemble $\mathcal{L}_K(V,V)$ des endomorphismes de V est noté $\mathcal{L}_K(V)$ (ou $\mathcal{L}(V)$).
Le sous-ensemble de $\mathcal{L}_K(V)$ formé des automorphismes est noté $\mathcal{GL}_K(V)$ ou $\mathcal{GL}(V)$.
On utilise également les notations $\mathrm{Hom}_K(V,V')$, $\mathrm{End}_K(V)$ et $\mathrm{Aut}_K(V)$ pour ces trois ensembles.

Soient maintenant f et g des applications linéaires de V dans V' et $\alpha \in K$ un scalaire. On peut alors définir les applications $f+g$ et αf de V dans V' : il s'agit là d'opérations dans le K-espace vectoriel $\mathcal{F}(V,V')$.

Exemples. L'homothétie H_λ de V est égale à $\lambda\,\mathrm{Id}_V$. Pour tout projecteur p de V, l'application $\mathrm{Id}_V - p$ est un projecteur de V.

Lemme 16. Les applications $f+g$ et αf sont linéaires.

Démonstration. Cela résultera des deux calculs suivants :

$$\begin{aligned}(f+g)(\lambda x+\mu y) &= f(\lambda x+\mu y)+g(\lambda x+\mu y)\\ &= \big(\lambda f(x)+\mu f(y)\big)+\big(\lambda g(x)+\mu g(y)\big)\\ &= \lambda(f+g)(x)+\mu(f+g)(y), \text{ et}\end{aligned}$$

$$(\alpha f)(\lambda x+\mu y) = \alpha f(\lambda x+\mu y) = \alpha\big(\lambda f(x)+\mu f(y)\big) = \lambda(\alpha f)(x)+\mu(\alpha f)(y),$$

d'où respectivement la linéarité de $f+g$, et celle de αf. ∎

Proposition 17. L'ensemble $\mathcal{L}_K(V,V')$ muni de ces deux opérations est un K-espace vectoriel.

Démonstration. On sait déjà que l'ensemble $\mathcal{F}(V,V')$ de toutes les applications de V dans V' est un K-espace vectoriel. Le lemme nous dit que $\mathcal{L}_K(V,V')$ en est un sous-espace vectoriel. ∎

Exemples.

1. De la description des applications linéaires de K dans V, on déduit la bijection : $\mathcal{L}_K(K,V) \to V$ qui, à $f \in \mathcal{L}_K(K,V)$, associe $f(1)$.
   De l'égalité $(\alpha f+\beta g)(1) = \alpha f(1)+\beta g(1)$, on déduit que cette bijection est linéaire. On a donc un isomorphisme $\mathcal{L}_K(K,V) \simeq V$.
2. De la description des applications linéaires de K^2 dans V, on déduit la bijection : $\mathcal{L}_K(K^2,V) \to V^2$ qui, à $f \in \mathcal{L}_K(K^2,V)$, associe $\big(f(1,0), f(0,1)\big)$. On démontre de même que c'est un isomorphisme. Il découlera de la section 3.3 que l'on a, plus généralement, $\mathcal{L}_K(K^n,V) \simeq V^n$.

Définition 15. L'espace vectoriel $\mathcal{L}_K(V, K)$ des formes linéaires sur l'espace vectoriel V est appelé *espace dual* ou simplement *dual* de V. On le note V^*.

On prendra garde aux risques de confusion avec d'autres usages de l'astérisque. Par exemple, il vaut mieux éviter de noter K^* le dual du K-espace vectoriel K, la signification $K^* = K \setminus \{0\}$ étant d'usage plus courant.

Exemple. En faisant $V = K$ dans les exemples précédents, on voit que le dual de K^n s'identifie naturellement à K^n. Par exemple, si $n = 1$, $\lambda \in K$ correspond à la forme linéaire $x \mapsto \lambda x$ de K dans K. Si $n = 2$, $(\lambda, \mu) \in K^2$ correspond à la forme linéaire $(x, y) \mapsto \lambda x + \mu y$ de K^2 dans K. Ces exemples sont d'ailleurs exceptionnels : il n'y pas de façon naturelle d'identifier un espace vectoriel quelconque avec son dual.

Nous allons énoncer deux règles de distributivité de la composition par rapport aux combinaisons linéaires. Soient V, V', V'' trois espaces vectoriels sur K, $f, g \in \mathcal{L}_K(V, V')$ et $f', g' \in L_K(V', V'')$ des applications linéaires et $\lambda, \mu \in K$ des scalaires. Alors :

$$(\lambda f' + \mu g') \circ f = \lambda(f' \circ f) + \mu(g' \circ f),$$

$$f' \circ (\lambda f + \mu g) = \lambda(f' \circ f) + \mu(f' \circ g).$$

Remarquons d'ailleurs que la première de ces deux règles n'exige en fait aucune hypothèse sur les applications f, f', g', alors que la deuxième règle repose sur la linéarité de f'.

Proposition 18. Muni de l'addition et de la composition, $\mathcal{L}_K(V)$ est un anneau.

Démonstration. On sait déjà que c'est un groupe pour l'addition. Si l'on se restreint à l'espace vectoriel $\mathcal{L}_K(V)$ des endomorphismes de V, la composition est une loi interne. Elle est associative et Id_V en est élément neutre. Elle est enfin distributive par rapport à l'addition d'après les règles ci-dessus. ∎

Corollaire 19. L'ensemble $\mathcal{GL}_K(V)$ (définition 14 de la page précédente) muni de la composition est un groupe.

Démonstration. C'est en effet le groupe des éléments inversibles de l'anneau $\mathcal{L}_K(V)$. ∎

Définition 16. Le groupe $\mathcal{GL}_K(V)$ est appelé *groupe linéaire* de l'espace vectoriel V.

Exemples.

1. Si V est trivial, $\mathcal{L}_K(V)$ est l'anneau trivial.
2. Si V est une droite vectorielle, les endomorphismes de V sont les homothéties, et l'anneau $\mathcal{L}_K(V)$ est isomorphe au corps K et le groupe linéaire $\mathcal{GL}_K(V)$ est isomorphe au groupe multiplicatif K^*.
3. Si $V = K^2$, l'anneau $\mathcal{L}_K(V)$ est isomorphe à l'anneau des matrices carrées $M_2(K)$ et le groupe linéaire $\mathcal{GL}_K(V)$ est isomorphe au groupe $GL_2(K)$ des matrices inversibles (module II.4).

3 Familles de vecteurs

On sait, depuis Descartes, que repérer les points du plan par leurs coordonnées permet d'appliquer à la géométrie toute la puissance de l'algèbre et de l'analyse. Ce repérage suppose le choix d'une origine et de deux vecteurs non proportionnels. La question du choix de l'origine sera l'un des objets de la géométrie affine ; ce qui compte, pour la géométrie *vectorielle*, c'est le choix d'une *base*, c'est-à-dire d'une famille de vecteurs *génératrice* et *libre*.

3.1 Familles génératrices

Définition 17. Une famille $(x_i)_{i\in I}$ d'éléments du K-espace vectoriel V est dite *génératrice* si tout élément de V est combinaison linéaire des x_i :

$$\forall x \in V \ , \ \exists (\lambda_i)_{i\in I} \in K^{(I)} \ : \ x = \sum_{i\in I} \lambda_i x_i.$$

Autrement dit (section 1.3), V est engendré par la famille $(x_i)_{i\in I}$: $V = \mathrm{Vect}(x_i)_{i\in I}$. Une *partie génératrice* de V est une partie A telle que la famille de ses éléments est génératrice : $V = \mathrm{Vect}(A)$.

Exemple. Dans l'espace vectoriel trivial $\{0\}$, toute famille est génératrice. La partie vide est génératrice. Dans l'espace vectoriel K, une famille est génératrice si, et seulement si, l'un de ses éléments au moins est non nul. Cela reste vrai dans toute droite vectorielle.

Même si l'on ne mentionne pas toujours explicitement le corps de base, la définition d'une partie génératrice en dépend de manière essentielle. Par exemple, la partie $\{1\} \subset \mathbb{R}$ est une partie génératrice du $\mathbb{R}$-espace vectoriel $\mathbb{R}$, mais pas du $\mathbb{Q}$-espace vectoriel $\mathbb{R}$: le sous-espace du $\mathbb{Q}$-espace vectoriel $\mathbb{R}$ engendré par $\{1\}$ est $\mathbb{Q}$.

Le lecteur prendra garde que, dans l'écriture de $x \in V$ sous la forme $x = \sum \lambda_i x_i$, les coefficients λ_i ne sont en général pas uniques.

Exercice 13.
Montrer que la famille $\big((1,0),(0,1),(1,1)\big)$ d'éléments de l'espace vectoriel K^2 est génératrice et déterminer toutes les écritures de $(a,b) \in K^2$ comme combinaison linéaire des éléments de cette famille.
Solution. On cherche tous les triplets (x,y,z) d'éléments de K tels que :

$$(a,b) = x(1,0) + y(0,1) + z(1,1),$$

ce qui revient à résoudre le système de 2 équations à 3 inconnues :

$$\begin{cases} x + z = a \\ y + z = b \end{cases}$$

On peut, par exemple, résoudre en fonction de z et obtenir $(x,y,z) = (a-z, b-z, z)$, $z \in K$ étant arbitraire. Cela fournit l'écriture : $(a,b) = (a-z)(1,0)+(b-z)(0,1)+z(1,1)$. On ne peut donc sans ambiguïté désigner « les » coefficients de l'écriture de (a,b) comme combinaison linéaire des éléments de cette famille.

Cet exercice suggère l'existence d'un lien entre familles génératrices et systèmes d'équations. Considérons le système de n équations à p inconnues :

$$\begin{cases} a_{1,1}x_1 + \cdots + a_{1,p}x_p = b_1, \\ \vdots \\ a_{n,1}x_1 + \cdots + a_{n,p}x_p = b_n. \end{cases} \tag{5}$$

La conjonction de ces n égalités scalaires équivaut à l'égalité vectorielle :

$$x_1 \begin{pmatrix} a_{1,1} \\ \vdots \\ a_{n,1} \end{pmatrix} + \cdots + x_p \begin{pmatrix} a_{1,p} \\ \vdots \\ a_{n,p} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}.$$

Cette égalité a lieu dans l'espace $M_{n,1}(K)$ des vecteurs colonnes.

Notons $A_1 = \begin{pmatrix} a_{1,1} \\ \vdots \\ a_{n,1} \end{pmatrix}, \ldots, A_p = \begin{pmatrix} a_{1,p} \\ \vdots \\ a_{n,p} \end{pmatrix}$ et $B = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$.

Pour que le système (5) admette au moins une solution, il faut, et il suffit, que B soit dans $\mathrm{Vect}(A_1, \ldots, A_p)$. Pour qu'il admette une solution *pour tout second membre* B, il faut, et il suffit, que la famille $(A_1, \ldots, A_p)$ soit une famille génératrice de $M_{n,1}(K)$.

Voici maintenant des exemples de familles génératrices infinies. Une particularité de ces exemples est que, dans chaque cas, il n'existe pas de famille génératrice finie.

Exemples. Dans le K-espace vectoriel $K[X]$, la famille $(X^n)_{n\in\mathbb{N}}$ est génératrice. Dans le $\mathbb{R}$-espace vectoriel $\mathbb{R}^{(\mathbb{N})}$ des suites à support fini, la famille des $\delta^{(n)}$ (exemple 4 de la page 166). est génératrice. Soit maintenant $(u^{(1)}, \ldots, u^{(n)})$ une famille finie d'éléments de $\mathbb{R}^{(\mathbb{N})}$. Chacune des suites $u^{(i)}$ est, par hypothèse, à support fini : il existe donc un entier N_i tel que tous les termes de la suite $u^{(i)}$ d'indices strictement supérieurs à N_i sont nuls. Notons N le plus grand des entiers N_i. Soit maintenant $u = \lambda_1 u^{(1)} + \cdots + \lambda_n u^{(n)}$ un élément de $\mathrm{Vect}(u^{(1)}, \ldots, u^{(n)})$: tous ses termes d'indices strictement supérieurs à N sont donc nuls. Comme les éléments de $\mathbb{R}^{(\mathbb{N})}$ ne vérifient pas tous cette propriété (par exemple $\delta^{(N+1)}$), le sous-espace vectoriel $\mathrm{Vect}(u^{(1)}, \ldots, u^{(n)})$ n'est pas $\mathbb{R}^{(\mathbb{N})}$ et la famille $(u^{(1)}, \ldots, u^{(n)})$ n'est pas génératrice. Ainsi, aucune famille finie de $\mathbb{R}^{(\mathbb{N})}$ n'est génératrice.

Exercice 14.

Montrer que le K-espace vectoriel $K[X]$ n'admet aucune famille génératrice finie.

Solution. Soit $(P_1, \ldots, P_k)$ une famille finie de polynômes et soit d le maximum de leurs degrés. Toute combinaison linéaire des P_i est de degré inférieur ou égal à d.
La famille $(P_1, \ldots, P_k)$ n'est donc pas génératrice.

La proposition qui suit est formulée pour des parties génératrices : l'adaptation au cas d'une

famille ne pose pas de problème.

Proposition 20. Toute partie de V contenant une partie génératrice est elle-même une partie génératrice. L'image par une application linéaire surjective d'une partie génératrice est une partie génératrice.

Démonstration. Soit A une partie génératrice du K-espace vectoriel V et soit B une partie de V contenant A. Si $x \in V$ est une combinaison linéaire d'éléments de A, c'est aussi une combinaison linéaire d'éléments de B : il suffit en effet d'attribuer aux éléments de A les mêmes coefficients et à ceux de $B \setminus A$ le coefficient 0. Autrement dit, partant de l'écriture $x = \sum_{a \in A} \lambda_a a$, on en tire l'écriture $x = \sum_{b \in B} \mu_b b$, en posant $\mu_a := \lambda_a$ pour $a \in A$ et $\mu_b := 0$ pour $b \notin A$. Cela montre bien que B est une partie génératrice. Soit par ailleurs f une application linéaire surjective de V sur V'. Tout élément x' de V' s'écrit $x' = f(x)$ avec $x \in V$ (surjectivité de f). Mais x est combinaison linéaire des $a \in A$ (qui est génératrice) donc $f(x)$ est combinaison linéaire des $f(a), a \in A$ (linéarité de f).
Avec les notations précédentes : $x' = \sum_{a \in A} \lambda_a f(a) = \sum_{a' \in A'} \lambda'_{a'} a'$, où l'on a noté $A' := f(A)$ et, pour $a' \in A'$: $\lambda'_{a'} := \sum_{f(a)=a'} \lambda_a$. Cette somme a bien un sens, parce que tous les λ_a sont nuls sauf un nombre fini ; et l'on en déduit d'ailleurs que tous les $\lambda'_{a'}$ sont nuls sauf un nombre fini. Ainsi, $A' = f(A)$ est bien une partie génératrice de V'. ■

3.2 Familles libres

Définition 18. La famille $(x_i)_{i \in I}$ d'éléments de V est dite *libre* s'il n'existe aucune relation linéaire non triviale entre les x_i :

$$\forall (\lambda_i)_{i \in I} \in K^{(I)} \ : \ \sum_{i \in I} \lambda_i x_i = 0 \Longrightarrow \forall i \in I \ , \ \lambda_i = 0.$$

On dit également que les vecteurs x_i sont *linéairement indépendants* .
Une partie de V est dite libre si la famille de ses éléments l'est.

La notion de famille libre dépend du corps de base. Par exemple, la partie $\{1, \sqrt{2}\}$ de $\mathbb{R}$ est-elle libre ? Dans le $\mathbb{R}$-espace vectoriel $\mathbb{R}$, aucune partie à 2 éléments n'est libre. Dans le $\mathbb{Q}$-espace vectoriel $\mathbb{R}$, cette partie est bien libre, car aucun de ces deux éléments n'est multiple de l'autre *au sens de la multiplication externe* : il faudrait en effet que le facteur soit un rationnel.

Des vecteurs linéairement indépendants ne vérifient aucune relation linéaire non triviale. Une famille ou une partie qui contient le vecteur nul 0_V ne peut donc être libre (considérer la combinaison linéaire qui affecte le coefficient 1 au vecteur 0_V et le coefficient $0 \in K$ à tous les autres vecteurs). Ainsi, même si la partie A est libre, la partie $A \cup \{0_V\}$ ne l'est jamais. À l'opposé, la partie A est génératrice si, et seulement si,

la partie $A \cup \{0_V\}$ l'est. S'il y a répétition dans une famille $(x_i)_{i\in I}$, elle ne peut être libre. Supposons en effet que $x_j = x_k$ avec deux indices distincts $j, k \in I$. On pose alors (par exemple) $\lambda_j = 1$, $\lambda_k = -1$ et $\lambda_i = 0$ pour $i \neq j, k$, d'où une relation linéaire non triviale. Ainsi, si la partie $\{x, y\}$ est libre, la famille (x, y) l'est également mais pas la famille (x, x, y). Rappelons d'ailleurs (module I.1) que la famille associée à un ensemble ne comporte, par définition, pas de répétition.

La définition d'une famille génératrice $(x_i)_{i\in I}$ affirme *l'existence*, pour tout $x \in V$, d'une combinaison linéaire des x_i égale à x ; mais elle n'en garantit pas *l'unicité*. La définition d'une famille libre garantit au contraire cette unicité (mais, bien sûr, pas l'existence). Supposons en effet la famille libre. Si l'on a deux écritures : $x = \sum\limits_{i\in I} \lambda_i x_i = \sum\limits_{i\in I} \mu_i x_i$, on en tire la relation linéaire $\sum\limits_{i\in I}(\lambda_i - \mu_i)x_i = 0$, d'où, par hypothèse : $\forall i \in I\ ,\ \lambda_i - \mu_i = 0$, et $\forall i \in I\ ,\ \lambda_i = \mu_i$. C'est bien l'unicité annoncée de l'écriture.

Exemples.

1. La seule partie libre de l'espace trivial $\{0\}$ est la partie vide $\varnothing$.

2. Les seules parties libres d'une droite vectorielle sont les $\{x\}$, où $x \neq 0$.

3. Dans tout espace vectoriel V, les singletons $\{x\}$, où $x \neq 0$ sont des parties libres. Le couple (famille de deux éléments) (x, y) est libre si, et seulement si, $x \neq y$ et si la paire (partie à deux éléments) $\{x, y\}$ l'est. La paire $\{x, y\}$ est une partie libre si, et seulement si, aucun des deux vecteurs x,y n'est multiple de l'autre. Pour le voir, prouvons que la paire $\{x, y\}$ est une partie *non* libre si, et seulement si, l'un des deux vecteurs x,y est multiple de l'autre. Si $y = \lambda x$, resp. $x = \mu y$, on a la relation linéaire non triviale $\lambda.x + (-1).y = 0_V$, resp. la relation linéaire non triviale $(-1).x + \mu.y = 0_V$.

   Réciproquement, si $\lambda x + \mu y = 0_V$ est une relation linéaire non triviale, λ ou μ est non nul. Si $\lambda \neq 0$, on a $x = \dfrac{-\mu}{\lambda} y$; si $\mu \neq 0$, on a $y = \dfrac{-\lambda}{\mu} x$.

4. Soient (a, b) et (c, d) deux éléments de K^2. Ils sont linéairement indépendants si, et seulement si, $ad - bc \neq 0$. Pour le voir, nous invoquons l'équivalence précédente. Si $(c, d) = \lambda(a, b)$ ou $(a, b) = \mu(c, d)$, il est immédiat que $ad - bc = 0$. Réciproquement, si $ad - bc = 0$, on a $d(a, b) = b(c, d)$ et $c(a, b) = a(c, d)$. Si l'un des quatre coefficients a, b, c, d est non nul, on a une relation linéaire non triviale ; si $a = b = c = d = 0$, les vecteurs ne sont évidemment pas indépendants.

5. Dans le K-espace vectoriel $K[X]$, la famille $(X^n)_{n\in\mathbb{N}}$ est libre.

La proposition qui suit s'étend sans peine au cas d'une famille.

Proposition 21. Soit A une partie libre du K-espace vectoriel V. Toute partie de A est libre. L'image de A par une application linéaire injective est libre.

Démonstration. Soit B une partie de A. De toute relation linéaire non triviale $\sum_{b\in B} \lambda_b b = 0$ entre les éléments de B, on déduirait une relation linéaire non triviale $\sum_{a\in A} \mu_a a = 0$ entre les éléments de A en posant $\mu_a := \lambda_a$ si $a \in B$, $\mu_a := 0$ si $a \in A \setminus B$. Cela démontre la première affirmation. Soit $f : V \to V'$ une application linéaire injective. Tout $a' \in A' = f(A)$ s'écrit de manière unique $a' = f(a)$, avec $a \in A$. Toute combinaison linéaire d'éléments de A' peut donc s'écrire : $\sum_{a'\in A'} \lambda'_{a'} a' = \sum_{a\in A} \lambda_a f(a) = f\left(\sum_{a\in A} \lambda_a a\right)$ (par linéarité de f), en posant $\lambda_a := \lambda'_{f(a)}$ pour tout $a \in A$. D'une relation linéaire $\sum_{a'\in A'} \lambda_{a'} a' = 0_{V'}$ entre les éléments de A', on déduit donc : $f\left(\sum_{a\in A} \lambda_a a\right) = 0_{V'}$, d'où $\sum_{a\in A} \lambda_a a = 0_V$ (par injectivité de f), puis, A étant libre, $\forall a \in A\ ,\ \lambda_a = 0$. Par définition, cela entraîne $\forall a \in A\ ,\ \lambda'_{f(a)} = 0$, puis $\forall a' \in A'\ ,\ \lambda'_{a'} = 0$, car tout $a' \in A'$ est un $f(a)$. ■

Exercice 15.

Notons, pour tout $k \in \mathbb{N}$, $u^{(k)}$ la suite de réels dont le $n^{ème}$ terme est $u_n^{(k)} = n^k$. Démontrer que la famille $(u^{(k)})_{k\in\mathbb{N}}$ est une famille libre du $\mathbb{R}$-espace vectoriel $\mathbb{R}^{\mathbb{N}}$.

Solution. Soit $u = a_1 u^{(k_1)} + \cdots + a_p u^{(k_p)}$ une combinaison linéaire non triviale u des $u^{(k)}$. On suppose les indices k_i deux à deux distincts et rangés en ordre croissant : $k_1 < \cdots < k_p$, et les coefficients a_i tous non nuls. Alors, lorsque $n \to \infty$, u_n est équivalent à son terme dominant $a_p n^{k_p}$, car $n^{k_i} = o(n^{k_p})$ pour $i < p$. La suite u n'est donc pas nulle.

La famille $(u^{(k)})_{k\in\mathbb{N}}$ est donc libre.

Exercice 16.

Pour tout $k \in \mathbb{N}$, soit f_k la fonction $x \mapsto x^k$. Démontrer que la famille $(f_k)_{k\in\mathbb{N}}$ est une famille libre du $\mathbb{R}$-espace vectoriel $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$. Le sous–espace vectoriel engendré par cette famille est l'espace vectoriel des *applications polynomiales* (module II.6).

Solution. Soit $\sum a_k f_k = 0$ une relation linéaire non triviale. Par définition, les coefficients a_k sont presque tous nuls et $\sum a_k X^k$ est un polynôme. Par hypothèse, ce polynôme s'annule sur $\mathbb{R}$ tout entier, donc il a une infinité de racines, donc il est nul, c'est-à-dire que tous les a_k sont nuls.

Comme la notion de famille génératrice, la notion de famille libre est liée à la résolution d'un système d'équations linéaires.

Exemple. Considérons le système de n équations à p inconnues :

$$\begin{cases} a_{1,1}x_1 + \cdots + a_{1,p}x_p = 0, \\ \vdots \\ a_{n,1}x_1 + \cdots + a_{n,p}x_p = 0. \end{cases} \tag{6}$$

La conjonction de ces n égalités scalaires équivaut à l'unique égalité vectorielle :

$$x_1 \begin{pmatrix} a_{1,1} \\ \vdots \\ a_{n,1} \end{pmatrix} + \cdots + x_p \begin{pmatrix} a_{1,p} \\ \vdots \\ a_{n,p} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Cette égalité est une relation linéaire dans l'espace $M_{n,1}(K)$. Le système (6) admet évidemment la solution triviale $x_1 = \cdots = x_p = 0$. Avec les notations antérieures, cela correspond à la relation linéaire triviale entre les vecteurs colonnes $A_1, \ldots, A_p$. Dire que le système n'admet que la solution triviale, c'est donc dire que les vecteurs $A_1, \ldots, A_p$ sont linéairement indépendants. Remarquons d'ailleurs que les solutions $(x_1, \ldots, x_p) \in K^p$ du système (6) forment, dans tous les cas, un sous-espace vectoriel de K^p : c'est le noyau de l'application linéaire $(x_1, \ldots, x_p) \mapsto x_1A_1 + \cdots + x_pA_p$ de K^p dans K^n. Dire que le système (6) n'admet que la solution triviale, c'est donc également dire que cette application linéaire est injective.

Définition 19. La famille $(x_i)_{i\in I}$ d'éléments de V est dite *liée* si elle n'est pas libre, autrement dit, s'il existe une relation linéaire non triviale entre les x_i. On dit également que les vecteurs x_i sont *linéairement dépendants*.
Une partie de V est dite liée si la famille de ses éléments l'est.

Toute partie contenant une partie liée est liée, de même que toute partie contenant le vecteur nul ou deux vecteurs proportionnels. La réciproque de ce dernier point est fausse : par exemple, dans K^2, les vecteurs $(1,0)$, $(0,1)$ et $(1,1)$ sont deux à deux indépendants mais liés dans leur ensemble.

Proposition 22. L'image d'une famille liée par toute application linéaire est liée.

Démonstration. L'implication : $\sum_{i\in I} \lambda_i x_i = 0 \Rightarrow \sum_{i\in I} \lambda_i f(x_i) = f\big(\sum_{i\in I} \lambda_i x_i\big) = 0$ montre que, de toute relation linéaire non triviale entre les x_i, on tire une relation linéaire non triviale entre les $f(x_i)$. ∎

Il en résulte qu'une famille dont l'image par une application linéaire *quelconque* est libre est elle-même libre. L'assertion correspondante pour une *partie* est fausse. Par exemple, soit f la forme linéaire $(x,y) \mapsto x$ sur $\mathbb{R}^2$ et soit $A := \{(1,y) \mid y \in \mathbb{R}\} \subset \mathbb{R}^2$. La partie A est liée mais $f(A) = \{1\}$ est une partie libre de $\mathbb{R}$.

Exemple. L'application qui, à une fonction numérique $f \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$, associe la suite $u \in \mathbb{R}^{\mathbb{N}}$ définie par $u_n = f(n)$ est linéaire. Ainsi, avec les notations de l'exercice 16 de la page précédente et de l'exemple qui le précède, si les f_k étaient liés, les $u^{(k)}$ le seraient. On en déduit une nouvelle solution du deuxième exercice, en supposant le premier résolu.

Proposition 23. Une famille (resp. une partie) est liée si, et seulement si, l'un de ses vecteurs est combinaison linéaire des autres.

Démonstration. Nous le démontrerons dans le cas d'une famille. Supposons donc la famille $(x_i)_{i\in I}$ liée. Il existe donc une relation linéaire non triviale : $\sum_{i\in I} \lambda_i x_i = 0$. L'un des coefficients de cette relation est non nul (sinon, ce serait la relation triviale) : soit $j \in I$ tel que $\lambda_j \neq 0$. On a alors : $x_j = \displaystyle\sum_{i\in I\setminus\{j\}} \frac{-\lambda_i}{\lambda_j} x_i$.

Réciproquement, supposons : $x_j = \sum_{i\in I\setminus\{j\}} \mu_i x_i$. Posons $\lambda_i = \mu_i$ pour $i \in I \setminus \{j\}$ et $\lambda_j = -1$. On a alors : $\sum_{i\in I} \lambda_i x_i = 0$. ■

Exercice 17.

Soient $a_1, \ldots, a_n$ des points deux à deux distincts de l'intervalle I de $\mathbb{R}$.
On note, pour $i = 1, \ldots, n$, f_i la fonction $x \mapsto |x - a_i|$ de $\mathcal{C}(I, \mathbb{R})$. Démontrer que les f_i sont linéairement indépendants.

Solution. S'ils sont liés, l'un d'eux est combinaison linéaire des autres. Quitte à renuméroter les a_i, on peut supposer que $|x - a_n| = \sum_{i=1}^{n-1} c_i\,|x - a_i|$. Mais le membre droit de cette égalité est une fonction dérivable en a_n et pas le membre gauche. En effet, la fonction numérique $x \mapsto |x - a|$ est dérivable en tout point sauf au point a. Voici un autre argument possible : dans un voisinage du point a_n, le membre droit de l'égalité est une fonction affine donc monotone, alors que le membre de gauche change de sens de variation.

Le critère ci-dessus est extrêmement utile, mais son maniement peut s'avérer dangereux *même dans le cas de deux vecteurs* ! Tout étudiant sait en effet que, si les vecteurs u et v sont liés, l'un est multiple de l'autre, et l'on en déduit souvent (mais bien à tort) que l'on peut écrire $v = \lambda u$. Il peut se produire que $u = 0_V$ et $v \neq 0_V$: les deux vecteurs sont alors liés mais v n'est pas multiple de u. En revanche $u = 0v$ est bien multiple de v. En fait, si u et v sont liés et tous deux non nuls, ou tous deux nuls, alors chacun est multiple de l'autre. Si l'un des deux est nul et pas l'autre, la situation n'est pas symétrique. Il sera donc souvent prudent d'accompagner le raisonnement ou le calcul d'une *discussion*. Soit par exemple x un vecteur non nul de V. Pour tout scalaire $t \in K$, notons $u_t = tx$ et $v_t = (1-t)x$. On a $tv_t + (t-1)u_t = 0_V$; comme t et $t-1$ ne sont pas tous deux nuls, ces deux vecteurs sont liés. Lequel est multiple de l'autre ? Si $t \notin \{0, 1\}$, les deux vecteurs sont non nuls et chacun est multiple de l'autre. Si $t = 0$, c'est u_t qui est nul, donc multiple de v_t. Si $t = 1$, la situation est symétrique.

3.3 Bases

Définition 20. Une *base* du K-espace vectoriel V est une famille libre et génératrice, autrement dit, une famille $(x_i)_{i\in I}$ telle que :

$$\forall x \in V \ , \ \exists ! \ (\lambda_i)_{i\in I} \in K^{(I)} \ : \ x = \sum_{i\in I} \lambda_i x_i.$$

Cette égalité est appelée *écriture de x dans la base $(x_i)_{i\in I}$*.
On appelle *coordonnées du vecteur $x \in V$ dans la base $(x_i)_{i\in I}$ de V* les coefficients λ_i de l'unique famille $(\lambda_i)_{i\in I} \in K^{(I)}$ telle que $x = \sum\limits_{i\in I} \lambda_i x_i$.
La famille $(\lambda_i)_{i\in I}$ des coordonnées de x dans la base $\mathcal{B}$ sera notée $(x)_{\mathcal{B}}$.
Une *partie basique* de V est une partie libre et génératrice.

Rappelons (module I.1) que le symbole $\exists!$ signifie « il existe un unique … », ici, une unique famille $(\lambda_i)_{i\in I}$ telle que … Selon cette définition, la base (x, y) et la base (y, x) sont distinctes, mais nous serons rarement aussi pointilleux. De même, on appellera souvent base une partie basique, par abus de langage simplificateur.

Exemples.

1. La seule base de l'espace vectoriel nul est la famille vide. Les bases du K-espace vectoriel K sont les familles formées d'un élément non nul.
2. Soit $((a, b), (c, d))$ une famille libre à deux éléments de K^2 : donc, $ad - bc \neq 0$.
   Si $(A, B) \in K^2$, la relation : $(A, B) = x(a, b) + y(c, d)$ équivaut au système :
   $$\begin{cases} ax + cy = A \\ bx + dy = B \end{cases} \qquad \text{donc à} \qquad x = \frac{Ad - Bc}{ad - bc} \ , \ y = \frac{aB - bA}{ad - bc} \ .$$
   Toute famille libre de deux éléments de K^2 en est donc une base. Il en découle que toute famille de plus de deux éléments est liée : en effet, si (a, b), (c, d) sont deux éléments d'une telle famille, la sous-famille $\big((a, b), (c, d)\big)$ est libre, donc c'est une base (on vient de le voir) donc les autres éléments de la famille sont des combinaisons linéaires de ces deux-là, ce qui contredit l'hypothèse que la famille est libre. Comme on a déjà vu qu'aucune famille de moins de 2 éléments n'est génératrice (elles engendrent $\{0\}$ ou une droite vectorielle), on conclut que *les bases de K^2 sont exactement les familles $\big((a, b), (c, d)\big)$ telles que $ad - bc \neq 0$*.
3. Le $\mathbb{R}$-espace vectoriel $\mathbb{C}$ admet pour base $\{1, \mathrm{i}\}$. Le $\mathbb{C}$-espace vectoriel $\mathbb{C}$ admet pour base $\{1\}$.
4. La famille des X^n est la *base canonique* de l'espace vectoriel $K[X]$.
5. Le système linéaire avec second membre (5) admet une solution unique pour tout second membre $B \in M_{n,1}(K)$ si, et seulement si, la famille $(A_1, \ldots, A_p)$ est une base de l'espace vectoriel $M_{n,1}(K)$.
6. Dans le $\mathbb{R}$-espace vectoriel $\mathbb{R}^{(\mathbb{N})}$ des suites *à support fini*, les suites $\delta^{(n)}$ (exemple 4 de la page 166) forment une base.

Remarque. Soit f un isomorphisme du K-espace vectoriel V sur le K-espace vectoriel V'. Alors l'image par f de toute base de V est une base de V'. On sait en effet que l'image d'une famille libre par une application linéaire injective est une famille libre et que l'image d'une famille génératrice par une application linéaire surjective est une famille génératrice. Les bases d'une droite vectorielle sont donc les familles formées d'un seul élément qui est non nul. En effet, toute droite vectorielle est isomorphe au K-espace vectoriel K, et cela a déjà été démontré pour K. De même, si V admet une base à deux éléments, il est isomorphe à K^2, c'est un plan vectoriel et toutes ses bases ont deux éléments. Les bases d'un plan vectoriel sont ses familles libres à deux éléments. En effet, cela a déjà été démontré pour K^2.

Pour $i = 1, \ldots, n$, on définit $e_i \in K^n$ comme le vecteur dont la $i^{ème}$ composante vaut 1 et les autres valent 0 ; avec la notation de Kronecker (voir exemple de la page 56) :

$$e_i := (\delta_{i,1}, \ldots, \delta_{i,n}).$$

Alors $(e_1, \ldots, e_n)$ est une base de K^n, appelée *base canonique*. L'unique écriture d'un élément de K^n dans la base canonique est :

$$(\lambda_1, \ldots, \lambda_n) = \lambda_1 e_1 + \cdots + \lambda_n e_n.$$

Ainsi, les coordonnées de $x = (\lambda_1, \ldots, \lambda_n) \in K^n$ dans la base canonique de K^n sont $(\lambda_1, \ldots, \lambda_n)$. Plus généralement la *base canonique* du K-espace vectoriel $K^{(I)}$ est la famille $(e_i)_{i \in I}$ définie comme suit : $\forall i \in I\ ,\ e_i = (\delta_{i,j})_{j \in I}$.
Les coordonnées de $(\lambda_i)_{i \in I} \in K^{(I)}$ dans la base canonique sont donc les λ_i.

Lorsque l'on change *un* vecteur de la base $\mathcal{B}$, en général, *toutes* les coordonnées sont modifiées.
Par exemple, les coordonnées de (x, y) dans la base canonique $\big((1,0),(0,1)\big)$ de K^2 sont x et y. Dans la base modifiée $\big((1,0),(1,-1)\big)$, les coordonnées sont x' et y' tels que :

$$(x, y) = x'\,(1,0) + y'\,(1,-1), \qquad \text{ce qui donne} \qquad x' = x + y \quad \text{et} \quad y' = -y.$$

Exercice 18.

Soient $n \geqslant 2$ et $a \neq 0$. Reprendre ce calcul avec

$$\mathcal{B} = (x_1, \ldots, x_n) \qquad \text{et} \qquad \mathcal{B}' = (x_1, \ldots, x_{n-1}, x_1 + \cdots + x_{n-1} + a x_n).$$

Solution. Les coordonnées $(\lambda_1, \ldots, \lambda_n)$ et $(\lambda'_1, \ldots, \lambda'_n)$ de x dans les bases $\mathcal{B}$ et $\mathcal{B}'$ sont reliées par les relations :

$$\lambda_1 x_1 + \cdots + \lambda_n x_n = \lambda'_1 x_1 + \cdots + \lambda'_{n-1} x_{n-1} + \lambda'_n (x_1 + \cdots + x_{n-1} + a x_n).$$

On en tire, par identification : $\lambda'_i = \lambda_i - a^{-1}\lambda_n$ pour $i \leqslant n-1$ et $\lambda'_n = a^{-1}\lambda_n$.

Théorème 24. Soit $(x_i)_{i\in I}$ une famille d'éléments de V.
(i) L'application $(\lambda_i)_{i\in I} \mapsto \sum_{i\in I} \lambda_i x_i$ de $K^{(I)}$ dans V est linéaire. Elle est injective (resp. surjective) si, et seulement si, la famille $(x_i)_{i\in I}$ est libre (resp. génératrice).
(ii) Si $(x_i)_{i\in I}$ est une base, on a ainsi un isomorphisme de $K^{(I)}$ sur V.
(iii) Si V admet une base de n éléments, il est isomorphe à K^n.

Démonstration. La linéarité traduit la règle de calcul (1) de la page 166. Les conditions de surjectivité et d'injectivité traduisent les caractérisations des familles génératrices et libres par existence et par unicité de l'écriture d'un vecteur quelconque comme combinaison linéaire de la famille. Les assertions (ii) et (iii) sont alors évidentes. ∎

L'isomorphisme $K^{(I)} \simeq V$ se traduit par les règles de calcul suivantes :

$$\begin{cases} (x)_{\mathcal{B}} = (\lambda_i)_{i\in I} \\ (y)_{\mathcal{B}} = (\mu_i)_{i\in I} \end{cases} \Longrightarrow \begin{cases} (x+y)_{\mathcal{B}} = (\lambda_i + \mu_i)_{i\in I} \\ (\lambda x)_{\mathcal{B}} = (\lambda\lambda_i)_{i\in I}. \end{cases}$$

En particulier, l'application $\pi_j : \sum_{i\in I} \lambda_i x_i \mapsto \lambda_j$ est une forme linéaire sur V.

Définition 21. Les formes linéaires π_j sont appelées *formes linéaires coordonnées* relatives à la base $\mathcal{B}$.

Dans le cas d'une base finie de n éléments $\mathcal{B} = (x_1, \ldots, x_n)$, il est de coutume de représenter $(x)_{\mathcal{B}}$ comme un vecteur *colonne* :

$$x = \lambda_1 x_1 + \cdots + \lambda_n x_n \Longleftrightarrow (x)_{\mathcal{B}} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \in M_{n,1}(K).$$

Dans ce cas, les règles de calcul sont les suivantes :

$$(x+y)_{\mathcal{B}} = (x)_{\mathcal{B}} + (y)_{\mathcal{B}} \text{ et } (\lambda x)_{\mathcal{B}} = \lambda\,(x)_{\mathcal{B}}.$$

Proposition et définition 25. Soit V un K-espace vectoriel. On suppose qu'il admet une base *finie* $\mathcal{B} = (x_i)_{i\in[\![1,n]\!]}$. Alors $\mathcal{B}^* := (\pi_i)_{i\in[\![1,n]\!]}$ (où les π_i sont les formes linéaires coordonnées) est une base du dual $E^* := \mathcal{L}(E, \mathbb{K})$.
On l'appelle *base duale* de la base $\mathcal{B}$.

Démonstration. Soit $\varphi \in E^*$. Pour tout $x = \lambda_1 x_1 + \cdots + \lambda_n x_n \in E$, on a :

$$\varphi(x) = \lambda_1\varphi(x_1) + \cdots + \lambda_n\varphi(x_n) = \varphi(x_1)\pi_1(x) + \cdots + \varphi(x_n)\pi_n(x).$$

Donc $\varphi = \varphi(x_1)\pi_1 + \cdots + \varphi(x_n)\pi_n$, ce qui prouve que la famille $(\pi_i)_{i\in[\![1,n]\!]}$ est *génératrice*.
Soient $\mu_1, \ldots, \mu_n \in \mathbb{K}$, tels que $\mu_1\pi_1 + \cdots + \mu_n\pi_n = 0$. Pour tout $j \in [\![1,n]\!]$, on a $0 = \mu_1\pi_1(x_j) + \cdots + \mu_n\pi_n(x_j) = \mu_j$. La famille $(\pi_i)_{i\in[\![1,n]\!]})$ est donc *libre*. ∎

Théorème 26. Soit V un K-espace vectoriel et soit $\mathcal{B} = (x_i)_{i\in I}$ une base de V. Soit V' un K-espace vectoriel et soit $(x'_i)_{i\in I}$ une famille quelconque dans V' (de même ensemble d'indices). Il existe alors une unique application linéaire f de V dans V' telle que, pour tout indice $i \in I$, on ait $f(x_i) = x'_i$. *Une application linéaire est uniquement déterminée par la connaissance de son effet sur une base de son espace de départ.*

Démonstration. En effet, si f est une telle application, et x un élément de V, de l'écriture $x = \sum_{i\in I} \lambda_i x_i$ de x dans la base $\mathcal{B}$, on tire, par linéarité, $f(x) = \sum_{i\in I} \lambda_i x'_i$, d'où l'unicité de f. Réciproquement, $\mathcal{B}$ étant une base, l'écriture $x = \sum_{i\in I} \lambda_i x_i$ de $x \in V$ est unique, et l'on peut poser : $f(x) = \sum_{i\in I} \lambda_i x'_i$. On définit ainsi une application de V dans V', dont le lecteur vérifiera qu'elle est linéaire et telle que $f(x_i) = x'_i$ pour tout indice $i \in I$. ■

Exemple. L'unique forme linéaire telle que : $\forall i \in I\ ,\ x_i \mapsto \delta_{i,j}$ est π_j.

3.3.1 Existence de bases

Nous allons maintenant démontrer l'existence de bases, *en admettant le point le plus délicat (c'est le lemme 29)*. Au module II.7, le cas le plus important (dimension finie) sera traité en détail.

Proposition 27. Soit B une base de V. Soient A et C des parties de V telles que $A \subset B \subset C$ (on sait donc que A est libre et que C est génératrice). Alors, si $A \neq B$, la partie A n'est pas génératrice, donc n'est pas une base. De même, si $C \neq B$, la partie C n'est pas libre, donc n'est pas une base. En d'autres termes, toute base est à la fois une *partie génératrice minimale* et une *partie libre maximale*.

Démonstration. Soit $x \in B \setminus A$. Si x était combinaison linéaire d'éléments de A, on obtiendrait une relation linéaire non triviale entre les éléments de B, qui ne serait donc pas libre. Soit $y \in C\setminus B$. C'est une combinaison linéaire d'éléments de B, ce qui fournit une relation linéaire non triviale entre des éléments de C, qui n'est donc pas libre. ■

Proposition 28. Toute partie libre maximale parmi les parties libres contenues dans une partie génératrice donnée C est une base.

Démonstration. Soit B une partie libre maximale parmi celles contenues dans C. On veut prouver que tout élément x de V est combinaison linéaire d'éléments de B. Comme C est génératrice, il suffit de le prouver pour un élément quelconque x de C. Si $x \in B$, c'est évident. Sinon, $B' = B \cup \{x\}$ contient strictement B, et est contenu dans C, donc n'est pas libre. Dans toute relation linéaire non triviale entre éléments de B', le coefficient de x est non nul (sinon, on en déduirait une relation linéaire non triviale entre éléments de B). On peut donc bien en tirer x comme combinaison linéaire d'éléments de B. ■

Lemme 29. Soient A une partie libre et C une partie génératrice de V, telles que $A \subset C$. Il existe alors une partie libre maximale B telle que $A \subset B \subset C$.

Nous admettrons ce lemme. Voir cependant une démonstration à l'exercice II.3.23 de la page 206.

Théorème 30. Soient A une partie libre et C une partie génératrice de V, telles que $A \subset C$. Il existe alors une base B de V telle que $A \subset B \subset C$.
En conséquence, tout espace vectoriel admet une base.

Démonstration. La première assertion résulte du lemme et de la proposition ci-dessus.
Pour la seconde, il suffit de prendre $A = \varnothing$ et $C = V$. ■

3.4 Dimension finie

Nous énonçons ici quelques résultats qui nous seront utiles aux modules suivants, mais dont la démonstration appartient au module II.7. Le lecteur attentif vérifiera qu'aucun cercle vicieux n'est ainsi créé.

Un espace vectoriel sera dit *de dimension finie* s'il admet une famille génératrice finie. On prouve alors que *dans un espace V de dimension finie, toutes les bases sont finies et ont même nombre d'éléments.* Ce nombre est appelé *dimension* de l'espace vectoriel V. L'espace K^n est donc de dimension n (voir sa base canonique). L'espace V est isomorphe à K^n si, et seulement s'il est de dimension n. Les espaces K^n et K^p ne sont isomorphes que si $n = p$.

4 Sommes directes et projections

4.1 Somme directe de deux sous-espaces vectoriels

Lemme 31. Soient E et F deux sous-espaces vectoriels du K-espace vectoriel V. L'application $(y, z) \mapsto y + z$ de $E \times F$ dans V est linéaire et son image est la somme $E + F$. Son noyau est isomorphe à $E \cap F$.

Démonstration. L'application considérée est la somme des deux projections $E \times F \to E$ et $E \times F \to F$, vues comme éléments de $\mathcal{L}_K(E \times F, V)$. L'assertion sur son image est évidente. Les éléments de son noyau sont les couples (y, z) tels que $y \in E$, $z \in F$ et $z = -y$: ce noyau est donc le sous-espace $\{(y, -y) \mid y \in E \cap F\}$ de $E \times F$. L'application $y \mapsto (y, -y)$ de $E \cap F$ dans $E \times F$ est un isomorphisme sur ce sous-espace. ■

Proposition 32.
Avec les notations du lemme, les trois affirmations suivantes sont équivalentes :

1. L'espace V est la somme $E + F$ et $E \cap F = \{0\}$.
2. L'application $(y, z) \mapsto y + z$ de $E \times F$ dans V est un isomorphisme.
3. Pour tout $x \in V$, il existe un unique couple $(y, z) \in E \times F$ tel que $x = y + z$.

Démonstration. L'équivalence des deux premières propriétés est une conséquence directe du lemme. La troisième propriété est une paraphrase de la deuxième. ■

Définition 22. Si les trois affirmations de la proposition sont vérifiées, on dit que V est la *somme directe de* E *et de* F et l'on écrit $V = E \oplus F$.

Bien sûr, dans ce cas, on a aussi l'égalité $V = F \oplus E$.

Exemples.

1. Dans tout espace vectoriel V, on a $V = \{0\} \oplus V = V \oplus \{0\}$.
2. Soient D et D' deux droites vectorielles distinctes du plan $\Pi = \mathbb{R}^2$.
   Alors $\Pi = D \oplus D' = D' \oplus D$.
3. Soient respectivement D et Π une droite vectorielle et un plan vectoriel ne la contenant pas dans l'espace $V = \mathbb{R}^3$. Alors $V = D \oplus \Pi = \Pi \oplus D$.
4. Soient respectivement D et H une droite et un hyperplan ne la contenant pas dans l'espace vectoriel V. On a alors $V = D \oplus H = H \oplus D$. Cet exemple admet comme cas particuliers les deux exemples précédents.

Lemme 33. (i) On suppose que $V = E + F$. Soient respectivement $\mathcal{C}$ et $\mathcal{D}$ des parties génératrices de E et de F. Alors $\mathcal{C} \cup \mathcal{D}$ est une partie génératrice de V.
(ii) On suppose que $E \cap F = \{0\}$. Soient respectivement $\mathcal{C}$ et $\mathcal{D}$ des parties libres de E et de F. Alors $\mathcal{C}$ et $\mathcal{D}$ sont disjointes et $\mathcal{C} \cup \mathcal{D}$ est libre.

Démonstration. (i) Tout $x \in V$ s'écrit $y + z$, où y est combinaison linéaire d'éléments de $\mathcal{C}$ et où z est combinaison linéaire d'éléments de $\mathcal{D}$; donc x est combinaison linéaire d'éléments de $\mathcal{C} \cup \mathcal{D}$.
(ii) Tout élément commun à $\mathcal{C}$ et $\mathcal{D}$ serait commun à E et F donc nul, contredisant la liberté de $\mathcal{C}$ et $\mathcal{D}$. Toute relation linéaire $\sum_{b \in \mathcal{C} \cup \mathcal{D}} \lambda_b b = 0$ entraîne :

$$\sum_{b \in \mathcal{C}} \lambda_b b = -\sum_{b \in \mathcal{D}} \lambda_b b \in E \cap F, \qquad \text{d'où} \qquad \sum_{b \in \mathcal{C}} \lambda_b b = \sum_{b \in \mathcal{D}} \lambda_b b = 0$$

et les coefficients λ_b sont nuls pour tout $b \in \mathcal{C}$ et pour tout $b \in \mathcal{D}$. ∎

La proposition suivante découle immédiatement de ce lemme :

Proposition 34. On suppose que $V = E \oplus F$. Soient respectivement $\mathcal{C}$ et $\mathcal{D}$ des bases de E et de F. Alors $\mathcal{C}$ et $\mathcal{D}$ sont disjointes et leur réunion $\mathcal{B} := \mathcal{C} \cup \mathcal{D}$ est une base de V.

Le lemme et la proposition s'étendent à des familles. Dans ce cas, on suppose que les bases $\mathcal{C} = (y_i)_{i \in I}$ et $\mathcal{D} = (z_j)_{j \in J}$ sont indexées par des ensembles d'indices I et J *disjoints*. On pose $L = I \cup J$, et, pour tout $\ell \in L$, $x_\ell = y_\ell$ si $\ell \in I$ et $x_\ell = z_\ell$ si $\ell \in J$. La famille $\mathcal{B} = (x_\ell)_{\ell \in L}$ est alors une base de V. La définition suivante concerne aussi bien les parties basiques que les bases qui sont des familles.

Définition 23. La base $\mathcal{B}$ obtenue par *juxtaposition* des bases $\mathcal{C}$ et $\mathcal{D}$ de E et de F. est dite *adaptée* à la décomposition en somme directe $V = E \oplus F$.

On peut construire des décompositions en sommes directes et des bases adaptées de la manière suivante. Soit $\mathcal{B}$ une base de V et supposons que $\mathcal{C}$ et $\mathcal{D}$ en sont une partition, c'est-à-dire que $\mathcal{B} = \mathcal{C} \cup \mathcal{D}$ et $\mathcal{C} \cap \mathcal{D} = \varnothing$. Alors les sous-espaces vectoriels E et F respectivement engendrés par les parties libres $\mathcal{C}$ et $\mathcal{D}$ admettent ces parties comme bases, et l'on a $V = E \oplus F$. La base $\mathcal{B}$ est adaptée à cette décomposition.

Définition 24. On dit que les sous-espaces vectoriels E et F de V sont *supplémentaires* ou que F est *un supplémentaire* de E si $V = E \oplus F$.

On précise parfois que F est un supplémentaire de E *dans* V. Bien sûr, E est alors un supplémentaire de F. On prendra garde à ne pas confondre *supplémentaire* et *complémentaire* : on n'a ici ni $E \cap F = \varnothing$ (le sous-espace trivial $\{0\}$ n'est pas l'ensemble vide) ni $E \cup F = V$ (la réunion de E et de F n'est en général pas un espace vectoriel).

Exemples. Le seul supplémentaire de $\{0\}$ dans V est V ; le seul supplémentaire de V dans V est $\{0\}$. Les supplémentaires d'une droite de $\mathbb{R}^2$ sont toutes les autres droites. Les supplémentaires d'une droite de $\mathbb{R}^3$ sont tous les plans ne la contenant pas. Les supplémentaires d'une droite de V sont tous les hyperplans ne la contenant pas.

Théorème 35. Tout sous-espace E de V admet au moins un supplémentaire.

Démonstration. Soit $\mathcal{C}$ une base de E ; elle est libre, donc (théorème 30 de la page 196), elle s'étend en une base $\mathcal{B}$ de V. Soit alors $\mathcal{D} = \mathcal{B} \setminus \mathcal{C}$. Le sous-espace F engendré par $\mathcal{D}$ est un supplémentaire de E dans V. ∎

Notons que cette preuve repose indirectement sur le lemme 29 de la page 195 qui a été admis. En général, un sous-espace admet beaucoup de supplémentaires : le seul cas d'unicité est celui de $\{0\}$ et V (voir l'exercice II.3.26 de la page 206). On prouvera plus loin (corollaire 38 de la page ci-contre) que ces supplémentaires sont tous isomorphes entre eux.

Théorème 36. Soit $f : V \to V'$ linéaire. Alors f induit, par restriction, un isomorphisme de tout supplémentaire de $\operatorname{Ker} f$ sur $\operatorname{Im} f$.

Démonstration. Soit F un supplémentaire de $\operatorname{Ker} f$ et soit $g = f_{|F}$ la restriction de f à F. Le noyau de g est $\operatorname{Ker} g = \operatorname{Ker} f \cap F = \{0\}$, et g est injective. D'autre part :

$$\operatorname{Im} f = f(V) = f(\operatorname{Ker} f + F) = f(\operatorname{Ker} f) + f(F),$$

d'où $\operatorname{Im} f = f(F) = g(F) = \operatorname{Im} g$, et g est surjective de F sur $\operatorname{Im} f$. ∎

4.2 Projections

Théorème 37. Supposons $V = E \oplus F$. Pour tout $x \in V$, notons $p(x) \in E$ et $q(x) \in F$ les éléments (uniquement déterminés) tels que $x = p(x) + q(x)$. Alors les applications p et q de V dans V sont des projecteurs tels que $E = \operatorname{Im} p = \operatorname{Ker} q$ et $F = \operatorname{Ker} p = \operatorname{Im} q$.

Démonstration. Par définition, $q = \mathrm{Id}_V - p$. Nous allons démontrer que p est un pro-

jecteur d'image E et de noyau F, le reste en découlant immédiatement par symétrie. Soient $\lambda, \mu \in K$ des scalaires et $x, y \in V$ des vecteurs. Alors :

$$\lambda x + \mu y = \big(\lambda p(x) + \mu p(y)\big) + \big(\lambda q(x) + \mu q(y)\big).$$

Comme $\lambda p(x) + \mu p(y) \in E$ et $\lambda q(x) + \mu q(y) \in F$, on a là la décomposition de $\lambda x + \mu y$ dans $E \oplus F$, et, par définition : $\lambda p(x) + \mu p(y) = p(\lambda x + \mu y)$ et $\lambda q(x) + \mu q(y) = q(\lambda x + \mu y)$, d'où la linéarité. Par définition, pour tout x de V, $p(x) \in E$. Enfin, si $x \in E$, sa décomposition est $x = x + 0$, d'où $p(x) = x$. On reconnaît les conditions (4) de la page 180, donc p est bien un projecteur. ■

Corollaire 38. Soit E un sous-espace vectoriel de V. Alors tous les supplémentaires de E sont isomorphes entre eux.

Démonstration. Soit F un supplémentaire de E choisi arbitrairement. Nous reprenons les notations du théorème : $E = \operatorname{Ker} q$. D'après le théorème 36 de la page ci-contre, tout supplémentaire de E est isomorphe à $\operatorname{Im} q$, c'est-à-dire à F. Plus précisément, si F' est un supplémentaire quelconque de E, la restriction de q à F' est un isomorphisme de F' sur F. ■

Corollaire 39. Pour tout sous-espace $E \subset V$, il existe un projecteur p d'image E.

Démonstration.
D'après le théorème 35 de la page précédente, E admet un supplémentaire. ■

Définition 25. L'endomorphisme p du théorème est appelé *projection sur E parallèlement à F* (et q est alors la projection sur F parallèlement à E).

Exemples.

1. Soient D et D' deux droites distinctes du plan vectoriel $\Pi = \mathbb{R}^2$. Nous avons déjà rencontré la projection sur D parallèlement à D'.
2. Soit f une forme linéaire non triviale sur V, et soit H l'hyperplan $\operatorname{Ker} f$.
   Pour tout $x_0 \notin H$, la droite vectorielle $D = Kx_0$ est un supplémentaire de H. La projection sur D parallèlement à H a été calculée à la section 2.3 : elle est donnée par la formule $p(x) := \dfrac{f(x)}{f(x_0)} x_0$.

Le sous-espace E étant donné, la projection p sur E dépend du supplémentaire F choisi. Dans l'exemple précédent, prenons pour $D := \mathbb{R} \times \{0\}$ l'axe des abscisses. Prenons pour D' la droite d'équation $y = ax$: $D' := \{(x, ax) \mid x \in \mathbb{R}\}$. Ces droites sont supplémentaires si et seulement si $a \neq 0$. Pour décomposer $(x, y) \in \mathbb{R}^2$ selon $D \oplus D'$, on écrit :

$$(x, y) = (t, 0) + (u, au), \qquad \text{d'où} \qquad u = y/a \quad \text{et} \quad t = x - y/a.$$

On a donc $p(x, y) = (x - y/a, 0)$, qui dépend de a, donc du choix de D'.

4.2.1 Sommes directes et applications linéaires

On suppose que $V = E \oplus F$ et l'on note p et q les projections associées. Soit f une application linéaire de V dans un espace vectoriel V' et soient $f_{|E}$ et $f_{|F}$ ses restrictions à E et à F. Le couple $(f_{|E}, f_{|F})$ est élément de $\mathcal{L}(E, V') \times \mathcal{L}(F, V')$.

Théorème 40. L'application $f \mapsto (f_{|E}, f_{|F})$ est un isomorphisme de l'espace vectoriel $\mathcal{L}(V, V')$ sur l'espace vectoriel $\mathcal{L}(E, V') \times \mathcal{L}(F, V')$: *une application linéaire de source V est uniquement déterminée par ses restrictions à E et F.*

Démonstration. Le lecteur vérifiera sans peine que l'application considérée est linéaire et qu'elle admet pour réciproque l'application $(g, h) \mapsto g \circ p + h \circ q$. ■

Proposition 41. Pour que l'endomorphisme f de V commute avec p, il faut, et il suffit, que les sous-espaces E et F soient stables par f : $f(E) \subset E$ et $f(F) \subset F$.

Démonstration. Supposons que l'endomorphisme f commute avec p : $f \circ p = p \circ f$.
Si $x \in E$, on a $x = p(x)$ et par suite $p\big(f(x)\big) = f\big(p(x)\big) = f(x)$, autrement dit, $f(x) \in E$ et nous avons bien démontré l'inclusion $f(E) \subset E$. Si $x \in F$, on a $p(x) = 0$ et par suite $p\big(f(x)\big) = f\big(p(x)\big) = f(0) = 0$, autrement dit $f(x) \in F$ ce qui donne l'inclusion $f(F) \subset F$.
Supposons maintenant que $f(E) \subset E$ et $f(F) \subset F$. Alors, pour tout $x \in V$, $y = f\big(p(x)\big) \in E$ et $z = f\big(q(x)\big) \in F$, de sorte que $y + z = f\big(p(x)\big) + f\big(q(x)\big) = f\big(p(x) + q(x)\big) = f(x)$ est une décomposition selon la somme directe $E \oplus F$. Ainsi, $p\big(f(x)\big) = y = f\big(p(x)\big)$: nous avons bien démontré que $f \circ p = p \circ f$. ■

Théorème 42. Soit f un endomorphisme *central* de V, autrement dit, qui commute avec tous les endomorphismes de V. Alors f est une homothétie.

Démonstration. D'après le corollaire 39 de la page précédente et la proposition ci-dessus, tout sous-espace vectoriel de V est stable par f. Nous allons en fait montrer que *tout endomorphisme f de V laissant stable toutes les droites vectorielles est une homothétie.* L'hypothèse signifie que, pour toute droite vectorielle $D \subset V$, on a $f(D) \subset D$. Cela revient à dire que, pour tout $x \in V$ non nul, $f(x) \in \mathrm{Vect}(x)$. On peut donc écrire $f(x) = L(x)x$, où $L(x) \in K$ est parfaitement défini. Il suffit de montrer que $L(x)$ a la même valeur pour tout $x \in V$ non nul. Or, de l'égalité $f(\lambda x) = \lambda f(x)$, on tire que $L(\lambda x) = L(x)$: si x et y sont liés, on a donc $L(x) = L(y)$. D'autre part, si x et y sont linéairement indépendants, de l'égalité $f(x + y) = f(x) + f(y)$, on tire :

$$\big(L(x + y) - L(x)\big)x + \big(L(x + y) - L(y)\big)y = 0.$$

Comme x et y sont linéairement indépendants, on en déduit $L(x) = L(x + y) = L(y)$. ■

4.2.2 Symétries

On suppose, dans cette section, que la caractéristique du corps K n'est pas 2 (*cf.* la définition 39 de la page 142). Rappelons que, dans le plan Π, la symétrie d'axe une droite D et de direction une droite D' a été définie ainsi : soit $x \in \Pi$ et soit y sa projection sur D parallèlement à D'. Le symétrique de x par rapport à D parallèlement à D' est l'unique élément $x' \in \Pi$ tel que y est le milieu de x et de x' : $y = \dfrac{x + x'}{2}$; ainsi, $x' = 2y - x$. Si l'on écrit $x = y + z$, avec $y \in D$ et $z \in D'$, on constate que $x' = y - z$. La définition de la symétrie est donc liée à la décomposition en somme directe $\Pi = D \oplus D'$, ce qui justifie la définition plus générale suivante :

Définition 26. Notons p et q les projecteurs associés à la décomposition en somme directe $V = E \oplus F$. L'endomorphisme $p - q$ est appelé *symétrie d'axe E et de direction F*, ou encore *symétrie par rapport à E et parallèlement à F*.

C'est donc l'application s qui, à $x = y + z$, où $y \in E$ et $z \in F$, associe $x' = s(x) = y - z$. Ainsi, $s(x) - x = -2z \in F$: l'action de s « déplace x dans la direction F », d'où la terminologie.

Remarque. Si le corps K était de caractéristique 2, on aurait $p + q = p - q$, $y + z = y - z$ et s serait l'identité.

Proposition 43. La symétrie d'axe E et de direction F est un endomorphisme *involutif*, c'est à dire tel que $s^2 = \mathrm{Id}_V$. Son axe est formé des $x \in V$ tels que $s(x) = x$. Sa direction est formée des $x \in V$ tels que $s(x) = -x$.

Démonstration. De même que, si $x = y + z$, où $y \in E$ et $z \in F$, on a $x' = s(x) = y - z$, on voit que $s(x') = y - (-z) = y + z = x$, d'où $s\big(s(x)\big) = x$. L'endomorphisme s est bien involutif. Avec les mêmes notations, l'égalité $s(x) = x$ équivaut à $z = -z$, donc à $2z = 0$, donc à $z = 0$ *puisque la caractéristique du corps K n'est pas* 2 ; enfin, $z = 0$ signifie que $x \in E$. De la même manière, $s(x) = -x$ équivaut à $y = -y$, donc à $y = 0$, c'est-à-dire à $x \in F$. ■

L'axe et la direction d'une symétrie s sont donc bien déterminés par les formules suivantes :

$$\text{Axe de } s = \{x \in V \mid s(x) = x\}, \tag{7}$$

$$\text{Direction de } s = \{x \in V \mid s(x) = -x\}. \tag{8}$$

Notons également qu'une symétrie est son propre inverse : $s^{-1} = s$, c'est donc en particulier un automorphisme de V.

On a vu qu'une projection peut être caractérisée par une condition purement algébrique, l'idempotence : $p^2 = p$. On peut également caractériser les symétries par une condition algébrique, l'involutivité.

Proposition 44. Soit s une involution linéaire de V : $s \in \mathcal{L}_K(V)$ et $s^2 = \mathrm{Id}_V$ (c'est donc exactement la même chose qu'un endomorphisme involutif). Alors c'est une symétrie, dont l'axe E et la direction F sont donnés par les équations (7) et (8). Les projecteurs associés à la décomposition en somme directe $V = E \oplus F$ sont $p := \dfrac{\mathrm{Id}_V + s}{2}$ et $q := \dfrac{\mathrm{Id}_V - s}{2}$.

Démonstration. Puisque s est linéaire, p et q le sont. Notons

$$E := \{x \in V \mid s(x) = x\} \qquad \text{et} \qquad F := \{x \in V \mid s(x) = -x\}.$$

Soient $x \in V$ et $x' = s(x)$, d'où $x = s(x')$ (car s est une involution). Alors $y = p(x) = \dfrac{x + x'}{2}$ est tel que $s(y) = \dfrac{s(x) + s(x')}{2} = y$, c'est-à-dire que $y = p(x) \in E$. Si $x \in E$, de l'égalité $x = s(x)$, on tire $p(x) = \dfrac{x + x'}{2} = x$. On reconnaît les conditions (4) de la page 180 qui assurent que p est un projecteur d'image E. Son noyau est formé des x tels que $p(x) = \dfrac{x + s(x)}{2} = 0$, c'est donc F. Comme $p + q = \dfrac{\mathrm{Id}_V + s}{2} + \dfrac{\mathrm{Id}_V - s}{2} = \mathrm{Id}_V$, l'autre projecteur associé à la décomposition $V = E \oplus F$ est $\mathrm{Id}_V - p = q$. La symétrie d'axe E et de direction F est $p - q = \dfrac{\mathrm{Id}_V + s}{2} - \dfrac{\mathrm{Id}_V - s}{2} = s$. ■

Exemple. Considérons le $\mathbb{R}$-espace vectoriel $V = \mathcal{F}(\mathbb{R}, \mathbb{R})$ formé de toutes les applications de $\mathbb{R}$ dans $\mathbb{R}$. Pour toute application $f : \mathbb{R} \to \mathbb{R}$, notons $g := s(f)$ l'application $x \mapsto g(x) = f(-x)$. Comme $g(-x) = f(x)$, on a $s(g) = f$ et s est une involution : $s^2 = \mathrm{Id}_V$. Nous laissons au lecteur le soin de démontrer que s est linéaire. C'est donc une symétrie. La fonction f appartient à l'axe E de s (resp. à sa direction F) si $\forall x \in \mathbb{R}\ ,\ f(-x) = f(x)$ (resp. $\forall x \in \mathbb{R}\ ,\ f(-x) = -f(x)$).
Ainsi, E est l'ensemble des fonctions paires et F est l'ensemble des fonctions impaires. Les projections p et q associent respectivement à f les fonctions :

$$x \mapsto \frac{f(x) + f(-x)}{2} \qquad \text{et} \qquad x \mapsto \frac{f(x) - f(-x)}{2}\ .$$

L'égalité $V = E \oplus F$ signifie que toute application $f : \mathbb{R} \to \mathbb{R}$ s'écrit de manière unique comme somme d'une fonction paire et d'une fonction impaire, et cette écriture est donnée par la formule :

$$\forall x \in \mathbb{R}\ ,\ f(x) = \frac{f(x) + f(-x)}{2} + \frac{f(x) - f(-x)}{2}\ .$$

Exercice type corrigé

II.3.0 1) Soient E un K-espace vectoriel et $E^* = \mathcal{L}_K(E, K)$ son dual (définition 15 de la page 184). Soit $\mathcal{B} = (x_i)_{i \in I}$ une base de E. On note π_i les formes linéaires coordonnées associées (définition 21 de la page 194). Démontrer que la famille $\mathcal{B}^* = (\pi_i)_{i \in I}$ est libre dans E^*.
2) Démontrer que les éléments de $\mathrm{Vect}(\mathcal{B}^*)$ sont les $\varphi \in E^*$ telles que la famille $\big(\varphi(x_i)\big)_{i \in I}$ est à support fini.
3) En déduire que $\mathcal{B}^*$ est une base de E^* si, et seulement si, I est fini.
4) On prend $E = \mathbb{R}_n[X]$ (polynômes sur $\mathbb{R}$ de degré inférieur ou égal à n) et pour $\mathcal{B}$ la base canonique. Soient $k \in [\![0, n]\!]$ et $a \in K$. Exprimer la forme linéaire $P \mapsto P^{(k)}(a)$ dans la base $\mathcal{B}^*$.
5) On prend pour E le $\mathbb{R}$-espace vectoriel $\mathbb{R}^{(\mathbb{N})}$ des suites de réels à support fini et pour $\mathcal{B}$ la base $(\delta^{(n)})_{n \in \mathbb{N}}$ (voir l'exemple 4 de la page 166, complété à la section 3.3).
Démontrer que l'application $\varphi \mapsto \big(\varphi(\delta^{(n)})\big)_{n \in \mathbb{N}}$ de E^* dans $\mathbb{R}^{\mathbb{N}}$ est un isomorphisme.
Quelle est l'image de $\mathcal{B}^*$ par cet isomorphisme ?
Solution.
1) Soit $\sum\limits_{i \in I} \lambda_i \pi_i = 0$ une relation linéaire. Pour $j \in I$ arbitraire, en appliquant cette forme linéaire à x_j, on trouve $\lambda_j = \sum\limits_{i \in I} \lambda_i \pi_i(x_j) = 0$ (car $\pi_i(x_j) = \delta_{i,j}$).
La famille $\mathcal{B}^*$ est donc bien libre.
2) Tout élément de $\mathrm{Vect}(\mathcal{B}^*)$ s'écrit $\varphi = \sum\limits_{i \in J} \lambda_i \pi_i$, où J est fini : alors $\varphi(x_i)$ est nul pour $i \notin J$ et la famille $\big(\varphi(x_i)\big)_{i \in I}$ est bien à support fini. Réciproquement, si la famille $\big(\varphi(x_i)\big)_{i \in I}$ est à support fini, la forme linéaire $\psi := \sum\limits_{i \in I} \varphi(x_i) \pi_i$ est bien définie et l'on a $\psi(x_j) = \varphi(x_j)$ pour tout $j \in I$.
D'après le théorème 26 de la page 195, $\psi = \varphi$, d'où $\varphi \in \mathrm{Vect}(\mathcal{B}^*)$.
3) Si $\mathcal{B}^*$ est une base de E^*, l'unique forme linéaire φ telle que $\varphi(x_i) = 1$ pour tout $i \in I$ est dans $\mathrm{Vect}(\mathcal{B}^*)$, donc la famille constante 1 est à support fini, donc I est fini. Si I est fini, toute forme linéaire φ s'écrit $\sum\limits_{i \in I} \varphi(x_i) \pi_i$ d'après le calcul ci-dessus, et $\mathcal{B}^*$ est bien une base de E^*.

4) Si $P = a_0 + \cdots + a_n X^n$, on a $\pi_i(P) = a_i$ et $P^{(k)}(a) = \displaystyle\sum_{i=k}^{n} a_i \frac{i!}{(i-k)!} a^{i-k}$.

La forme linéaire $P \mapsto P^{(k)}(a)$ est donc $\displaystyle\sum_{i=k}^{n} \frac{i!}{(i-k)!} a^{i-k} \pi_i = \sum_{i=0}^{n-k} \frac{(i+k)!}{i!} a^i \pi_{i+k}$.

5) Il est clair que l'application $\varphi \mapsto \left(\varphi(\delta^{(n)})\right)_{n\in\mathbb{N}}$ est linéaire. Le théorème 26 de la page 195 dit qu'elle est bijective. Notons encore $(\pi_n)_{n\in\mathbb{N}}$ la famille $\mathcal{B}^*$.
L'image d'un π_m $(m \in \mathbb{N})$ est la suite $\left(\pi_m(\delta^{(n)})\right)_{n\in\mathbb{N}} = (\delta_{m,n})_{n\in\mathbb{N}} = \delta^{(m)}$. L'image de $\mathcal{B}^*$ est donc $\mathcal{B}$. Cela a bien un sens, puisque c'est une famille d'éléments de $\mathbb{R}^{(\mathbb{N})}$, donc de $\mathbb{R}^{\mathbb{N}}$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

II.3.1 Montrer que, dans la définition d'un espace vectoriel, il n'est pas nécessaire de supposer le groupe abélien : cette propriété est en effet une conséquence des autres axiomes. (Développer de deux manières le produit $(1+1).(x+y)$.) Montrer qu'en revanche l'axiome $1.x = x$ n'est pas une conséquence des autres axiomes. *Indication : prendre un corps K et un groupe abélien V quelconques et poser* $\forall \lambda \in K, \forall x \in V\ ,\ \lambda.x = 0$.

II.3.2 Soit V un $\mathbb{Q}$-espace vectoriel. Montrer que, pour tout entier $m \in \mathbb{Z}$ et pour tout $x \in V$, on a l'égalité $mx = m.x$. Le premier membre de l'égalité correspond à la définition correspondante dans tout groupe noté additivement ; le deuxième membre correspond à la loi de composition externe de V (il a un sens, puisque $m \in \mathbb{Q}$) ; par exemple, $x + x = 2.x$. Montrer que le groupe abélien V est *sans torsion*, autrement dit, pour tout entier $m \in \mathbb{Z}$ non nul et pour tout $x \in V$, on a $mx = 0 \Rightarrow x = 0$. Montrer que le groupe V est de plus *divisible* , autrement dit, pour tout entier $m \in \mathbb{Z}$ non nul et pour tout $x \in V$, il existe $y \in V$ tel que $x = my$.

II.3.3 $*$ Montrer que tout groupe abélien sans torsion et divisible peut être muni d'une unique structure de $\mathbb{Q}$-espace vectoriel.

II.3.4 Soient E un ensemble et $F \subset E$ un sous-ensemble. Démontrer que $\mathcal{P}(F)$ est un sous-espace vectoriel du $\mathbb{F}_2$-espace vectoriel $\mathcal{P}(E)$. Déterminer le sous-espace vectoriel de $\mathcal{P}(E)$ engendré par les singletons.

II.3.5 Soit V un $\mathbb{F}_2$-espace vectoriel. Montrer que tout élément du groupe V est son propre opposé. Soit réciproquement V un groupe noté additivement dans lequel tout élément est son propre opposé. Montrer que V est abélien. Montrer que le groupe V admet une unique structure de $\mathbb{F}_2$-espace vectoriel.

II.3.6 $*$ Soit V un $\mathbb{Q}$-espace vectoriel et soit A une partie dénombrable de V. Démontrer que le sous-espace vectoriel engendré par A est dénombrable (commencer par supposer A fini). En déduire que le $\mathbb{Q}$-espace vectoriel $\mathbb{R}$ n'est engendré par aucune partie dénombrable.

Exercices sur la section 2

II.3.7 Vérifier que l'application $x \mapsto \sqrt{2}\,x$ est un endomorphisme du $\mathbb{Q}$-espace vectoriel $\mathbb{R}$. Est-ce une homothétie ?

II.3.8 Soient V et V' deux espaces vectoriels sur $\mathbb{Q}$ et soit $f : V \to V'$ un morphisme de groupes. Montrer que f est $\mathbb{Q}$-linéaire.
Indication : dans l'égalité $f(mx) = mf(x)$, remplacer x par $(n/m)x$.

II.3.9 (i) Soient f et g deux formes linéaires non nulles sur le K-espace vectoriel V, de même noyau H. On suppose qu'il existe $x_0 \notin H$ tel que $f(x_0) = g(x_0)$. Démontrer que $f = g$. On pourra d'abord considérer les restrictions de f et g à H et à $\mathrm{Vect}(x_0)$.
(ii) On ne fait plus l'hypothèse $f(x_0) = g(x_0)$. Que peut-on cependant conclure au sujet de f et g ? On pourra remplacer g par l'application $h : x \mapsto \lambda g(x)$ pour un choix convenable de λ.

II.3.10 (i) On fixe $a \in K$. Montrer que l'application $P \mapsto P(a)$ est une forme linéaire non nulle sur le K-espace vectoriel $K[X]$ des polynômes sur K. On note H son noyau : c'est donc un hyperplan. Montrer que l'application $P \mapsto P - P(a)$ est un projecteur d'image H.
(ii) On suppose désormais K de caractéristique nulle. Montrer que l'application $D : P \mapsto P'$ est un endomorphisme de $K[X]$ et préciser son noyau et son image.
(iii) On note $I(P)$ l'unique polynôme dont le terme constant est nul et dont le polynôme dérivé est P ; ainsi : $I\big(\sum_{k=0}^{n} a_k X^k\big) = \sum_{k=0}^{n} \dfrac{a_k}{k+1} X^{k+1}$. Montrer que l'application $I : P \mapsto I(P)$ est un endomorphisme de $K[X]$ et préciser son noyau et son image. Calculer $I \circ D$ et $D \circ I$.

II.3.11 (i) Soit $p \in \mathcal{L}_K(V)$ un projecteur. Calculer $(\mathrm{Id}_V - p)^2$ et en déduire une nouvelle preuve du fait que $q = \mathrm{Id}_V - p$ est un projecteur.
(ii) Soit $s := p - q$. Exprimer s en fonction de p seul et calculer s^2.
(iii) Soit réciproquement $s \in \mathcal{L}_K(V)$ tel que $s^2 = \mathrm{Id}_V$. On suppose que le corps K n'est pas de caractéristique 2. Démontrer l'existence d'un unique projecteur p tel que $s = 2p - \mathrm{Id}_V$.

II.3.12 Soit $V = \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$. On définit deux endomorphismes de V : D qui, à f, associe sa dérivée f' ; et I qui, à f, associe celle de ses primitives qui s'annule en 0 : $x \mapsto \displaystyle\int_0^x f(t)\,\mathrm{d}t$. Démontrer que $D, I \in \mathcal{L}(V)$, $DI = \mathrm{Id}_V$ et $ID \neq \mathrm{Id}_V$. Est-ce que $D, I \in GL(V)$?

II.3.13 Soient n un entier naturel non nul, $E = [\![1, n]\!]$ et V le $\mathbb{F}_2$-espace vectoriel $\mathcal{P}(E)$. En posant $\chi(\lambda_1, \ldots, \lambda_n) := \{i \in [\![1, n]\!] \mid \lambda_i = 1\}$, on définit une application $\chi : \mathbb{F}_2^n \to V$. Démontrer que l'application χ est linéaire et bijective. Plus généralement, définir pour tout ensemble E une bijection linéaire $\chi : \mathbb{F}_2^E \to \mathcal{P}(E)$. Quelle est l'image par χ du sous-espace $\mathbb{F}_2^{(E)}$ de $\mathbb{F}_2^E$?

II.3.14 Soient $f : V \to V'$ et $f' : V' \to V''$ des applications linéaires. Montrer que $f' \circ f = 0$ (l'application linéaire nulle) si, et seulement si, $\mathrm{Im}\, f \subset \mathrm{Ker}\, f'$. Montrer que $\mathrm{Im}(f' \circ f) = f'(\mathrm{Im}\, f)$ et que $\mathrm{Ker}(f' \circ f) = f^{-1}(\mathrm{Ker}\, f')$.

II.3.15 (i) Soient $f : V \to V'$ une application linéaire et $(W'_i)_{i \in I}$ une famille de sous-espaces vectoriels de V'. Démontrer l'égalité : $f^{-1}\left(\bigcap_{i \in I} W'_i\right) = \bigcap_{i \in I} f^{-1}(W'_i)$.
(ii) Soient $W_1, \ldots, W_p$ des sous-espaces vectoriels de V. Démontrer l'égalité :

$$f\left(W_1 + \cdots + W_p\right) = f(W_1) + \cdots + f(W_p).$$

(iii) On suppose f surjective ; soient $W'_1, \ldots, W'_p$ des sous-espaces de V'. Démontrer l'égalité :

$$f^{-1}\left(W'_1 + \cdots + W'_p\right) = f^{-1}(W'_1) + \cdots + f^{-1}(W'_p).$$

Soit $p : \mathbb{R}^2 \to \mathbb{R}^2$ la projection $(x, y) \mapsto (x, 0)$. Soient D'_1 et D'_2 deux droites vectorielles distinctes de $\mathbb{R}^2$ que l'on suppose non horizontales. Comparer $p^{-1}(D'_1 + D'_2)$ et $p^{-1}(D'_1) + p^{-1}(D'_2)$. *On vérifiera que* $p^{-1}(D'_1) + p^{-1}(D'_2) = \{0\} \times \mathbb{R}$ *et* $p^{-1}(D'_1 + D'_2) = \mathbb{R}^2$.
(iv) On suppose f injective ; soient $(W_i)_{i\in I}$ des sous-espaces de V. Démontrer l'égalité :

$$f\left(\bigcap_{i\in I} W_i\right) = \bigcap_{i\in I} f(W_i).$$

Soit $p : \mathbb{R}^2 \to \mathbb{R}^2$ la projection $(x, y) \mapsto (x, 0)$. Soient D_1 et D_2 deux droites vectorielles distinctes de $\mathbb{R}^2$ que l'on suppose non verticales. Comparer $p(D_1 \cap D_2)$ et $p(D_1) \cap p(D_2)$. (On vérifiera que $p(D_1 \cap D_2) = \{0\}$ et que $p(D_1) \cap p(D_2) = \mathbb{R} \times \{0\}$.)

Exercices sur la section 3

II.3.16 Soit $(x_i)_{i\in I}$ une famille génératrice du K-espace vectoriel V. Soient f et g deux applications linéaires de V dans V' qui coïncident sur la famille : $\forall i \in I\ ,\ f(x_i) = g(x_i)$. Montrer que $f = g$.

II.3.17 $*$ Soit $V = \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ le $\mathbb{R}$-espace vectoriel des fonctions indéfiniment dérivables. Soit, pour $a \in \mathbb{R}$, f_a la fonction e^{ax}. Montrer que les f_a $(a \in \mathbb{R})$ sont linéairement indépendants.
Indication : considérer une relation linéaire non triviale comportant le plus petit nombre de termes possible et la dériver pour rabaisser ce nombre de termes.

II.3.18 $*$ (i) Soit $V = \mathbb{R}^\mathbb{N}$ le $\mathbb{R}$-espace vectoriel des suites numériques. Soit, pour $q \in \mathbb{R}$, $g^{(q)}$ la suite des q^n $(n \in \mathbb{N})$. Montrer que les $g^{(q)}$ $(q \in \mathbb{R})$ sont linéairement indépendants.
Indication : pour toute combinaison linéaire $\underline{u} = \sum a_q g^{(q)}$, *on pourra étudier le comportement de* u_n *lorsque* $n \to +\infty$.
(ii) On note maintenant $V = K^\mathbb{N}$ et, pour $q \in \mathbb{K}$, $g^{(q)}$ la suite des q^n $(n \in \mathbb{N})$.
Montrer que les $g^{(q)}$ $(q \in K)$ sont linéairement indépendants.
Indication : considérer une relation linéaire non triviale comportant le plus petit nombre de termes possible et chercher à rabaisser ce nombre de termes.
(iii) On suppose maintenant le corps K de caractéristique 0, de sorte que l'on peut identifier $\mathbb{N}$ à une partie de K. Pour tout $q \in K^*$ et tout $k \in \mathbb{N}$, on note $g^{(q,k)}$ la suite des $n^k q^n$ $(n \in \mathbb{N})$. Montrer que les $g^{(q,k)}$ $((q, k) \in K^* \times \mathbb{N})$ sont linéairement indépendants.
Même méthode que ci-dessus.

II.3.19 (i) Soient $a_1, \ldots, a_p$ des éléments du corps K. Dans le K-espace vectoriel $K^\mathbb{N}$ des suites d'éléments de K, on considère le sous-espace vectoriel suivant :

$$V := \{\underline{u} = (u_n)_{n\in\mathbb{N}} \mid \forall n \in \mathbb{N}\ ,\ u_{n+p} = a_1 u_{n+p-1} + \cdots + a_p u_n\}.$$

On dit que les suites de V satisfont une relation de récurrence linéaire à p pas. Démontrer que V est un sous-espace vectoriel de $K^\mathbb{N}$.
(ii) Démontrer que l'application $\underline{u} \mapsto (u_0, \ldots, u_{p-1})$ de $K^\mathbb{N}$ dans K^p est linéaire et que sa restriction à V est un isomorphisme de V sur K^p. En déduire la dimension de V.

II.3.20 Soit $\mathcal{B} = (x_i)_{i\in I}$ une base du K-espace vectoriel V et soit f une application linéaire de V dans V'. Démontrer que f est injective (resp. surjective, resp. bijective) si, et seulement si, la famille $f(\mathcal{B}) = (f(x_i))_{i\in I}$ de V' est libre (resp. génératrice, resp. une base).

II.3.21 Démontrer que, si I est infini, le K-espace vectoriel $K^{(I)}$ n'est pas de dimension finie.

II.3.22 Soient $K \subset L \subset M$ trois corps commutatifs. Soient $(x_i)_{i\in I}$ une base du K-espace vectoriel L et $(y_j)_{j\in J}$ une base du L-espace vectoriel M. Démontrer que la famille des $x_i y_j$ est une base du K-espace vectoriel M.

II.3.23 $*$ (i) Soit $(B_i)_{i\in I}$ une famille de parties libres d'un espace vectoriel V telles que deux d'entre elles sont *comparables pour l'inclusion*, autrement dit, si $i, j \in I$, alors $B_i \subset B_j$ ou $B_j \subset B_i$. Démontrer que $\bigcup_{i\in I} B_i$ est libre.

Indication : si $x_1, \ldots, x_p \in \bigcup_{i\in I} B_i$, *alors* $\exists i \in I \ : \ x_1, \ldots, x_p \in B_i$.

(ii) Soient A une partie libre et C une partie génératrice de V, telles que $A \subset C$. Soit E l'ensemble des parties libres B telles que $A \subset B \subset C$, que l'on ordonne par inclusion. Démontrer que E est inductif (module I.1). En déduire, à l'aide du lemme de Zorn, une preuve du lemme 29.

II.3.24 $*$ On appelle *base de Hamel* une base du $\mathbb{Q}$-espace vectoriel $\mathbb{R}$. Démontrer qu'une telle base n'est pas dénombrable. Utiliser l'existence d'une base de Hamel pour construire un morphisme de groupe non trivial $\mathbb{R} \to \mathbb{R}$ qui n'est pas bijectif. Montrer que ce morphisme n'est pas continu. En admettant que, si E est infini, $\operatorname{card} E + \operatorname{card} E = \operatorname{card} E$ (module I.1, section 6), en déduire que le groupe $\mathbb{R}$ est isomorphe au groupe $\mathbb{R} \times \mathbb{R}$.

Exercices sur la section 4

II.3.25 Soient D, D', D'' trois droites vectorielles deux à deux distinctes de $\mathbb{R}^2$. Calculer leurs intersections deux à deux. Calculer leur somme. Y a-t-il unicité de la décomposition d'un $x \in \mathbb{R}^2$?

II.3.26 Montrer qu'un sous-espace vectoriel de V autre que 0 et V a une infinité de supplémentaires (on suppose le corps infini).

II.3.27 (i) Montrer que $f \in \mathcal{L}_K(V, V')$ est injective si, et seulement s'il existe $g \in \mathcal{L}_K(V', V)$ telle que $g \circ f = \mathrm{Id}_V$.
(ii) Montrer que $f \in \mathcal{L}_K(V, V')$ est surjective si, et seulement s'il existe $h \in \mathcal{L}_K(V', V)$ telle que $f \circ h = \mathrm{Id}_{V'}$.
(iii) Si f est bijective, $g = h = f^{-1}$.

II.3.28 Les sous-espaces $V_1 \times \{0\}$ et $\{0\} \times V_2$ sont supplémentaires dans $V_1 \times V_2$ et les projections associées sont $s_1 \circ p_1$ et $s_2 \circ p_2$.

II.3.29 Si V n'est ni trivial ni une droite vectorielle, il admet des endomorphismes qui ne sont pas des homothéties.

II.3.30 Donner un exemple d'endomorphisme d'un espace vectoriel V et de sous–espace vectoriel E de V tels que $f(E) = E$ mais aucun supplémentaire de E n'est stable.
Indication : si (u, v) *est une base de* V, *poser* $f(u) = u$ *et* $f(v) = u + v$.

II.3.31 Soit E un sous-espace vectoriel de V. Soit F' un sous-espace de V tel que $E \cap F' = \{0\}$. Démontrer qu'il existe un supplémentaire F de E tel que $F' \subset F$. (Appliquer le théorème 30 de la page 196.) Soit $x \in V \setminus E$. Déduire de ce qui précède l'existence d'une forme linéaire f nulle sur E et telle que $f(x) = 1$. Démontrer que E est l'intersection des hyperplans qui le contiennent.

II.3.32 Nous supposons que $V = E \oplus F$ et que p et q sont les projecteurs associés. Soit f un endomorphisme de V.
Démontrer que, pour que E soit stable par f, il faut, et il suffit, que $q \circ f \circ p = 0$.

Calcul matriciel élémentaire

II.4

Nous avons déjà rencontré des matrices en deux occasions au moins : représentation matricielle d'une permutation (module II.2), « codage » des endomorphismes de K^2 (module II.3). On peut également citer la matrice des pixels d'une description d'écran.

1 Algèbre matricielle

1.1 Définitions et généralités

Soient $m, n \in \mathbb{N}^*$ deux entiers naturels non nuls et soit E un ensemble quelconque.

Définition 1. Une matrice *de type* (m, n) (ou à m lignes et n colonnes) *à coefficients dans* E (ou *sur* E) est une famille $(a_{i,j})_{1\leqslant i\leqslant m, 1\leqslant j\leqslant n} \in E^{[\![1,m]\!]\times[\![1,n]\!]}$ d'éléments de E, indexée par l'ensemble produit $[\![1, m]\!] \times [\![1, n]\!]$.
L'ensemble de ces matrices est noté $M_{m,n}(E)$.
Une *matrice ligne* est une matrice à une seule ligne, donc un élément de $M_{1,n}(E)$.
Une *matrice colonne* est une matrice à une seule colonne, donc un élément de $M_{m,1}(E)$.

Par exemple, $M_{1,1}(E) = E$. On convient de représenter par un tableau rectangulaire la matrice $A = (a_{i,j})_{1\leqslant i\leqslant m, 1\leqslant j\leqslant n}$:

$$A = \begin{pmatrix} a_{1,1} & \dots & a_{1,j} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1} & \dots & a_{i,j} & \dots & a_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m,1} & \dots & a_{m,j} & \dots & a_{m,n} \end{pmatrix}.$$

Le *coefficient d'indices* (i, j) *de* A, qui est ici $a_{i,j}$ mais que l'on peut également noter $(A)_{i,j}$, est donc situé à l'intersection de la $i^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne du tableau, les lignes étant numérotées de 1 à m de haut en bas, et les colonnes étant numérotées de 1 à n de gauche à droite.

On appelle alors $i^{ème}$ *ligne de* A (pour $1 \leqslant i \leqslant m$) la matrice ligne :

$$L_i := \begin{pmatrix} a_{i,1} & \dots & a_{i,j} & \dots & a_{i,n} \end{pmatrix} \in M_{1,n}(E);$$

on l'identifie avec le n-uplet $(a_{i,1}, \dots, a_{i,n}) \in E^n$.

On appelle $j^{ème}$ colonne de A (pour $1 \leqslant j \leqslant n$) la matrice colonne :

$$C_j := \begin{pmatrix} a_{1,j} \\ \vdots \\ a_{i,j} \\ \vdots \\ a_{m,j} \end{pmatrix} \in M_{m,1}(E).$$

Pour indiquer que A est formée des lignes $L_1, \dots, L_m$ (resp. des colonnes $C_1, \dots, C_n$), on pourra écrire :

$$A = \begin{bmatrix} L_1 \\ \vdots \\ L_m \end{bmatrix} = [C_1; \dots; C_n]. \tag{1}$$

Tableau horaire d'occupation des salles.. On veut représenter le tableau suivant d'occupation de 3 salles le matin et l'après-midi par les 3 promotions :

	Salle 1	Salle 2	Salle 3
Matin :	vide	L3	L1
Après-midi :	L2	L1	vide

Soient $E = \{\text{L1}, \text{L2}, \text{L3}, \text{vide}\}$ et $m = 2, n = 3$. La matrice $A = (a_{i,j})_{1 \leqslant i \leqslant 2, 1 \leqslant j \leqslant 3}$ telle que $a_{1,1} = a_{2,3} = \text{vide}$, $a_{1,3} = a_{2,2} = \text{L1}$, $a_{1,2} = \text{L3}$ et $a_{2,1} = \text{L2}$ est représentée par $\begin{pmatrix} \text{vide} & \text{L3} & \text{L1} \\ \text{L2} & \text{L1} & \text{vide} \end{pmatrix}$. Ses lignes sont : $L_1 = \begin{pmatrix} \text{vide} & \text{L3} & \text{L1} \end{pmatrix}$ (occupation des salles le matin) et $L_2 = \begin{pmatrix} \text{L2} & \text{L1} & \text{vide} \end{pmatrix}$ (occupation des salles l'après-midi) et ses colonnes $C_1 = \begin{pmatrix} \text{vide} \\ \text{L2} \end{pmatrix}$ (horaire d'occupation de la salle 1), $C_2 = \begin{pmatrix} \text{L3} \\ \text{L1} \end{pmatrix}$ (horaire d'occupation de la salle 2) et $C_3 = \begin{pmatrix} \text{L1} \\ \text{vide} \end{pmatrix}$ (horaire d'occupation de la salle 3).

L'ensemble $M_{m,n}(E)$ est par définition identique à $E^{[\![1,m]\!] \times [\![1,n]\!]}$ et il s'identifie donc à $\mathcal{F}([\![1,m]\!] \times [\![1,n]\!], E)$. Plus généralement, I et J étant deux ensembles finis, les matrices de type (I, J) sur E sont les éléments de $E^{I \times J}$. Pour les représenter graphiquement, il faut introduire une relation d'ordre total sur I et sur J. Pratiquement, la plupart des définitions et des résultats de ce module se transposent sans difficulté au cas de matrices dont les ensembles d'indices sont autres que $[\![1,m]\!]$ et $[\![1,n]\!]$.

Exemple. L'espace vectoriel $K_n[X]$ des polynômes de degré inférieur ou égal à n sur le corps K (module II.6, proposition 5 de la page 287) admet pour base $(1, X, \dots, X^n)$. Le vecteur colonne des coordonnées d'un polynôme $P \in K_n[X]$ dans cette base admet donc $(n+1)$ composantes, qu'il est naturel d'indexer de 0 à n. Si l'on forme une matrice ayant pour colonnes de tels vecteurs colonnes, les lignes de cette matrice seront naturellement numérotées de 0 à n.

Nous considérerons dorénavant des matrices à coefficients dans un corps commutatif K. Cependant, la plupart des définitions et une partie des résultats seraient tout aussi valides sur un anneau commutatif quelconque. L'identification de l'ensemble $M_{m,n}(K)$ à $K^{[\![1,m]\!]\times[\![1,n]\!]}$ permet de le munir d'une structure de K-espace vectoriel (module II.3), avec les opérations :

$$\lambda \begin{pmatrix} a_{1,1} & \dots & a_{1,j} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1} & \dots & a_{i,j} & \dots & a_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m,1} & \dots & a_{m,j} & \dots & a_{m,n} \end{pmatrix} + \lambda' \begin{pmatrix} a'_{1,1} & \dots & a'_{1,j} & \dots & a'_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a'_{i,1} & \dots & a'_{i,j} & \dots & a'_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a'_{m,1} & \dots & a'_{m,j} & \dots & a'_{m,n} \end{pmatrix} :=$$

$$\begin{pmatrix} \lambda a_{1,1} + \lambda' a'_{1,1} & \dots & \lambda a_{1,j} + \lambda' a'_{1,j} & \dots & \lambda a_{1,n} + \lambda' a'_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \lambda a_{i,1} + \lambda' a'_{i,1} & \dots & \lambda a_{i,j} + \lambda' a'_{i,j} & \dots & \lambda a_{i,n} + \lambda' a'_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \lambda a_{m,1} + \lambda' a'_{m,1} & \dots & \lambda a_{m,j} + \lambda' a'_{m,j} & \dots & \lambda a_{m,n} + \lambda' a'_{m,n} \end{pmatrix}.$$

De manière plus compacte : $(\lambda A + \lambda' A')_{i,j} = \lambda\,(A)_{i,j} + \lambda'\,(A')_{i,j}$.

L'espace vectoriel $M_{1,n}(E)$ des matrices lignes s'identifie ainsi à l'espace vectoriel K^n. Nous avons déjà rencontré (au module II.3) l'espace vectoriel $M_{m,1}(E)$ des vecteurs colonnes. Plus généralement, on peut calculer les combinaisons linéaires de matrices ligne par ligne ou colonne par colonne :

$$\lambda \begin{bmatrix} L_1 \\ \vdots \\ L_m \end{bmatrix} + \lambda' \begin{bmatrix} L'_1 \\ \vdots \\ L'_m \end{bmatrix} = \begin{bmatrix} \lambda L_1 + \lambda' L'_1 \\ \vdots \\ \lambda L_m + \lambda' L'_m \end{bmatrix},$$

$$\lambda[C_1;\dots;C_n] + \lambda'[C'_1;\dots;C'_n] = [\lambda C_1 + \lambda' C'_1;\dots;\lambda C_n + \lambda' C'_n].$$

Définition 2. Les *matrices élémentaires* de $M_{m,n}(K)$ sont les matrices $E_{i,j}^{(m,n)}$ $(1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n)$ ainsi définies : le coefficient d'indices (i,j) de $E_{i,j}^{(m,n)}$ est égal à 1, tous ses autres coefficients sont nuls.
On a donc, pour $1 \leqslant k \leqslant m$ et $1 \leqslant l \leqslant n$:

$(E_{i,j}^{(m,n)})_{k,l} := \delta_{i,k}\delta_{j,l}$ (avec la notation de Kronecker, voir exemple de la page 56).

Proposition et définition 1. Le couple d'entiers naturels (m,n) étant fixé, les $E_{i,j}^{(m,n)}$, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$, forment une base de $M_{m,n}(K)$, appelée la *base canonique*. La dimension de $M_{m,n}(K)$ est donc mn.

Démonstration. La base canonique d'un espace vectoriel de la forme $K^{(X)}$ a été décrite au module II.3 à la page 193 : nous avons ici le cas particulier $X = [\![1,m]\!] \times [\![1,n]\!]$.

Pratiquement, les coordonnées d'une matrice dans la base canonique sont ses coefficients :

$$(a_{i,j})_{1\leqslant i\leqslant m, 1\leqslant j\leqslant n} = \sum_{1\leqslant i\leqslant m, 1\leqslant j\leqslant n} a_{i,j} E_{i,j}^{(m,n)}. \qquad \blacksquare$$

La notation $E_{i,j}^{(m,n)}$ des matrices élémentaires sera allégée en $E_{i,j}$ lorsqu'il n'y aura pas de risque de confusion. On pourra donc, par exemple, distinguer les matrices

$$E_{1,1} = \begin{pmatrix} 1 & 0 \end{pmatrix} \in M_{1,2}(K)\,,\ E_{1,1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \in M_{2,2}(K) \quad \text{et} \quad E_{1,1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in M_{2,1}(K)$$

en se fiant au contexte.

Appelons *hauteur* d'une matrice colonne non nulle $C = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \in M_{m,1}(K)$ et notons $h(C)$ l'entier $i \in [\![1, m]\!]$ tel que $a_1 = \cdots = a_{m-i} = 0$ et $a_{m-i+1} \neq 0$. On appellera *pivot* de C le premier élément non nul : a_{m-i+1}. Si $C = 0$, nous posons $h(C) := 0$ (et C n'a pas de pivot). Appelons de même *largeur* d'une matrice ligne non nulle $L = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix} \in M_{1,n}(K)$ et notons $\ell(L)$ l'entier $j \in [\![1, n]\!]$ tel que $a_1 = \cdots = a_{n-j} = 0$ et $a_{n-j+1} \neq 0$. On appellera *pivot* de L le premier élément non nul : a_{n-j+1}. Si $L = 0$, nous posons $\ell(L) := 0$ (et L n'a pas de pivot).

Exemple. Les matrices colonnes $\begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix}$ et $\begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix}$ ont pour hauteur 2.

Les matrices lignes $\begin{pmatrix} 0 & 0 & 2 & 1 & 4 \end{pmatrix}$ et $\begin{pmatrix} 0 & 0 & 2 & 0 & 0 \end{pmatrix}$ ont pour largeur 3.

Définition 3. La matrice de lignes $L_1, \dots, L_m$ est dite *échelonnée en lignes* si elle est nulle ou s'il existe un indice $k \in [\![1, m]\!]$ tel que $\ell(L_1) > \cdots > \ell(L_k) > 0$ et $L_i = 0$ pour $k < i \leqslant m$. La matrice de colonnes $C_1, \dots, C_n$ est dite *échelonnée en colonnes* si elle est nulle ou s'il existe un indice $\ell \in [\![1, n]\!]$ tel que $h(C_1) > \cdots > h(C_\ell) > 0$ et $C_j = 0$ pour $\ell < j \leqslant n$.

Exemple. Nous représentons ici par $\star$ des coefficients quelconques et par $\underline{\star}$ les pivots (qui sont donc non nuls). Les matrices $\begin{pmatrix} 0 & \underline{\star} & \star & \star & \star \\ 0 & 0 & 0 & \underline{\star} & \star \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ et $\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \underline{\star} & 0 & 0 & 0 & 0 \\ \star & \underline{\star} & 0 & 0 & 0 \\ \star & \star & \underline{\star} & 0 & 0 \end{pmatrix}$ sont respectivement échelonnées en lignes de largeurs $4, 2, 0$ et en colonnes de hauteurs $3, 2, 1, 0, 0$.

Exercice 1.

La matrice $A = E_{i,j}^{(m,n)}$ est-elle échelonnée ?

Solution. Ses lignes sont nulles, sauf la $i^{ème}$ qui est de largeur $n-j+1$. Elle est donc échelonnée en lignes si, et seulement si, $i=1$. De même, elle est échelonnée en colonnes si, et seulement si, $j=1$.

Exercice 2.

À quoi ressemble une matrice échelonnée en lignes et en colonnes ?

Solution. Si $a_{1,1}=0$, $h(C_1)<m$ donc $h(C_j)<m$ pour tout j, donc $L_1=0$, donc $L_i=0$ pour tout i et $A=0$. Si A n'est pas nulle, on a donc $a_{1,1}\neq 0$, donc $h(C_1)=m>h(C_2),\dots,h(C_n)$ et $l(L_1)=n>\ell(L_2),\dots,\ell(L_m)$. On voit donc que $L_1=\begin{pmatrix}a_{1,1} & 0 & \dots & 0\end{pmatrix}$ et $C_1=\begin{pmatrix}a_{1,1}\\0\\\vdots\\0\end{pmatrix}$. Réitérant le raisonnement, on trouve que A a des coefficients $a_{1,1},\dots,a_{k,k}$ non nuls et tous les autres coefficients nuls :

$$A=\begin{pmatrix}a_{1,1} & \dots & 0 & \dots & 0\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ 0 & \dots & a_{k,k} & \dots & 0\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ 0 & \dots & 0 & \dots & 0\end{pmatrix}.$$

Soient $m,n,p\in\mathbb{N}^*$. On peut définir le *produit* des matrices

$$A=(a_{i,j})_{1\leqslant i\leqslant m,1\leqslant j\leqslant n}\in M_{m,n}(K)\qquad\text{et}\qquad B=(a_{j,k})_{1\leqslant j\leqslant n,1\leqslant k\leqslant p}\in M_{n,p}(K)$$

comme la matrice $AB=(c_{i,k})_{1\leqslant i\leqslant m,1\leqslant k\leqslant p}\in M_{m,p}(K)$ donnée par les formules suivantes :

$$\forall(i,k)\in[\![1,m]\!]\times[\![1,p]\!]\ ,\ c_{i,k}:=\sum_{j=1}^{n}a_{i,j}b_{j,k}.$$

Il est commode de poser les calculs comme suit :

$$\begin{array}{cc} & \begin{pmatrix} & b_{1,k} & \\ & \vdots & \\ & b_{j,k} & \\ & \vdots & \\ & b_{n,k} & \end{pmatrix}\\ \begin{pmatrix} & & & & \\ & & & & \\ a_{i,1} & \dots & a_{i,j} & \dots & a_{i,n}\\ & & & & \end{pmatrix} & \begin{pmatrix} & \vdots & \\ & \vdots & \\ \dots & \dots & c_{i,k} & \\ & & & \end{pmatrix}\end{array}$$

Exemple. On prend ici $m = 3, n = 4, p = 2$:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \end{pmatrix} \times \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \\ b_{3,1} & b_{3,2} \\ b_{4,1} & b_{4,2} \end{pmatrix} =$$

$$\begin{pmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} + a_{1,3}b_{3,1} + a_{1,4}b_{4,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} + a_{1,3}b_{3,2} + a_{1,4}b_{4,2} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} + a_{2,3}b_{3,1} + a_{2,4}b_{4,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} + a_{2,3}b_{3,2} + a_{2,4}b_{4,2} \\ a_{3,1}b_{1,1} + a_{3,2}b_{2,1} + a_{3,3}b_{3,1} + a_{3,4}b_{4,1} & a_{3,1}b_{1,2} + a_{3,2}b_{2,2} + a_{3,3}b_{3,2} + a_{3,4}b_{4,2} \end{pmatrix}.$$

La multiplication des matrices A et B n'est donc possible que si le nombre de colonnes de A est égal au nombre de lignes de B. Par exemple, le produit de la matrice ligne $L = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix}$ par la matrice colonne $C = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$ n'est défini que si $m = n$, et il vaut alors $LC = a_1b_1 + \cdots + a_nb_n$; c'est un élément de $M_{1,1}(K) = K$. Plus généralement, si l'on note $L_1, \dots, L_m$ les lignes de A et $C_1, \dots, C_p$ les colonnes de B, on voit que les coefficients de AB sont aussi donnés par la formule : $c_{i,k} = L_iC_k$:

$$\begin{bmatrix} L_1 \\ \vdots \\ L_m \end{bmatrix} \times [C_1; \dots; C_p] = \begin{pmatrix} L_1C_1 & \dots & L_1C_p \\ \vdots & \ddots & \vdots \\ L_mC_1 & \dots & L_mC_p \end{pmatrix}.$$

Quelques conséquences de ces règles sont extrêmement utiles dans le calcul matriciel. Les colonnes de AB sont les AC_j :

$$A[C_1; \dots; C_p] = [AC_1; \dots; AC_p].$$

Les lignes de AB sont les L_iB :

$$\begin{bmatrix} L_1 \\ \vdots \\ L_m \end{bmatrix} B = \begin{bmatrix} L_1B \\ \vdots \\ L_mB \end{bmatrix}.$$

Le produit AX de A par un vecteur colonne X est une combinaison linéaire des colonnes de A :

$$A = [C_1; \dots; C_n] \Longrightarrow A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1C_1 + \cdots + x_nC_n. \tag{2}$$

Le produit YA d'un vecteur ligne Y par A est une combinaison linéaire des lignes de A :

$$A = \begin{bmatrix} L_1 \\ \vdots \\ L_m \end{bmatrix} \Longrightarrow \begin{pmatrix} y_1 & \dots & y_m \end{pmatrix} A = y_1L_1 + \cdots + y_mL_m.$$

Par exemple, si l'on note $(\Lambda_1, \ldots, \Lambda_m)$ la base canonique de $M_{1,m}(K)$ et $(\Gamma_1, \ldots, \Gamma_n)$ celle de $M_{n,1}(K)$, on en déduit :

$$A = [C_1; \ldots; C_n] \Longrightarrow A\Gamma_j = C_j,$$

$$A = \begin{bmatrix} L_1 \\ \vdots \\ L_m \end{bmatrix} \Longrightarrow \Lambda_i A = L_i.$$

Le lecteur est invité à contrôler la possibilité de toutes ces formules (compatibilité des nombres de lignes et de colonnes) pour en tester la vraisemblance. De même, il devrait effectuer soigneusement les calculs dont les résultats sont donnés ci-dessous.

Exemples.

1. Soit $A = [C_1; \ldots; C_n] \in M_{m,n}(K)$. Alors $AE_{k,\ell}^{(n,p)} = [0; \ldots; C_k; \ldots; 0]$, la colonne C_k étant placée en $\ell^{\text{ème}}$ position.

2. Soit $B = \begin{bmatrix} L_1 \\ \vdots \\ L_n \end{bmatrix} \in M_{n,p}(K)$. Alors $E_{i,j}^{(m,n)} B = \begin{bmatrix} 0 \\ \vdots \\ L_j \\ \vdots \\ 0 \end{bmatrix}$, la ligne L_j étant placée en $i^{\text{ème}}$ position.

3. Le produit de deux matrices élémentaires de formats compatibles est donné par la formule : $E_{i,j}^{(m,n)} E_{k,\ell}^{(n,p)} = \delta_{j,k} E_{i,\ell}^{(m,p)}$.

4. On a $[C_1; \ldots; C_n] \times \begin{bmatrix} L_1 \\ \vdots \\ L_n \end{bmatrix} = C_1 L_1 + \cdots + C_n L_n$. En effet, le terme $C_j L_j$ est la matrice de type (m, p) de coefficient général $c_{i,k}^{(j)} = a_{i,j} b_{j,k}$ et l'on a bien (dans la définition du produit AB) l'égalité $c_{i,k} = c_{i,k}^{(1)} + \cdots + c_{i,k}^{(n)}$.

Exercice 3.

À quoi ressemble le produit d'une colonne par une ligne ?

Solution. Le produit de $C \in M_{m,1}(K)$, de coefficients notés a_i (plutôt que $a_{i,1}$), par $L \in M_{1,p}(K)$, de coefficients notés b_k (plutôt que $b_{1,k}$), est toujours possible. On obtient une matrice (m, p) de coefficient général $a_i b_k$.

Nous noterons $I_n := (\delta_{i,j})_{1\leqslant i,j\leqslant n} = \sum_{i=1}^{n} E_{i,i}^{(n,n)}$ la *matrice identité d'ordre* n ; ses coefficients $a_{i,i}$ valent 1 et les autres 0 :

$$I_n = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Proposition 2. La multiplication est associative ; elle est distributive à droite et à gauche par rapport à l'addition ; les matrices identités sont neutres pour la multiplication :

$$\forall A \in M_{m,n}(K), \forall B \in M_{n,p}(K), \forall C \in M_{p,q}(K)\ ,\ (AB)C = A(BC)$$
$$\forall A, A' \in M_{m,n}(K), \forall B \in M_{n,p}(K)\ ,\ (A+A')B = AB + A'B,$$
$$\forall A \in M_{m,n}(K), \forall B, B' \in M_{n,p}(K)\ ,\ A(B+B') = AB + AB',$$
$$\forall A \in M_{m,n}(K)\ ,\ I_m A = A,$$
$$\forall A \in M_{m,n}(K)\ ,\ AI_n = A.$$

Démonstration. Notant $a_{i,j}, b_{j,k}, c_{k,\ell}$ les coefficients généraux respectifs de A, B, C, le calcul donne $\sum_{j=1}^{n}\sum_{k=1}^{p} a_{i,j}b_{j,k}c_{k,\ell}$ pour coefficient d'indices (i,ℓ) de $(AB)C$ comme de $A(BC)$, d'où l'associativité. L'égalité $(A+A')B = AB + A'B$ découle de l'égalité $\sum_{j=1}^{n}(a_{i,j}+a'_{i,j})b_{j,k} = \sum_{j=1}^{n} a_{i,j}b_{j,k} + \sum_{j=1}^{n} a'_{i,j}b_{j,k}$ entre les coefficients d'indices (i,k) des deux matrices. L'égalité $A(B+B') = AB + AB'$ se démontre de manière analogue. Les égalités $I_m A = A = AI_n$ découlent du fait que les lignes de I_m et les colonnes de I_n forment respectivement les bases canoniques $(\Lambda_1, \dots, \Lambda_m)$ de $M_{1,m}(K)$ et $(\Gamma_1, \dots, \Gamma_n)$ de $M_{n,1}(K)$. ∎

Les matrices $\lambda I_n = \begin{pmatrix} \lambda & 0 & \dots & 0 & 0 \\ 0 & \lambda & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda & 0 \\ 0 & 0 & \dots & 0 & \lambda \end{pmatrix}$ sont appelées *matrices scalaires*. On vérifie aisément que $(\lambda I_m)A = \lambda A = A(\lambda I_n)$, ce qui, joint à la proposition, entraîne une généralisation de la distributivité :

Corollaire 3. La multiplication des matrices est *bilinéaire* : pour tous $\lambda, \lambda' \in K$, $A, A' \in M_{m,n}(K)$ et $B, B' \in M_{n,p}(K)$:

$$(\lambda A + \lambda' A')B = \lambda AB + \lambda' A'B, \qquad \text{et} \qquad A(\lambda B + \lambda' B') = \lambda AB + \lambda' AB'.$$

Exercice 4.
Peut-on calculer A^2 ? Peut-on comparer AB et BA ?
Solution. On ne peut calculer A^2 que si A est de type (m,m). On ne peut comparer AB et BA que si A et B sont de type (m,m).

Coût des opérations sur les matrices. Il s'agit du nombre d'opérations (additions et multiplications) dans K qui sont nécessaires pour additionner et multiplier deux matrices. Chacun des mn coefficients de la somme de deux matrices de type (m,n) est une somme de deux coefficients. Additionner deux matrices de type (m,n) coûte donc mn additions élémentaires (c'est-à-dire dans le corps K). Chacun des mp coefficients du produit de deux matrices de types (m,n) et (n,p) s'obtient en additionnant n produits de deux coefficients : son calcule coûte donc n multiplications élémentaires et $(n-1)$ additions élémentaires. Il y a donc au total mnp multiplications élémentaires et $mp(n-1)$ additions élémentaires.

La définition et les premières propriétés de la transposition sont valables pour des matrices sur un ensemble E quelconque.

Définition 4. La *transposée* de la matrice $A = (a_{i,j})_{1\leqslant i\leqslant m, 1\leqslant j\leqslant n}$ de $M_{m,n}(E)$ est la matrice $B := (a_{j,i})_{1\leqslant j\leqslant n, 1\leqslant i\leqslant m}$ de $M_{n,m}(E)$ définie en posant $b_{j,i} = a_{i,j}$.
On la note $B = {}^tA$ et on l'appelle encore *matrice transposée*

Ainsi, la transposée d'une matrice ligne est une matrice colonne, et réciproquement. Avec les notations de l'équation (1) de la page 208 :

$$ {}^t[C_1;\ldots;C_n] = \begin{bmatrix} {}^tC_1 \\ \vdots \\ {}^tC_n \end{bmatrix} \qquad \text{et} \qquad {}^t\begin{bmatrix} L_1 \\ \vdots \\ L_m \end{bmatrix} = [{}^tL_1;\ldots;{}^tL_m]. $$

On vérifie facilement que ${}^t({}^tA) = A$. Autrement dit, les applications $A \mapsto {}^tA$ de $M_{m,n}(E)$ dans $M_{n,m}(E)$ et $B \mapsto {}^tB$ de $M_{n,m}(E)$ dans $M_{m,n}(E)$ sont réciproques l'une de l'autre.
Revenons au cas de matrices sur un corps commutatif K.

Proposition 4.
(i) L'application $A \mapsto {}^tA$ de $M_{m,n}(K)$ dans $M_{n,m}(K)$ est un isomorphisme.
(ii) Soient $A \in M_{m,n}(K)$ et $B \in M_{n,p}(K)$. Alors ${}^t(AB) = {}^tB\,{}^tA$.

Démonstration.
Nous laisserons au lecteur le plaisir de démontrer l'égalité ${}^t(\lambda_1 A_1 + \lambda_2 A_2) = \lambda_1\,{}^tA_1 + \lambda_2\,{}^tA_2$.
Pour la deuxième égalité, notons $C = AB$, $A' = {}^tA$, $B' = {}^tB$ et $C' = {}^tC$. Avec des notations évidentes, $c'_{k,i} = c_{i,k} = \sum_{j=1}^{n} a_{i,j}b_{j,k} = \sum_{j=1}^{k} b'_{k,j}a'_{j,i}$, de sorte que $C' = B'A'$. ∎

On peut remarquer enfin que la transposée d'une matrice échelonnée en lignes est une matrice échelonnée en colonnes et réciproquement.

1.2 Matrices carrées

On appellera *matrice carrée d'ordre* n une matrice de type (n,n) et l'on notera $M_n(K) = M_{n,n}(K)$. Les coefficients d'indices (i,i) $(1 \leqslant i \leqslant n)$ de la matrice $A = (a_{i,j})_{1\leqslant i,j\leqslant n}$ de $M_n(K)$ sont appelés *coefficients diagonaux*; les $a_{i,j}$ tels que $i \leqslant j$ (resp. $i < j$, resp. $i \geqslant j$, resp. $i > j$) sont dits *au dessus de la diagonale* (resp. *strictement au dessus de la diagonale*, resp. *au dessous de la diagonale*, resp. *strictement au dessous de la diagonale*). Parmi les matrices carrées, divers sous-ensembles méritent un nom spécifique; notant, comme ci-dessus, $A = (a_{i,j})_{1\leqslant i,j\leqslant n}$:

1. La matrice A est dite *triangulaire supérieure* si tous ses coefficients strictement au dessous de la diagonale sont nuls. Par exemple, les matrices carrées échelonnées en lignes sont triangulaires supérieures. L'ensemble $T_n(K)$ des matrices triangulaires supérieures est un sous–espace vectoriel de $M_n(K)$. Les $E_{i,j}$ tels que $1 \leqslant i \leqslant j \leqslant n$ en forment une base. Sa dimension est $n(n+1)/2$.
2. La matrice A est dite *triangulaire supérieure stricte* si tous ses coefficients au dessous de la diagonale sont nuls. L'ensemble $T_n^*(K)$ des matrices triangulaires supérieures strictes est un sous–espace vectoriel de $M_n(K)$. Les $E_{i,j}$ tels que $1 \leqslant i < j \leqslant n$ en forment une base. Sa dimension est $n(n-1)/2$.
3. On définit de même le sous–espace vectoriel $\mathcal{L}_n(K)$ (resp. $\mathcal{L}_n^*(K)$) des matrices *triangulaires inférieures* (resp. *triangulaires inférieures strictes*), dont une base est formée des $E_{i,j}$ tels que $1 \leqslant j \leqslant i \leqslant n$ (resp. tels que $1 \leqslant j < i \leqslant n$) et dont la dimension est $n(n+1)/2$ (resp. $n(n-1)/2$).
4. La matrice A est dite *diagonale* si ses coefficients non diagonaux sont nuls, autrement dit, si elle est à la fois triangulaire supérieure et triangulaire inférieure. La matrice diagonale de coefficients diagonaux $a_1, \ldots, a_n$ est notée $\mathrm{Diag}(a_1, \ldots, a_n)$. Par exemple, les matrices scalaires sont diagonales : $\lambda I_n = \mathrm{Diag}(\lambda, \ldots, \lambda)$.

   L'ensemble $D_n(K)$ des matrices diagonales est un sous–espace vectoriel de $M_n(K)$. Les $E_{i,i}$ tels que $1 \leqslant i \leqslant n$ en forment une base. Sa dimension est n.

Des diverses partitions de la base de $M_n(K)$ formée des $E_{i,j}$, on déduit autant de décompositions en sommes directes de sous–espace vectoriels :

$$M_n(K) = T_n(K) \oplus \mathcal{L}_n^*(K) = \mathcal{L}_n(K) \oplus T_n^*(K),$$
$$T_n(K) = T_n^*(K) \oplus D_n(K),$$
$$\mathcal{L}_n(K) = \mathcal{L}_n^*(K) \oplus D_n(K).$$

Deux sous-ensembles de $M_n(K)$ méritent un traitement à part. La matrice A est dite *symétrique* (resp. *antisymétrique*) si elle est égale (resp. si elle est opposée) à sa transposée, autrement dit, si l'on a, pour $1 \leqslant i \leqslant n$: $a_{i,j} = a_{j,i}$ (resp. $a_{i,j} = -a_{j,i}$). Par exemple, les matrices diagonales sont symétriques.

Les coefficients diagonaux d'une matrice antisymétrique vérifient $a_{i,i} = -a_{i,i}$. Si le corps K n'est pas de caractéristique 2, cela entraîne qu'ils sont nuls. En revanche, si le corps K est de caractéristique 2, il n'y a pas de différence entre matrices symétriques et matrices antisymétriques.

Les matrices symétriques forment un sous–espace vectoriel de $M_n(K)$, que l'on note $S_n(K)$. Une base de $S_n(K)$ est formée par les $E_{i,i}$ pour $1 \leqslant i \leqslant n$ et les $E_{i,j} + E_{j,i}$ pour $1 \leqslant i < j \leqslant n$. Sa dimension est donc $n(n+1)/2$. Les matrices antisymétriques forment également un sous–espace vectoriel de $M_n(K)$, que l'on note $A_n(K)$. *Si l'on suppose que le corps K n'est pas de caractéristique* 2, une base de $A_n(K)$ est formée par les $E_{i,j} - E_{j,i}$ pour $1 \leqslant i < j \leqslant n$. Sa dimension est donc $n(n-1)/2$ dans ce cas. Si le corps K est de caractéristique 2, on a déjà remarqué que $A_n(K) = S_n(K)$.

Proposition 5. On suppose que le corps K n'est pas de caractéristique 2. Les sous-espaces $S_n(K)$ et $A_n(K)$ sont supplémentaires dans $M_n(K)$ et les projecteurs associés sont les applications $A \mapsto \dfrac{A + {}^tA}{2}$ et $A \mapsto \dfrac{A - {}^tA}{2}$.

Démonstration. En effet, l'application $A \mapsto {}^tA$ est un automorphisme involutif de $M_n(K)$, donc une symétrie selon le module II.3 (définition 26 de la page 201), dont on peut appliquer les résultats. ∎

Exercice 5.
Déterminer $S_n(K) \cap T_n^*(K)$, $S_n(K) \cap \mathcal{L}_n^*(K)$, $A_n(K) \cap T_n(K)$ et $A_n(K) \cap \mathcal{L}_n(K)$. Que peut-on en conclure ?
Solution. Tous ces sous-espaces sont triviaux (pour les deux derniers, on doit supposer que car $K \neq 2$). En anticipant sur le module II.7, proposition 17 de la page 339, on en déduit autant de décompositions de $M_n(K)$ en somme directe : on y prouvera en effet que, si W_1 et W_2 sont des sous-espaces de V tels que $W_1 \cap W_2 = \{0\}$ et $\dim W_1 + \dim W_2 = \dim V$, alors $V = W_1 \oplus W_2$.

Il découle de la proposition 2 de la page 214 que, muni de l'addition et de la multiplication des matrices, $M_n(K)$ *est un anneau*, dont l'élément neutre est I_n. De plus, l'application $\lambda \mapsto \lambda I_n$ est un isomorphisme de K sur le sous-anneau de $M_n(K)$ formé des matrices scalaires. Enfin, ce dernier est *central* : une matrice scalaire commute avec tout élément de $M_n(K)$, en vertu des égalités $(\lambda I_n)A = \lambda A = A(\lambda I_n)$. On résume ces propriétés en disant que $M_n(K)$ est une K*-algèbre*, notion qui sera étudiée en L2.

Exemple. Puisque $M_n(K)$ est à la fois un anneau et un K-espace vectoriel, il est naturel de former la table de multiplication d'une base de $M_n(K)$. On choisit, bien sûr, la base des matrices élémentaires $E_{i,j}$, et l'on a les formules :

$$\forall i, j, k, l \in [\![1, n]\!] \ , \ E_{i,j} E_{k,l} = \delta_{j,k} E_{i,l}.$$

Lorsque $n = 1$, l'anneau $M_n(K)$ n'est pas bien mystérieux, puisqu'il s'identifie au corps commutatif K. Dès que $n \geqslant 2$, il est plein de pièges, principalement parce qu'il n'est ni commutatif ni intègre. Par exemple, on a $E_{2,2}E_{1,2} = 0$, d'où la non intégrité, puisque $E_{2,2}$ et $E_{1,2}$ sont non nulles ; et $E_{1,2}E_{2,2} = E_{1,2} \neq 0$, d'où la non commutativité.

Exercice 6.

Déterminer le centre de $M_n(K)$, c'est-à-dire le sous-anneau formé des matrices qui commutent avec toute matrice. (On explicitera les conditions $AE_{i,j} = E_{i,j}A$.)

Solution. Notons L_i, C_j les lignes et les colonnes de A.

On a déjà calculé $AE_{i,j} = [0; \dots; C_i; \dots; 0]$ (C_i en $j^{ème}$ position) et $E_{i,j}A = \begin{bmatrix} 0 \\ \vdots \\ L_j \\ \vdots \\ 0 \end{bmatrix}$ (L_j en $i^{ème}$ position). On en déduit que le seul coefficient non nul de C_i est $a_{i,i}$, que le seul coefficient non nul de L_j est $a_{j,j}$ et qu'ils sont égaux. Ainsi, les coefficients non diagonaux de A sont nuls et tous ses coefficients diagonaux sont égaux entre eux : elle est donc scalaire. Réciproquement, on a déjà vu qu'une matrice scalaire est centrale.

Exercice 7.

A-t-on $(A+B)(A-B) = A^2 - B^2$ et $(A+B)^2 = A^2 + 2AB + B^2$ dans $M_n(K)$?

Solution. En développant soigneusement, on trouve :

$$(A+B)(A-B) = A^2 + BA - AB - B^2 \quad \text{et} \quad (A+B)^2 = A^2 + AB + BA + B^2.$$

Les identités ci-dessus sont donc valables si, et seulement si, $AB = BA$.

Plus généralement, si A et B commutent, les identités classiques suivantes sont valables :

$$(A+B)^N = \sum_{i=0}^{N} \binom{N}{i} A^i B^{N-i} \quad \text{(formule du binôme de Newton)},$$

$$A^N - B^N = (A-B) \sum_{i=0}^{N-1} A^i B^{N-1-i} = \left(\sum_{i=0}^{N-1} A^i B^{N-1-i} \right) (A-B).$$

C'est en particulier le cas si A ou B est I_n.

Proposition et définition 6.
Nous dirons que A est *nilpotente d'ordre* N si $A^N = 0$. Nous dirons que B est *unipotente* si $B - I_n$ est nilpotente. Toute matrice unipotente est inversible.

Démonstration. Écrivons $B = I_n - A$ avec $(-A)^N = 0$, d'où $A^N = 0$. En vertu de la deuxième identité remarquable ci-dessus, on a :

$$I_n = I_n - A^N = (I_n - A)(I_n + A + \cdots + A^{N-1}) = (I_n + A + \cdots + A^{N-1})(I_n - A).$$

■

Par exemple, $(I_n + E_{i,j})^{-1} = I_n - E_{i,j}$ pour $i \neq j$ car $E_{i,j}^2 = 0$.

Lemme 7. Soient A et B deux matrices triangulaires supérieures (resp. inférieures) ; alors AB est triangulaire supérieure (resp. inférieure) et les coefficients diagonaux de AB s'obtiennent en multipliant les coefficients diagonaux correspondants de A et de B.

Démonstration. Traitons le cas où $A, B \in T_n(K)$, l'autre cas étant analogue (ou s'en déduisant par transposition). Si $i > j$, chacun des termes de la somme $c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}$ est tel que $k < i$ ou $j < k$ (car sinon, on aurait $i \leqslant k$ et $k \leqslant j$, donc $i \leqslant j$), donc tel que $a_{i,k} = 0$ ou $a_{k,j} = 0$, car $A, B \in T_n(K)$. Leur somme $c_{i,j}$ est donc nulle et $AB \in T_n(K)$. De plus, le seul terme éventuellement non nul de la somme $c_{i,i} = \sum_{k=1}^{n} a_{i,k} b_{k,i}$ est $a_{i,i} b_{i,i}$, pour la même raison. On a donc $c_{i,i} = a_{i,i} b_{i,i}$. ■

Proposition 8.
Les sous-ensembles $T_n(K)$, $\mathcal{L}_n(K)$ et $D_n(K)$ de $M_n(K)$ sont des sous-anneaux.

Démonstration. On sait déjà que ce sont des sous-groupes (et même des sous–espaces vectoriels) et il est clair qu'ils contiennent I_n. La stabilité pour la multiplication de $T_n(K)$ et $\mathcal{L}_n(K)$ est conséquence du lemme, et $D_n(K)$ est leur intersection. ■

Exercice 8.
Les sous-ensembles $S_n(K)$ et $A_n(K)$ de $M_n(K)$ sont-ils des sous-anneaux ?
Solution. Si car $K \neq 2$, $A_n(K)$ ne contient pas I_n et si car $K = 2$, $A_n(K) = S_n(K)$: on n'examinera donc que la stabilité par multiplication du sous-groupe $S_n(K)$. Si $n = 1$, $S_n(K) = M_n(K) = K$ et la réponse est triviale. Supposons donc $n \geqslant 2$. Alors les matrices symétriques $E_{1,1}$ et $E_{1,2} + E_{2,1}$ ont pour produit $E_{1,2}$ qui n'est pas symétrique ; $S_n(K)$ et $A_n(K)$ ne sont donc pas des sous-anneaux. En fait, si A et B sont symétriques, ${}^t(AB) = {}^tB\,{}^tA = BA$; dire que AB est symétrique est donc équivalent à dire que A et B commutent.

1.2.1 Le groupe linéaire

Rappelons (module I.1, section 4) que, dans un ensemble muni d'une loi associative et dotée d'un élément neutre e, tout élément a inversible à gauche (resp. à droite) est simplifiable à gauche (resp. à droite). Si un élément a admet à la fois un inverse à gauche et un inverse à droite, ceux-ci sont égaux. L'élément a est alors dit inversible. Une matrice $A \in M_n(K)$ est donc inversible (on dit aussi *régulière*) s'il existe $A' \in M_n(K)$ telle que $AA' = A'A = I_n$; son inverse A' est alors noté A^{-1}. Toute matrice inversible est donc inversible à gauche et à droite et simplifiable à gauche et à droite. Les réciproques de toutes ces implications sont vraies, bien que non triviales (corollaire 25 de la page 235). Grâce à la correspondance entre matrices et applications linéaires (module II.7), on pourra également le déduire du théorème 19 de la page 340.

Rappelons également (module II.2, section 3.1, p. 137) que l'ensemble des éléments inversibles d'un anneau est un groupe pour la multiplication de l'anneau.

Définition 5.
Le *groupe linéaire* $GL_n(K)$ est le groupe des éléments inversibles de l'anneau $M_n(K)$.

Si $n = 1$, ce groupe s'identifie à K^*. Si $n \geqslant 2$, il n'est jamais commutatif.
Par exemple, $I_n + E_{1,2}$ et $I_n + E_{2,1}$ sont inversibles (d'inverses $I_n - E_{1,2}$ et $I_n - E_{2,1}$) et ne commutent pas : $(I_n + E_{1,2})(I_n + E_{2,1}) = I_n + E_{1,2} + E_{2,1} + E_{1,1}$,

$$\text{alors que} \quad (I_n + E_{2,1})(I_n + E_{1,2}) = I_n + E_{1,2} + E_{2,1} + E_{2,2}.$$

Exemples.

1. Une matrice inversible ne peut avoir ni ligne ni colonne nulle. En effet, si par exemple la $i^{ème}$ ligne de A est nulle, il en est de même de toute matrice AB et l'on ne peut avoir $AB = I_n$. De même, si une colonne de A est nulle, on ne peut avoir $CA = I_n$.
2. La matrice diagonale $D := \mathrm{Diag}(a_1, \ldots, a_n)$ est inversible si, et seulement si, les a_i sont non nuls, et son inverse est alors $D' = \mathrm{Diag}(a_1^{-1}, \ldots, a_n^{-1})$.
   En effet, si $a_i = 0$ la $i^{ème}$ ligne de D est nulle et si les a_i sont tous non nuls, alors :
   $$DD' = D'D = \mathrm{Diag}(1, \ldots, 1) = I_n.$$
3. Si A est une matrice nilpotente d'ordre N, c'est-à-dire si $A^N = 0$, alors $I_n - A$ admet pour inverse $I_n + A + \cdots + A^{N-1}$.
4. La transposée d'une matrice inversible est inversible et ${}^t(A^{-1}) = ({}^tA)^{-1}$.
   En effet, Si $AB = BA = I_n$, alors ${}^tB{}^tA = {}^tA{}^tB = {}^tI_n = I_n$.

Définition 6. Les matrices A et B de $M_{m,n}(K)$ sont dites *équivalentes* s'il existe $P \in GL_m(K)$ et $Q \in GL_n(K)$ telles que $B = PAQ$.

Notons que, par définition, deux matrices équivalentes sont de même format. On vérifie facilement que la relation « être équivalentes » entre matrices est une relation d'équivalence ! Elle est réflexive : $A = I_m A I_n$. Elle est symétrique : si $B = PAQ$, alors $A = P^{-1}BQ^{-1}$. Elle est enfin transitive : si $B = PAQ$ et $C = RBS$, alors $C = (RP)A(QS)$ où, bien sûr, $RP \in GL_m(K)$ et $QS \in GL_n(K)$.
La seule matrice équivalente à la matrice $0 \in M_{m,n}(K)$ est 0 elle-même. À l'opposé, la classe d'équivalence de la matrice $I_n \in M_n(K)$ est $GL_n(K)$.
D'après le lemme 7 de la page 218, les coefficients diagonaux d'une matrice triangulaire inversible sont nécessairement non nuls. Nous allons démontrer la réciproque de ce fait. À titre auxiliaire, pour tout $p \in \mathbb{N}$, nous noterons N_p l'ensemble des matrices $A = (a_{i,j})_{1 \leqslant i,j \leqslant n} \in M_n(K)$ telles que $j - i < p \Rightarrow a_{i,j} = 0$. Par exemple, $N_0 = T_n(K)$, $N_1 = T_n^*(K)$ et $N_p = \{0\}$ pour $p \geqslant n$ (vérifiez ces affirmations !).

Lemme 9. Si $A \in N_p$ et $B \in N_q$, alors $AB \in N_{p+q}$.

Démonstration. Soient $A = (a_{i,j})_{1 \leqslant i,j \leqslant n}$, $B = (b_{i,j})_{1 \leqslant i,j \leqslant n}$ et $(c_{i,j})_{1 \leqslant i,j \leqslant n} = AB$.
Si $j - i < p + q$, tout terme $a_{i,k}b_{k,j}$ de la somme $c_{i,j} = \sum_{k=1}^{n} a_{i,k}b_{k,j}$ est tel que $k - i < p$ ou $j - k < q$ (car $k - i \geqslant p, j - k \geqslant q \Rightarrow j - i \geqslant p + q$), donc tel que $a_{i,k}b_{k,j} = 0$ (car $A \in N_p, B \in N_q$) ; on a donc $c_{i,j} = 0$. ∎

Lemme 10. Toute matrice triangulaire stricte est nilpotente d'ordre n.

Démonstration. Si $A \in T_n^*(K)$, alors $A \in N_1$, donc, d'après le lemme précédent, $A^k \in N_k$ pour tout $k \geqslant 1$ (par récurrence sur k) ; en particulier, $A^n = 0$. Si $A \in \mathcal{L}_n^*(K)$, alors ${}^tA \in T_n^*(K)$, donc ${}^t(A^n) = ({}^tA)^n = 0$. ■

Proposition 11. Pour qu'une matrice triangulaire (supérieure ou inférieure) soit inversible, il faut, et il suffit, que ses coefficients diagonaux soient non nuls.

Démonstration. On a déjà vu que la condition était nécessaire. Supposons que $A = D+N$ avec $D = \mathrm{Diag}(a_1, \ldots, a_n)$, les $a_i \in K^*$ et $N \in T_n^*(K)$ ou $\mathcal{L}_n^*(K)$. Alors D est inversible et $A = D(I_n - P)$, où $P = -D^{-1}N$ est triangulaire (supérieure ou inférieure) stricte, donc nilpotente d'ordre n. On a alors $A^{-1} = (I_n + P + \cdots + P^{n-1})D^{-1}$. ■

Exercice 9.

Calculer l'inverse de la matrice $A := \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 1 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$.

Solution. On a $A = I_n + P$, où $P = \sum_{i=1}^{n-1} E_{i,i+1}$.

On vérifie facilement que, pour $1 \leqslant k \leqslant n-1$, $P^k = \sum_{i=1}^{n-k} E_{i,i+k}$ (et $P^n = 0$). Donc :

$$A = I - P + P^2 + \cdots + (-1)^{n-1}P^{n-1} = \begin{pmatrix} 1 & -1 & 1 & \ldots & (-1)^{n-2} & (-1)^{n-1} \\ 0 & 1 & -1 & \ldots & (-1)^{n-3} & (-1)^{n-2} \\ 0 & 0 & 1 & \ldots & (-1)^{n-4} & (-1)^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

1.3 Matrices et applications linéaires

Il ne s'agira ici que d'applications linéaires de K^n dans K^m ; le cas de deux espaces vectoriels de dimension finie quelconques met en jeu des choix de bases, on l'étudiera au module II.7. Nous avons remarqué au module II.3 (section 2) que toute application linéaire de K^2 dans lui-même était de la forme $(x, y) \mapsto (ax + by, cx + dy)$ et nous lui avons associé la matrice $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Selon les définitions de 1.1, le lien est le suivant : $\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}$. C'est cette correspondance que nous allons généraliser.

Comme dans le module II.3, nous conviendrons de représenter les éléments de l'espace vectoriel K^n sous forme de vecteurs colonnes ; nous identifierons donc K^n à $M_{n,1}(K)$. Soit $A \in M_{m,n}(K)$. Pour tout vecteur colonne $X \in M_{n,1}(K)$, le produit AX est un vecteur colonne : $AX \in M_{m,1}(K)$:

$$\begin{array}{rccc} \Phi_A : & M_{n,1}(K) & \longrightarrow & M_{m,1}(K) \\ & X & \longmapsto & AX \end{array} .$$

Modulo l'identification ci-dessus, nous avons ainsi défini une application Φ_A de K^n dans K^m. En termes de composantes, le vecteur $(x_1, \ldots, x_n) \in K^n$ a pour image le vecteur $(a_{1,1}x_1 + \cdots + a_{1,n}x_n, \ldots, a_{m,1}x_1 + \cdots + a_{m,n}x_n) \in K^m$. En termes de matrices, on peut écrire :

$$\begin{pmatrix} a_{1,1} & \ldots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \ldots & a_{m,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{1,1}x_1 + \cdots + a_{1,n}x_n \\ \vdots \\ a_{m,1}x_1 + \cdots + a_{m,n}x_n \end{pmatrix} .$$

Proposition 12. L'application Φ_A est linéaire. Son image est le sous–espace vectoriel de $M_{m,1}(K)$ engendré par les colonnes de A.

Démonstration. La linéarité de Φ_A découle de la bilinéarité de la multiplication matricielle (corollaire 3 de la page 214). Par ailleurs, si $A = [C_1; \ldots; C_n]$, le vecteur $X \in M_{n,1}(K)$ de composantes $x_1, \ldots, x_n$ a pour image $x_1C_1 + \cdots + x_nC_n$ (équation (2) de la page 212) et l'image de Φ_A est l'ensemble des vecteurs de cette forme. ■

Corollaire 13. Les conditions suivantes sont équivalentes :
(i) L'application Φ_A est surjective.
(ii) Les colonnes de A engendrent $M_{m,1}(K)$.
(iii) Il existe une matrice $B \in M_{n,m}(K)$ telle que $AB = I_m$.

Démonstration.
L'équivalence de (i) et de (ii) est une conséquence immédiate de la proposition 12. Supposons que (i) soit vraie. Alors les vecteurs $E_1, \ldots, E_m$ de la base canonique de $M_{m,1}(K)$ sont dans l'image de Φ_A : notons $E_i = AX_i$. Soit $B = [X_1; \ldots; X_m] \in M_{n,m}(K)$. On a $AB = [AX_1; \ldots; AX_m] = [E_1; \ldots; E_m] = I_m$.
Réciproquement, si $AB = I_m$, tout $Y \in M_{m,1}(K)$ est égal à $A(BY) = \Phi_A(X)$ avec $X = BY \in M_{n,1}(K)$. ■

Remarque. Le vecteur $X \in M_{n,1}(K)$ est dans le noyau de Φ_A si, et seulement si, $x_1C_1 + \cdots + x_nC_n = 0$. Pour que Φ_A soit injective, il est donc nécessaire et suffisant que les C_j soient linéairement indépendantes.

Théorème 14. L'application $A \mapsto \Phi_A$ de $M_{m,n}(K)$ dans $\mathcal{L}_K(K^n, K^m)$ est un isomorphisme d'espaces vectoriels.

Pour résumer, toute application linéaire de K^n dans K^m est codée par une matrice $A \in M_{m,n}(K)$ et cette correspondance est compatible avec les combinaisons linéaires.

Démonstration.
L'application $\Phi_{\lambda A+\lambda' A'}$ associe à $X \in M_{n,1}(K)$ le vecteur $(\lambda A+\lambda' A')X = \lambda AX+\lambda' A'X$. C'est donc l'application $\lambda\Phi_A + \lambda'\Phi_{A'}$, et $A \mapsto \Phi_A$ est bien linéaire. Si Φ_A est l'application linéaire nulle, son image est triviale ; comme elle contient les colonnes de A, celles-ci sont nulles et $A = 0$, d'où l'injectivité. Montrons enfin la surjectivité. Toute application linéaire $\varphi : K^n \to K^m$ est déterminée par son effet sur la base canonique de K^n. Notons donc $(a_{1,j}, \ldots, a_{m,j}) \in K^m$ l'image par φ du $j^{ème}$ vecteur de cette base canonique ($1 \leqslant j \leqslant n$). L'image par φ de $(x_1, \ldots, x_n) \in K^n$ est alors le vecteur $(a_{1,1}x_1 + \cdots + a_{1,n}x_n, \ldots, a_{m,1}x_1 + \cdots + a_{m,n}x_n) \in K^m$, et l'on voit que $\varphi = \Phi_A$ avec $A = (a_{i,j})_{1\leqslant i,j\leqslant n}$. ∎

Qu'en est-il de la multiplication ?
Soient $A \in M_{m,n}(K)$, $B \in M_{n,p}(K)$ et $C = AB \in M_{m,p}(K)$. De l'associativité du produit matriciel, on déduit que, si $X \in M_{n,1}(K)$:

$$\Phi_C(X) = (AB)X = A(BX) = \Phi_A\big(\Phi_B(X)\big),$$

c'est-à-dire $\Phi_C = \Phi_A \circ \Phi_B$. Ainsi, par la correspondance précédente, *la multiplication des matrices correspond à la composition des applications linéaires.*

Théorème 15.
L'application $A \mapsto \Phi_A$ de $M_n(K)$ dans $\mathcal{L}_K(K^n)$ est un isomorphisme d'anneaux.

Démonstration. Il nous reste seulement à vérifier qu'elle fait se correspondre les éléments neutres de la multiplication et de la composition : si $A = I_n$, il faut voir que $\Phi_A = \mathrm{Id}_{K^n}$, ce qui est évident puisque $I_nX = X$. ∎

Application : inversion d'une matrice triangulaire.. Soit $A = (a_{i,j})_{1\leqslant i,j\leqslant n}$ une matrice triangulaire supérieure (le cas d'une matrice triangulaire inférieure se traite exactement de la même manière). On a vu qu'elle est inversible si, et seulement si, ses coefficients diagonaux $a_{1,1}, \ldots, a_{n,n}$ sont non nuls. Pour calculer A^{-1}, nous partirons de la remarque que l'application linéaire $Y \mapsto A^{-1}Y$ est la réciproque de l'application linéaire $X \mapsto AX$. Nous sommes ramenés à résoudre un système :

$$\begin{array}{rrrrl} a_{1,1}x_1+ & a_{1,2}x_2+ & \cdots+ & a_{1,n}x_n = & y_1 \\ & a_{2,2}x_2+ & \cdots+ & a_{2,n}x_n = & y_2 \\ & & & \vdots & \\ & & & a_{n,n}x_n = & y_n \end{array}$$

On le résout de bas en haut :

$$\begin{array}{rcl} x_n & = & a_{n,n}^{-1}\, y_n \\ x_{n-1} & = & a_{n-1,n-1}^{-1}\,(y_n - a_{n-1,n}x_n) \\ & \vdots & \\ x_1 & = & a_{1,1}^{-1}\,(y_1 - a_{1,2}x_2 - \cdots - a_{1,n}x_n) \end{array}$$

On obtient ainsi des expressions $x_i = \sum\limits_{j=1}^{n} b_{i,j}y_j$ et $A^{-1} = (b_{i,j})_{1\leqslant i,j\leqslant n}$.

Exercice 10.

Calculer l'inverse de la matrice $A = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$ par cette méthode.

Solution. Le système associé est : $\begin{cases} x_1 + x_2 = y_1 \\ x_2 + x_3 = y_2 \\ \vdots \\ x_{n-1} + x_n = y_{n-1} \\ x_n = y_n \end{cases}$

On trouve donc $x_n = y_n$, $x_{n-1} = y_{n-1} - y_n$, …, $x_1 = y_1 - y_2 + \cdots + (-1)^{n-1} y_n$, d'où la matrice $\begin{pmatrix} 1 & -1 & 1 & \dots & (-1)^{n-2} & (-1)^{n-1} \\ 0 & 1 & -1 & \dots & (-1)^{n-3} & (-1)^{n-2} \\ 0 & 0 & 1 & \dots & (-1)^{n-4} & (-1)^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$.

Définition 7. On appelle *rang* de la matrice A, et l'on note $\operatorname{rg} A$ la dimension du sous–espace vectoriel engendré par les colonnes de A.

D'après la proposition 12 de la page 222, ce sous–espace vectoriel est également l'image de l'application Φ_A. Des propriétés générales de la dimension (module II.7, proposition 18 de la page 340), on déduit les deux inégalités suivantes :

- $\operatorname{rg}(A) \leqslant n$, avec égalité si, et seulement si, les colonnes de A sont linéairement indépendantes ; cela équivaut à l'injectivité de Φ_A.
- $\operatorname{rg}(A) \leqslant m$, avec égalité si, et seulement si, les colonnes de A engendrent $M_{m,1}(K)$; cela équivaut à la surjectivité de Φ_A.

Proposition 16. Des matrices équivalentes ont même rang.

Démonstration. Soient $A \in M_{m,n}(K)$, $P \in GL_m(K)$ et $Q \in GL_n(K)$. Il suffit évidemment de démontrer que $\operatorname{rg}(AQ) = \operatorname{rg} A = \operatorname{rg}(PA)$. Les colonnes de AQ sont des combinaisons linéaires de celles de A ; la réciproque est vraie, puisque $A = (AQ)Q^{-1}$. L'égalité $\operatorname{rg} A = \operatorname{rg}(AQ)$ s'en déduit immédiatement. Il est plus malaisé de démontrer que $\operatorname{rg} A = \operatorname{rg}(PA)$ par calcul direct. Nous raisonnerons donc sur l'espace image de $\Phi_{PA} = \Phi_P \circ \Phi_A$: il est l'image par l'isomorphisme Φ_P de l'espace image de Φ_A ; ces deux espaces images ont donc même dimension. ∎

Rang d'une matrice échelonnée en colonnes. Soit une matrice A échelonnée en colonnes : ses colonnes $C_1, \dots, C_k$ ont pour pivots $a_{i_1,1}, \dots, a_{i_k,k}$ (donc $1 \leqslant i_1 < \cdots < i_k \leqslant n$) et $C_j = 0$ pour $k < j \leqslant n$. Alors *le rang de A est k*.
Il suffit en effet de démontrer que les vecteurs colonnes $C_1, \dots, C_k$ sont linéairement indépendants (puisque les autres colonnes sont nulles). Supposons donc une relation linéaire $\lambda_1 C_1 + \cdots + \lambda_k C_k = 0$. Pour $j \geqslant 2$, les coefficients non nuls de C_j sont d'indice $k \geqslant i_j \geqslant i_2 > i_1$. Le coefficient d'indice i_1 de $\lambda_1 C_1 + \cdots + \lambda_k C_k$ est donc $\lambda_1 a_{i_1,1}$, et il est nul par hypothèse ; comme $a_{i_1,1} \neq 0$, on a $\lambda_1 = 0$. On recommence alors le même raisonnement à partir de la relation $\lambda_2 C_2 + \cdots + \lambda_k C_k = 0$, et l'on prouve successivement que $\lambda_2, \dots, \lambda_k$ sont nuls.

Rang d'une matrice échelonnée en lignes. Soit une matrice A échelonnée en lignes : ses lignes $L_1, \dots, L_k$ ont pour pivots $a_{1,j_1}, \dots, a_{k,j_k}$ (donc $1 \leqslant j_1 < \cdots < j_k \leqslant n$) et $L_i = 0$ pour $k < i \leqslant m$. Alors *le rang de A est k*.
La preuve est plus compliquée, et nous en donnerons seulement les arguments essentiels. On remarque tout d'abord que le dernier coefficient non nul de chaque « colonne de pivot » C_{j_i} ($1 \leqslant i \leqslant k$) a pour indice i (car toutes les lignes d'indice $> i$ sont plus courtes et calées à droite). Le même argument que dans la proposition précédente s'applique et permet de voir que les colonnes de pivot $C_{j_1}, \dots, C_{j_k}$ sont linéairement indépendantes. De plus, notant temporairement $E_1, \dots, E_m$ la base canonique de l'espace $M_{m,1}(K)$ des colonnes, on voit de la même manière que $\mathrm{Vect}(C_{j_1}, \dots, C_{j_k}) = \mathrm{Vect}(E_1, \dots, E_k)$. Comme, pour toute colonne C_j de A, les seuls coefficients non nuls de C_j sont d'indice $\leqslant k$ (les lignes au delà de L_k étant nulles), $C_j \in \mathrm{Vect}(C_{j_1}, \dots, C_{j_k})$. L'espace engendré par les colonnes est donc de dimension k.

Exemple. Les colonnes de pivot de la matrice $\begin{pmatrix} 0 & \underline{\star} & \star & \star & \star & \star & \star & \star \\ 0 & 0 & 0 & \underline{\star} & \star & \star & \star & \star \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \underline{\star} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ sont C_2, C_4 et C_8. Elles sont linéairement indépendantes et engendrent le sous–espace vectoriel $\mathrm{Vect}(E_1, E_2, E_3)$, auquel appartiennent également C_1, C_3, C_5, C_6, C_7.

On voit donc que, pour une matrice A échelonnée en lignes ou en colonnes, $\mathrm{rg}\, A = \mathrm{rg}\, {}^tA$. On démontrera plus loin que c'est vrai pour toute matrice (corollaire 23 de la page 234).

2 Opérations élémentaires et algorithmes de Gauß

Dans cette section, nous adoptons un point de vue résolument algorithmique, ce qui aura quelques conséquences sur notre vocabulaire. Notre activité principale consistera à *transformer* des matrices. Cela signifie, par exemple, que l'on partira d'une matrice A. Par un calcul algébrique, on en déduira une matrice A'. On dira alors que l'on a transformé A en A', puis, si cela simplifie le discours, *c'est cette dernière que l'on appellera A*.

Exemple. Si l'on multiplie à gauche $A = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}$ par $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, elle est transformée en la nouvelle valeur $A = \begin{pmatrix} d & e & f \\ a & b & c \end{pmatrix}$.

Remarque. Ces formules apparemment contradictoires signifient que nous considérons ici le membre gauche de l'égalité, A, comme une variable *au sens de l'informatique*, autrement dit, comme une « case » susceptible de contenir successivement des « valeurs » qui varient au cours du temps. Le symbole d'égalité ne désigne alors plus une relation symétrique, ni même transitive : le membre gauche désigne toujours cette variable susceptible de transformations et le membre droit une valeur prise par cette variable. Voir là-dessus le module II.8.

2.1 Opérations élémentaires sur les lignes et sur les colonnes d'une matrice

2.1.1 Définition des opérations élémentaires

Soient $A = \begin{bmatrix} L_1 \\ \vdots \\ L_m \end{bmatrix}$ et $A' = \begin{bmatrix} L_1 \\ \vdots \\ L'_m \end{bmatrix}$ deux matrices de $M_{m,n}(K)$. Nous dirons que A' a été obtenue à partir de A par :

- *permutation* ou *échange* des lignes L_i et L_j (avec $i, j \in [\![1, m]\!]$ et $i \neq j$) si $L'_i = L_j$, $L'_j = L_i$ et $L'_k = L_k$ pour $k \in [\![1, m]\!] \setminus \{i, j\}$; $\boxed{\text{notation : } L_i \leftrightarrow L_j}$
- *addition à L_i de αL_j* (avec $i, j \in [\![1, m]\!]$, $i \neq j$ et $\alpha \in K$) si $L'_i = L_i + \alpha L_j$ et $L'_k = L_k$ pour $k \in [\![1, m]\!] \setminus \{i\}$ (on parle alors de *transvection sur les lignes*) ;

$$\boxed{\text{notation : } L_i \leftarrow L_i + \alpha L_j}$$

- *multiplication de L_i par λ* (avec $i \in [\![1, m]\!]$ et $\lambda \in K^*$) si $L'_i = \lambda L_i$ et $L'_k = L_k$ pour $k \in [\![1, m]\!] \setminus \{i\}$ (on parle alors de *dilatation sur les lignes*) ; $\boxed{\text{notation : } L_i \leftarrow \lambda L_i.}$

Ces opérations sont appelées *opérations élémentaires sur les lignes de A*.

La dilatation $L_i \leftarrow \lambda L_i$ n'est autorisée que si $\lambda \neq 0$.

Exercice 11.

Soient $a, b, c \in K$. Transformer la matrice $\begin{pmatrix} 1 & a & a^2 \\ 1 & b & b^2 \\ 1 & c & c^2 \end{pmatrix}$ en une matrice triangulaire à l'aide d'opérations élémentaires sur les lignes.

Solution.

$$\begin{pmatrix}1 & a & a^2\\ 1 & b & b^2\\ 1 & c & c^2\end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - L_1} \begin{pmatrix}1 & a & a^2\\ 0 & b-a & b^2-a^2\\ 1 & c & c^2\end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - L_1} \begin{pmatrix}1 & a & a^2\\ 0 & b-a & b^2-a^2\\ 0 & c-a & c^2-a^2\end{pmatrix}$$

Supposons $a \neq b$:

$$\begin{pmatrix}1 & a & a^2\\ 0 & b-a & b^2-a^2\\ 0 & c-a & c^2-a^2\end{pmatrix} \xrightarrow{L_3 \leftarrow L_3 - \frac{c-a}{b-a}L_2} \begin{pmatrix}1 & a & a^2\\ 0 & b-a & b^2-a^2\\ 0 & 0 & (c-a)(c-b)\end{pmatrix}$$

Supposons $a = b$:

$$\begin{pmatrix}1 & a & a^2\\ 0 & b-a & b^2-a^2\\ 0 & c-a & c^2-a^2\end{pmatrix} = \begin{pmatrix}1 & a & a^2\\ 0 & 0 & 0\\ 0 & c-a & c^2-a^2\end{pmatrix} \xrightarrow{L_2 \leftrightarrow L_3} \begin{pmatrix}1 & a & a^2\\ 0 & c-a & c^2-a^2\\ 0 & 0 & 0\end{pmatrix}$$

Remarque. Dans le processus ci-dessus, les deux premières opérations

$$(L_2 \leftarrow L_2 - L_1 \quad \text{et} \quad L_3 \leftarrow L_3 - L_1)$$

sont permutables ; dans ce cas, on peut les exécuter simultanément et écrire directement :

$$\begin{pmatrix}1 & a & a^2\\ 1 & b & b^2\\ 1 & c & c^2\end{pmatrix} \xrightarrow[L_3 \leftarrow L_3 - L_1]{L_2 \leftarrow L_2 - L_1} \begin{pmatrix}1 & a & a^2\\ 0 & b-a & b^2-a^2\\ 0 & c-a & c^2-a^2\end{pmatrix}$$

Ce n'est pas toujours aussi simple : le lecteur vérifiera que les opérations $L_2 \leftarrow L_2 - L_1$ et $L_3 \leftarrow L_3 - L_2$, par exemple, ne sont pas permutables.

Soient $A = [C_1; \ldots; C_m]$ et $A' = [C'_1; \ldots; C'_m]$ deux matrices de $M_{m,n}(K)$. Nous dirons que A' a été obtenue à partir de A par :

- *permutation* ou *échange* des colonnes C_i et C_j (avec $i, j \in [\![1, m]\!]$ et $i \neq j$) si $C'_i = C_j$, $C'_j = C_i$ et $C'_k = C_k$ pour $k \in [\![1, m]\!] \setminus \{i, j\}$; $\boxed{\text{notation : } C_i \leftrightarrow C_j}$
- *addition à C_i de αC_j* (avec $i, j \in [\![1, m]\!]$, $i \neq j$ et $\alpha \in K$) si $C'_i = C_i + \alpha C_j$ et $C'_k = C_k$ pour $k \in [\![1, m]\!] \setminus \{i\}$ (on parle alors de *transvection sur les colonnes*) ;

$$\boxed{\text{notation : } C_i \leftarrow C_i + \alpha C_j}$$

- *multiplication de C_i par λ* (avec $i \in [\![1, m]\!]$ et $\lambda \in K^*$) si $C'_i = \lambda C_i$ et $C'_k = C_k$ pour $k \in [\![1, m]\!] \setminus \{i\}$ (on parle alors de *dilatation sur les colonnes*) ; $\boxed{\text{notation : } C_i \leftarrow \lambda C_i.}$

Ces opérations sont appelées *opérations élémentaires sur les colonnes de A*. Il revient donc au même d'effectuer des opérations élémentaires sur les lignes de M ou d'effectuer des opérations élémentaires sur les colonnes de ${}^tM$.

La dilatation $C_i \leftarrow \lambda C_i$ n'est autorisée que si $\lambda \neq 0$.

Exercice 12.

Transformer la matrice triangulaire obtenue plus haut en une matrice diagonale par une suite d'opérations élémentaires sur les colonnes.

Solution.

$$\begin{pmatrix} 1 & a & a^2 \\ 0 & b-a & b^2-a^2 \\ 0 & 0 & (c-a)(c-b) \end{pmatrix} \xrightarrow[]{\substack{C_2 \leftarrow C_2 - aC_1 \\ C_3 \leftarrow C_3 - a^2 C_1}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & b-a & b^2-a^2 \\ 0 & 0 & (c-a)(c-b) \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & b-a & b^2-a^2 \\ 0 & 0 & (c-a)(c-b) \end{pmatrix} \xrightarrow[]{C_3 \leftarrow C_3 - (a+b)C_2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & b-a & 0 \\ 0 & 0 & (c-a)(c-b) \end{pmatrix}$$

Remarquons que la dernière opération n'était pas permutable avec les précédentes.

Introduisons maintenant trois types de matrices carrées particuliers :

- Les *matrices d'échange* $S_{i,j}^{(n)} := I_n - E_{i,i}^{(n,n)} - E_{j,j}^{(n,n)} + E_{i,j}^{(n,n)} + E_{j,i}^{(n,n)}$ (avec $i, j \in [\![1, n]\!]$ et $i \neq j$) :

$$S_{i,j}^{(n)} = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \iddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix}$$

- Les *matrices de transvection* $T_{i,j}^{(n)}(\alpha) := I_n + \alpha E_{i,j}^{(n,n)}$ (avec $i, j \in [\![1, n]\!]$, $i \neq j$ et $\alpha \in K$) :

$$T_{i,j}^{(n)}(\alpha) = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & \dots & \alpha & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix}$$

- Les *matrices de dilatation* $D_i^{(n)}(\lambda) := I_n + (\lambda - 1)E_{i,i}^{(n,n)}$ (avec $i \in [\![1, n]\!]$ et $\lambda \in K^*$) :

$$D_i^{(n)}(\lambda) = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix}$$

Toutes ces matrices sont inversibles (*i.e.* inversibles à droite et à gauche) :

$$\left(S_{i,j}^{(n)}\right)^{-1} = S_{i,j}^{(n)}$$

$$\left(T_{i,j}^{(n)}(\alpha)\right)^{-1} = T_{i,j}^{(n)}(-\alpha)$$

$$\left(D_i^{(n)}(\lambda)\right)^{-1} = D_i^{(n)}(\lambda^{-1})$$

On vérifie alors sans peine que la matrice obtenue à partir de A par

- l'opération élémentaire $L_i \leftrightarrow L_j$ est la matrice $S_{i,j}^{(m)}A$,
- l'opération élémentaire $L_i \leftarrow L_i + \alpha L_j$ est la matrice $T_{i,j}^{(m)}(\alpha)A$,
- l'opération élémentaire $L_i \leftarrow \lambda L_i$ est la matrice $D_i^{(m)}(\lambda)A$,
- l'opération élémentaire $C_i \leftrightarrow C_j$ est la matrice $AS_{i,j}^{(n)}$,
- l'opération élémentaire $C_i \leftarrow C_i + \alpha C_j$ est la matrice $AT_{j,i}^{(n)}(\alpha)$,
- l'opération élémentaire $C_i \leftarrow \lambda C_i$ est la matrice $AD_i^{(n)}(\lambda)$.

La matrice A étant rectangulaire, les tailles des matrices carrées spéciales par lesquelles on la multiplie ne sont pas les mêmes selon que l'on multiplie à gauche ou à droite.

Proposition 17. Les opérations sur les lignes et les colonnes produisent des matrices équivalentes. Elles conservent le rang. Elles conservent également les propriétés suivantes : simplifiabilité à droite (resp. à gauche), inversibilité à droite (resp. à gauche).

Démonstration. C'est immédiat à partir des remarques précédentes. ∎

Exemple. La matrice $\begin{pmatrix} 1 & a & a^2 \\ 1 & b & b^2 \\ 1 & c & c^2 \end{pmatrix}$ est inversible si, et seulement si, a, b, c sont deux à deux distincts. Son rang est le nombre de valeurs distinctes parmi ces trois scalaires. Cela découle des exemples précédents.

2.2 Algorithmes de Gauß : lignes seules

Proposition 18. À l'aide d'échanges et de transvections sur les lignes, on peut transformer toute matrice $A \in M_{m,n}(K)$ en une matrice échelonnée en lignes.

Démonstration. Nous procéderons par récurrence sur $\inf(m,n)$. Si $m = 1$, A est une matrice ligne et il n'y a rien à faire. Si $n = 1$, c'est une matrice colonne. Si elle est nulle, il n'y a rien à faire. Sinon, une opération $L_1 \leftrightarrow L_i$ fait venir en première position un *pivot* non nul a_1 ; puis des opérations simultanées $L_i \leftarrow L_i - \dfrac{a_i}{a_1} L_1$ (pour $i \geqslant 2$) annulent tous les autres coefficients, et l'on a fini. Nous supposons maintenant que $m, n \geqslant 2$. Comme première étape, on examine la première colonne. Si elle est nulle, la première étape est finie. Sinon, une opération $L_1 \leftrightarrow L_i$ fait venir en première position un *pivot* non nul $a_{1,1}$; puis des opérations simultanées $L_i \leftarrow L_i - \dfrac{a_{i,1}}{a_{1,1}} L_1$ (pour $i \geqslant 2$) annulent tous les autres coefficients de la première colonne. À l'issue de la première étape, on a dans tous les cas une matrice de la forme suivante :

$$\begin{pmatrix} a_{1,1} & \star & \cdots & \star \\ 0 & b_{1,1} & \cdots & b_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & b_{m-1,1} & \cdots & b_{m-1,n-1} \end{pmatrix}$$

Par hypothèse de récurrence, des opérations du type indiqué permettent de transformer la matrice $\begin{pmatrix} b_{1,1} & \cdots & b_{1,n-1} \\ \vdots & \ddots & \vdots \\ b_{m-1,1} & \cdots & b_{m-1,n-1} \end{pmatrix}$ en une matrice échelonnée en lignes, donc également notre matrice. ∎

Corollaire 19. Toute matrice s'écrit sous la forme BC, où B est un produit de matrices d'échange et de matrices de transvection, et C une matrice échelonnée en lignes.

2.2.1 Application aux systèmes d'équations

Considérons le système :

$$\begin{cases} a_{1,1}x_1 + \cdots + a_{1,n}x_n = b_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m,1}x_1 + \cdots + a_{m,n}x_n = b_m \end{cases}$$

On peut l'écrire sous forme matricielle $AX = B$, avec $A \in M_{m,n}(K)$, $B \in M_{m,1}(K)$ et $X \in M_{n,1}(K)$. Soit $A' = PAQ$, où $P \in GL_m(K)$ et $Q \in GL_n(K)$, une matrice équivalente à A. Le système ci-dessus équivaut au système $A'X' = B'$, où $B' = PB$ et où $X = QX'$. Pratiquement, remplacer le système $AX = B$ par le système $(PA)X = (PB)$ revient à effectuer des combinaisons linéaires sur les équations de départ

pour obtenir un système équivalent en l'inconnue X ; alors que le remplacer par le système $A'X' = B'$ revient à effectuer des combinaisons linéaires sur les coefficients de X pour remplacer l'inconnue X par l'inconnue X', qui lui est équivalente parce que l'on peut passer de l'une à l'autre dans les deux sens : $X = QX'$ et $X' = Q^{-1}X$.

L'algorithme décrit plus haut montre que l'on peut se ramener, par des combinaisons linéaires simples sur les équations, à un système échelonné en lignes, donc de la forme suivante :

$$\begin{cases} a_{1,j_1}x_{j_1} + \sum\limits_{j>j_1} a_{1,j}x_j = b_1 \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ a_{k,j_k}x_{j_k} + \sum\limits_{j>j_k} a_{k,j}x_j = b_k \\ 0 = b_{k+1} \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ 0 = b_m \end{cases}$$

Ici, les coefficients pivots $a_{1,j_1}, \dots, a_{k,j_k}$ sont non nuls. Si $b_{k+1}, \dots, b_n$ sont non tous nuls, il n'y a pas de solution à ce système. Si $k = m$ ou si $b_{k+1}, \dots, b_n$ sont tous nuls, il y a des solutions et l'on peut toutes les trouver en résolvant du bas en haut :

1. Les x_j pour $j > j_k$ sont choisis arbitrairement et l'on pose :
$$x_{j_k} := \frac{1}{a_{k,j_k}}\Big(b_k - \sum_{j>j_k} a_{k,j}x_j\Big),$$
2. Pour chaque indice i de $k-1$ à 1 (dans cet ordre), les x_j pour $j_{i+1} > j > j_i$ sont choisis arbitrairement et l'on pose :
$$x_{j_i} := \frac{1}{a_{i,j_i}}\Big(b_i - \sum_{j>j_i} a_{i,j}x_j\Big).$$
3. Les x_j pour $j < j_1$ sont choisis arbitrairement.

Le nombre d'inconnues x_j choisies arbitrairement est :

$$(n - j_k) + \cdots + (j_k - j_{k-1} - 1) + \cdots + (j_2 - j_1 - 1) + (j_1 - 1) = n - k.$$

Les autres inconnues s'expriment linéairement en fonction de celles-ci et des b_k : comme on le verra au module II.7 (théorème 64 de la page 377), cela signifie que les solutions forment un *espace affine de dimension* $n - k$, et c'est lié au fait que $\operatorname{rg} A = k$ (comme on l'a prouvé précédemment).

Exemple. Nous supposons a, b, c deux à deux distincts. Le système :

$$\begin{cases} x + ay + a^2z = u \\ x + by + b^2z = v \\ x + cy + c^2z = w \end{cases}$$

est transformé par les opérations déjà vues en un système équivalent. Nous le coderons par la *matrice augmentée* :

$$\begin{pmatrix} 1 & a & a^2 & u \\ 1 & b & b^2 & v \\ 1 & c & c^2 & w \end{pmatrix}$$

et nous effectuerons les opérations sur les lignes sur cette matrice augmentée :

$$\begin{pmatrix} 1 & a & a^2 & u \\ 1 & b & b^2 & v \\ 1 & c & c^2 & w \end{pmatrix} \xrightarrow[]{\substack{L_2 \leftarrow L_2 - L_1 \\ L_3 \leftarrow L_3 - L_1}} \begin{pmatrix} 1 & a & a^2 & u \\ 0 & b-a & b^2-a^2 & v-u \\ 0 & c-a & c^2-a^2 & w-u \end{pmatrix}$$

$$\begin{pmatrix} 1 & a & a^2 & u \\ 0 & b-a & b^2-a^2 & v-u \\ 0 & c-a & c^2-a^2 & w-u \end{pmatrix} \xrightarrow[]{L_3 \leftarrow L_3 - \dfrac{c-a}{b-a} L_2}$$

$$\begin{pmatrix} 1 & a & a^2 & u \\ 0 & b-a & b^2-a^2 & v-u \\ 0 & 0 & (c-a)(c-b) & w-u-\dfrac{c-a}{b-a}(v-u) \end{pmatrix}$$

Le système de départ est donc équivalent au suivant :

$$\begin{cases} 1x + ay + a^2 z = u \\ 0x + (b-a)y + (b^2-a^2)z = v-u \\ 0x + 0y + (c-a)(c-b)z = w-u-\dfrac{c-a}{b-a}(v-u) \end{cases}$$

On peut alors résoudre ce système. On trouve successivement :

$$\begin{aligned} z &= \frac{u}{(c-a)(b-a)} + \frac{v}{(a-b)(c-b)} + \frac{w}{(a-c)(b-c)} \\ y &= -\frac{(b+c)u}{(c-a)(b-a)} - \frac{(a+c)v}{(a-b)(c-b)} - \frac{(a+b)w}{(a-c)(b-c)} \\ x &= \frac{cbu}{(c-a)(b-a)} + \frac{cav}{(a-b)(c-b)} + \frac{abw}{(a-c)(b-c)} \end{aligned}$$

Remarquons qu'en posant $P(T) = x + yT + zT^2$, on est ramené à chercher un polynôme P de degré inférieur ou égal à 2 tel que $P(a) = u$, $P(b) = v$, $P(c) = w$. Nous verrons au module II.6 (théorème 23 de la page 298) que ce problème d'*interpolation polynomiale* admet pour unique solution :

$$P(T) = u\frac{(T-b)(T-c)}{(a-b)(a-c)} + v\frac{(T-a)(T-c)}{(b-a)(b-c)} + w\frac{(T-a)(T-b)}{(c-a)(c-b)}.$$

Les coefficients de ce polynôme P sont x, y, z.

2.3 Algorithmes de Gauß : lignes et colonnes

Il serait encore plus simple de transformer une matrice quelconque en matrice échelonnée en lignes et en colonnes ; par exemple, transformer une matrice carrée en matrice diagonale. Cependant, cela n'est pas toujours possible avec les seules opérations sur les lignes (comme pourrait le faire croire l'exemple précédent) : essayez avec la matrice $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$.

Proposition 20. À l'aide d'échanges et de transvections sur les lignes et les colonnes, on peut transformer toute matrice $A \in M_{m,n}(K)$ en une matrice échelonnée en lignes et en colonnes.

Démonstration. À l'aide des seuls échanges et transvections sur les lignes, on peut se ramener au cas où A est échelonnée en lignes, ce que nous supposons. Il y a donc des indices $j_1 > \cdots > j_k$ tels que les pivots $a_{1,j_1}, \ldots, a_{k,j_k}$ sont non nuls, que $a_{i,j} = 0$ pour $i \leqslant k$ et $j < j_i$ et pour $i > k$ et j quelconque. On effectue alors, dans l'ordre, les transvections suivantes :

1. $C_j \leftarrow C_j - \dfrac{a_{1,j}}{a_{1,j_1}} C_{j_1}$, pour tous les $j > j_1$; ces transvections sont permutables et peuvent donc être simultanées. Après leur exécution, le seul coefficient non nul sur la première ligne est le pivot a_{1,j_1}.

2. $C_j \leftarrow C_j - \dfrac{a_{2,j}}{a_{2,j_2}} C_{j_2}$, pour tous les $j > j_2$; ces transvections sont permutables et peuvent donc être simultanées. Après leur exécution, le seul coefficient non nul sur la deuxième ligne est le pivot a_{2,j_2} ; de plus, la première ligne n'a pas été modifiée par ces opérations.

3. Etc., pour toutes les lignes d'indices $\leqslant k$.

À l'issue de toutes ces opérations, chacune des k premières lignes a un seul coefficient non nul et leurs indices de colonnes vont strictement croissant. On fait alors, dans cet ordre, les échanges $C_1 \leftrightarrow C_{j_1}$ (qui ramène le pivot de la première ligne en première colonne), puis $C_2 \leftrightarrow C_{j_2}$ (qui ramène le pivot de la deuxième ligne en deuxième colonne), etc. ∎

À l'issue de ces transformations, on a obtenu une matrice de la forme :

$$A = \sum_{i=1}^{k} E_{i,i}^{(m,n)} = \begin{pmatrix} a_{1,1} & \ldots & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \ldots & a_{k,k} & \ldots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \ldots & 0 \end{pmatrix}, \text{ avec } a_{1,1}, \ldots, a_{k,k} \neq 0.$$

Théorème 21. Toute matrice A de $M_{m,n}(K)$ est équivalente à une matrice de la forme suivante et à une seule :

$$I_r^{(m,n)} := \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \end{pmatrix} = \sum_{i=1}^{r} E_{i,i} = \begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

pour un certain $r \in [\![0, \min(m,n)]\!]$. Dans la dernière écriture, les *blocs* I_r, $0_{r,n-r}$, $0_{m-r,r}$ et $0_{m-r,n-r}$ désignent respectivement la matrice identité d'ordre r et les matrices nulles de types $(r, n-r)$, $(m-r, r)$ et $(m-r, n-r)$.

Démonstration. On peut partir de la forme obtenue précédemment. Pour chaque indice $i \in [\![1, k]\!]$, on effectue alors, au choix, la dilatation $L_i \leftarrow \dfrac{1}{a_{i,i}} L_i$ ou la dilatation $C_i \leftarrow \dfrac{1}{a_{i,i}} C_i$. Cela montre l'existence d'un $r \in [\![0, \min(m,n)]\!]$ tel que A est équivalente à $I_r^{(m,n)}$. Le couple (m,n) est nécessairement le type de A et l'entier r est nécessairement le rang de A (puisque l'on sait déjà que $\operatorname{rg} I_r^{(m,n)} = r$). ■

Une preuve géométrique de ce théorème sera donnée au module II.7 (théorème 34 de la page 351).

Corollaire 22.
Deux matrices de $M_{m,n}(K)$ sont équivalentes si, et seulement si, elles ont même rang.

Démonstration. On sait déjà que deux matrices équivalentes ont même rang. Il découle du théorème que deux matrices de même rang r de $M_{m,n}(K)$ sont équivalentes à $I_r^{(m,n)}$, donc entre elles. ■

Corollaire 23. Une matrice et sa transposée ont même rang. Le rang d'une matrice est la dimension du K-espace vectoriel engendré par ses lignes

Démonstration. Si $PAQ = I_r^{(m,n)}$, alors ${}^tQ\,{}^tA\,{}^tP = I_r^{(n,m)}$.
La deuxième affirmation est alors triviale. ■

Corollaire 24. Toute matrice carrée est équivalente à sa transposée.

Démonstration. Elles ont en effet même type et même rang. ■

Corollaire 25. Soit $M \in M_n(K)$. Alors les conditions suivantes sont équivalentes :
M est simplifiable à gauche ; M est simplifiable à droite ; M est inversible à gauche ;
M est inversible à droite ; M est inversible ; M est de rang n.

Démonstration. Chacune des propriétés énoncées est conservée par les opérations élémentaires (elles sont compatibles avec l'équivalence). Il suffit donc d'examiner le cas de la matrice $I_r = I_r^{(n,n)}$. Pour $r = n$, c'est la matrice identité, qui vérifie toutes ces propriétés. Pour $r < n$, on a $E_{n,n}I_r = I_rE_{n,n} = 0$; la matrice $I_r^{(n,n)}$ n'est donc ni simplifiable à gauche ni simplifiable à droite, et, *fortiori*, ni inversible à gauche, ni inversible à droite, ni inversible. ∎

Comme on l'a remarqué lors de l'étude du groupe linéaire, l'inverse à droite est donc aussi l'inverse à gauche.

Proposition 26. Toute matrice inversible, peut être transformée en la matrice I_n avec des opérations élémentaires sur les lignes.

Démonstration. Encore une fois, on peut partir d'une matrice échelonnée en lignes. Comme elle est carrée, elle est triangulaire supérieure. Comme elle est inversible, sa diagonale est formée de pivots $a_{1,1}, \dots, a_{n,n}$ non nuls. À partir de là :

1. La première étape est la transvection $L_1 \leftarrow L_1 - \dfrac{a_{1,2}}{a_{2,2}}L_2$, à la suite de laquelle la colonne C_2 n'a qu'un coefficient non nul (le pivot $a_{2,2}$) (et la colonne C_1 est inchangée).
2. On effectue ensuite les transvections $L_i \leftarrow L_i - \dfrac{a_{i,3}}{a_{3,3}}L_3$ pour $i < 3$, dans un ordre quelconque car elles sont permutables. À la suite de cette étape, la colonne C_3 n'a qu'un coefficient non nul (le pivot $a_{3,3}$), et les colonnes C_1 et C_2 sont inchangées.
3. Etc., jusqu'à la simplification de C_n.

On a maintenant une matrice diagonale à coefficients diagonaux non nuls, et l'on applique, pour chaque $i \in [\![1, n]\!]$, une dilatation $L_i \leftarrow \dfrac{1}{a_{i,i}}L_i$. ∎

Corollaire 27. Toute matrice inversible, peut être transformée en la matrice I_n avec des opérations élémentaires sur les colonnes

Algorithme d'inversion par opérations élémentaires.. L'algorithme repose sur la remarque suivante : une suite d'opérations sur les lignes qui transforme A en l'identité revient à une multiplication à gauche par P telle que $PA = I_n$, c'est-à-dire par $P = A^{-1}$. La *même* suite d'opérations sur les lignes transforme donc I_n en $PI_n = P$. De cette remarque, on tire la méthode suivante : appliquer des opérations sur les lignes à A et I_n en parallèle ; quand A s'est transformée en I_n, I_n s'est transformée en A^{-1}. D'après le corollaire, on peut tout aussi bien opérer sur les colonnes ; mais *on ne peut pas mélanger des opérations sur les lignes à des opérations sur les colonnes.*

Exercice 13.

On va inverser une matrice $A \in M_2(K)$ quelconque.

Solution. La méthode indiquée en 1.3 serait ici bien plus commode, mais notre but est d'illustrer l'algorithme précédent sur un exemple abordable ! On part de :

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Par commodité, on notera ci-dessous $\Delta = ad - bc$.

Si $a \neq 0$, on effectue la transvection $L_2 \leftarrow L_2 - \dfrac{c}{a} L_1$:

$$\begin{pmatrix} a & b \\ 0 & \Delta/a \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & 0 \\ -c/a & 1 \end{pmatrix}.$$

Si Δ était nul, A ne serait pas inversible ; voyez-vous pourquoi ? On peut donc effectuer la transvection $L_1 \leftarrow L_1 - \dfrac{ab}{\Delta} L_2$:

$$\begin{pmatrix} a & 0 \\ 0 & \Delta/a \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} ad/\Delta & -ab/\Delta \\ -c/a & 1 \end{pmatrix}.$$

Enfin, les dilatations $L_1 \leftarrow \dfrac{1}{a} L_1$ et $L_2 \leftarrow \dfrac{a}{\Delta} L_2$ achèvent le travail :

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad \frac{1}{\Delta} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = A^{-1}.$$

Le calcul ci-dessus n'est valable que si $a \neq 0$. Si $a = 0$, on a $c \neq 0$ et l'on a la suite d'opérations :

$$\begin{pmatrix} 0 & b \\ c & d \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \xrightarrow{L_1 \leftrightarrow L_2} \begin{pmatrix} c & d \\ 0 & b \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \xrightarrow{L_1 \leftarrow L_1 - \frac{d}{b} L_2}$$

$$\begin{pmatrix} c & 0 \\ 0 & b \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} -d/b & 1 \\ 1 & 0 \end{pmatrix} \xrightarrow[L_2 \leftarrow b^{-1} L_2]{L_1 \leftarrow c^{-1} L_1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} -d/bc & 1/c \\ 1/b & 0 \end{pmatrix}$$

On remarque que la formule $A^{-1} = \dfrac{1}{\Delta} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ est encore valable.

Remarque. La phase de triangularisation réussit toujours. À son issue, on peut voir si la matrice est inversible : une condition nécessaire et suffisante est que les coefficients diagonaux soient tous non nuls.

Exercice type corrigé

II.4.0 1) Calculer l'inverse $\begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$ de $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2(K)$. Calculer $\Delta' := a'd' - b'c'$ en fonction de $\Delta := ad - bc$.

À quelle condition la matrice $P := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z})$ admet-elle un inverse dans $M_2(\mathbb{Z})$?

2) On note $GL_2(\mathbb{Z}) := \{P \in M_2(\mathbb{Z}) \mid P^{-1} \in M_2(\mathbb{Z})\}$.
Quelles opérations élémentaires sur $M_2(\mathbb{Q})$ (section 2) correspondent-elles à des matrices de $GL_2(\mathbb{Z})$? Quel est l'effet de ces opérations sur les matrices de $M_2(\mathbb{Z})$?
3) On définit une relation d'équivalence sur $M_2(\mathbb{Z})$ par : $A \sim B$ si, et seulement si, l'on peut passer de A à B par des opérations élémentaires du type ci-dessus. Pour toute matrice A non nulle, on note $|A|$ la plus petite valeur absolue non nulle de ses coefficients.
Démontrer que, pour toute $A \in M_2(\mathbb{Z})$, il existe $B \sim A$ telle que B est diagonale ou $|B| < |A|$.
En déduire que toute matrice de $M_2(\mathbb{Z})$ est équivalente à une matrice diagonale.
4) Trouver une matrice diagonale $B \in M_2(\mathbb{Z})$ telle que $B \sim A = \begin{pmatrix} 2 & 3 \\ 4 & 5 \end{pmatrix}$.

Solution. 1) En résolvant $ax + by = x'$, $cx + dy = y'$, on trouve $x = a'x' + b'y'$, $y = c'x' + d'y'$ avec $a' = \dfrac{d}{\Delta}$, $b' = \dfrac{-b}{\Delta}$, $c' = \dfrac{-c}{\Delta}$, $d' = \dfrac{a}{\Delta}$ ·
On calcule alors $\Delta' = \dfrac{1}{\Delta}$ et l'on en déduit que, pour que $P \in M_2(\mathbb{Z})$ soit inversible dans $M_2(\mathbb{Z})$, il faut, et il suffit, que $\dfrac{1}{\Delta} \in \mathbb{Z}$, c'est-à-dire $\Delta = \pm 1$.
2) Reprenant la liste des matrices associées aux opérations élémentaires de la section 2, on voit que celles qui conviennent sont : les échanges de lignes et les échanges de colonnes ; les dilatations de lignes ou de colonnes de rapport ± 1 ; les transvections de lignes ou de colonnes à coefficient entier. Ces opérations transforment une matrice de $M_2(\mathbb{Z})$ en une matrice de $M_2(\mathbb{Z})$.
3) Par échanges de lignes et de colonnes, on peut ramener en position $(1,1)$ le coefficient de plus petite valeur absolue non nulle. Par dilatation, on peut le rendre positif. Par transvection $C_2 \leftarrow C_2 - bC_1$, on peut rendre le coefficient de position $(1,2)$ positif et strictement plus petit que celui de position $(1,1)$: c'est une division euclidienne ! De même pour le coefficient de position $(2,1)$, par transvection $L_2 \leftarrow L_2 - aL_1$. La matrice B obtenue est alors diagonale ou telle que $|B| < |A|$. On peut réitérer ce procédé, jusqu'à obtenir une matrice diagonale. En effet, si l'on tombait toujours sur le deuxième cas $|B| < |A|$, on aurait une suite strictement décroissante d'entiers naturels.
4) Les opérations $L_2 \leftarrow L_2 - 2L_1$, puis $C_2 \leftarrow C_2 - C_1$, puis échange de colonnes, puis $L_2 \leftarrow L_2 + L_1$ puis $C_2 \leftarrow C_2 - 2C_1$ donnent la matrice diagonale $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \sim \begin{pmatrix} 2 & 3 \\ 4 & 5 \end{pmatrix}$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

II.4.1 Soit $A \in M_n(K)$ une matrice triangulaire supérieure dont tous les coefficients diagonaux valent $\lambda \in K$. Calculer les puissances de A à l'aide de la formule du binôme.

II.4.2 $*$ Soient $A, B \in M_n(K)$ deux matrices nilpotentes qui commutent. Démontrer que $A+B$ est nilpotente à l'aide de la formule du binôme. Donner un exemple de matrices $A, B \in M_2(K)$ nilpotentes mais telles que $A+B$ n'est pas nilpotente.

II.4.3 Soit $A \in M_2(K)$.
Démontrer qu'il existe des scalaires $\tau, \delta \in K$ tels que $A^2 = \tau A - \delta I_2$. En déduire qu'il existe deux suites $(\alpha_n)_{n\in\mathbb{N}}$ et $(\beta_n)_{n\in\mathbb{N}}$ d'éléments de K telles que $\forall n \in \mathbb{N}, A^n = \alpha_n A + \beta_n I_2$.

II.4.4 Soient $a_1, \ldots, a_m, b_1, \ldots, b_n \in K$ des scalaires, et soit $A \in M_{m,n}(K)$ la matrice dont le coefficient d'indices (i,j) est $a_i b_j$. Démontrer que A est nulle si, et seulement si, $a_1 = \cdots = a_m = 0$ ou $b_1 = \cdots = b_n = 0$. Démontrer que, si la matrice A n'est pas nulle, elle est de rang 1. Démontrer enfin la réciproque : toute matrice de rang 1 est de la forme indiquée.

II.4.5 Démontrer que le centre de $GL_n(K)$ est l'ensemble des matrices scalaires inversibles λI_n, avec $\lambda \in K^*$.

II.4.6 Démontrer que, si $n = 3$ et $K = \mathbb{Z}/3\mathbb{Z}$, le groupe constitué des matrices triangulaires supérieures dont tous les éléments diagonaux sont égaux à 1 est un groupe non commutatif dans lequel $\forall x\ ,\ x^3 = 1$.

II.4.7 Démontrer qu'une application linéaire $\varphi : V \to V'$ est injective si, et seulement s'il existe $\psi : V' \to V$ telle que $\psi \circ \varphi = \mathrm{Id}_V$.
En déduire que, si $A \in M_{m,n}(K)$, l'application linéaire $\Phi_A : X \mapsto AX$ est injective si, et seulement si, A est inversible à gauche.

II.4.8 Soit σ une permutation de $[\![1, n]\!]$. On définit la matrice $A_\sigma \in M_n(K)$ comme celle dont le coefficient d'indices (i,j) vaut 1 si $j = \sigma(i)$, et 0 sinon. Calculer le produit de deux telles matrices. En déduire que l'ensemble de ces matrices est un sous-groupe de $GL_n(K)$ isomorphe au groupe symétrique.

Exercices sur la section 2

II.4.9 Inverser les matrices $\begin{pmatrix} 1 & a & a^2 \\ 1 & b & b^2 \\ 1 & c & c^2 \end{pmatrix}$ et $\begin{pmatrix} 1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 1 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$ par opérations sur les lignes ou sur les colonnes.

II.4.10 $*$ Cet exercice fait appel à des connaissances élémentaires sur les polynômes.
Soient $x_1, \ldots, x_n \in K$. On définit la *matrice de Vandermonde* : $A = (a_{i,j})_{1 \leqslant i,j \leqslant n} \in M_n(K)$ par $a_{i,j} = x_i^{j-1}$. Démontrer que, pour qu'elle soit inversible, il faut que les x_i soient deux à deux distincts. On le suppose désormais, et l'on pose, pour $i = 1, \ldots, n-1$:

$$(X - x_1) \cdots (X - x_i) = X^i + \sum_{j=0}^{i-1} \lambda_{i,j} X^j.$$

(i) On effectue l'opération suivante sur les colonnes :

$$C_n \leftarrow C_n + \lambda_{n-1,0} C_1 + \cdots + \lambda_{n-1,n-2} C_{n-1}.$$

Démontrer que cela équivaut à effectuer les $n-1$ transvections permutables :

$$C_n \leftarrow C_n + \lambda_{n-1,i-1} C_i.$$

Décrire la matrice obtenue à la suite de ces opérations. (Seule la dernière colonne a changé, tous ses coefficients sont nuls, sauf le dernier qui vaut maintenant $(x_n - x_1) \cdots (x_n - x_{n-1})$.)
(ii) On effectue les opérations analogues :

$$C_k \leftarrow C_k + \lambda_{k-1,0} C_1 + \cdots + \lambda_{k-1,k-2} C_{k-1},$$

pour k allant (dans cet ordre) de $n-1$ à 2. Décrire la matrice obtenue. (Elle est triangulaire inférieure, ses coefficients diagonaux sont explicitement calculables). En déduire que la matrice de Vandermonde est inversible et calculer son déterminant (celui-ci est défini au module II.7).

II.4.11 $*$ L'algorithme d'Euclide pour le calcul du pgcd $\delta > 0$ de deux entiers $u \geqslant v > 0$, peut être décrit ainsi. On définit, par récurrence à deux pas, une suite $(x_n)_{n \geqslant 0}$ en posant $x_0 := u$, $x_1 := v$ et, tant que $x_i > 0$, $x_{i+1} :=$ le reste de la division euclidienne : $x_{i-1} := q_i x_i + x_{i+1}$. Il existe alors un entier k tel que $x_k \neq 0$ et $x_{k+1} = 0$; le pgcd de u et v est alors $\delta := x_k$. Démontrer que, pour $i = 1, \ldots, k$:

$$\begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_i \end{pmatrix} \begin{pmatrix} x_{i-1} \\ x_i \end{pmatrix}.$$

En déduire que $x_i = a_i u + b_i v$, où :

$$\begin{pmatrix} a_i & b_i \\ \star & \star \end{pmatrix} := \begin{pmatrix} 0 & 1 \\ 1 & -q_i \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix}.$$

Que valent les $\star$ de la deuxième ligne ?
Donner une définition par récurrence mutuelle des suites $(a_n)_{n \geqslant 0}$ et $(b_n)_{n \geqslant 0}$, puis une méthode de calcul des coefficients de Bézout $a, b \in \mathbb{Z}$ tels que $au + bv = \delta$.

Le corps des nombres complexes

II.5

Les thèmes de ce module et du suivant : nombres complexes et polynômes, recouvrent ensemble les bases du calcul algébrique ; on entend ici le mot *algébrique* dans ses deux acceptions classiques, celle liée à la prescription d'algorithmes d'une part, celle qui concerne la structure de l'ensemble des objets du calcul d'autre part. L'étude du corps $\mathbb{C}$, comme celle de $\mathbb{R}$, ne relève cependant pas de l'algèbre pure : elle fait appel à des outils de topologie et d'analyse (suites convergentes, parties fermées ou bornées du plan, fonctions continues ou dérivables). Cela concerne, par exemple, la fonction exponentielle et la trigonométrie (dans ce module), et l'existence et la localisation des racines des polynômes (au module suivant). Sur ces points, on s'aidera librement des modules correspondants de ce livre.

1 Construction et axiomes

Aucun paradoxe et aucune conséquence néfaste ne peuvent être déduits de l'égalité $\mathrm{i}^2 = -1$ jointe aux règles usuelles du calcul, *à la seule exception de la règle des signes* [1]. Les anciens avaient constaté que l'égalité $a^2 = 2$ était incompatible avec leur *conception* des nombres comme *grandeurs commensurables à l'unité* ; mais leurs successeurs ont pu développer le calcul dans $\mathbb{R}$, y compris les irrationnels, sans jamais rencontrer de contradiction. Et c'est seulement au XIXe siècle que l'on a pu construire rigoureusement les nombres réels en accord avec leurs propriétés observées. De même, les modernes ont utilisé le nombre $\mathrm{i} = \sqrt{-1}$ et rassemblé ses propriétés bien avant de savoir construire le corps des nombres complexes. L'application fondamentale de celui-ci, le théorème de d'Alembert-Gauß (possibilité de résoudre *toutes* les équations algébriques, voir la section 5.3) a été pressentie, énoncée, justifiée par des arguments probants bien avant d'être démontrée.

Après avoir construit le corps $\mathbb{C}$ dans la présente section, nous en exposerons le mode d'emploi à la section suivante. Le lecteur pragmatique et doué de foi peut donc, s'il le préfère, passer directement à la section 2 de la page 245.

[1] Il faut aussi mentionner les règles relatives à l'emploi du symbole $\sqrt{\ }$, mais celles-ci sont sans danger si l'on se cantonne aux réels positifs et, dans le cas contraire, exigent le recours à la règle des signes.

1.1 Approche axiomatique

Supposons donnés le corps $\mathbb{R}$ des nombres réels d'un côté, un nombre imaginaire noté i de l'autre, le tout étant englobé dans un domaine de calcul noté $\mathbb{C}$, dont on suppose seulement que les règles usuelles concernant l'addition et la multiplication y ont cours : c'est donc en fait un anneau commutatif $(\mathbb{C}, +, \times)$. Nous *postulons* les propriétés suivantes :

1. Les règles du calcul sur $\mathbb{R}$ sont englobées dans celles de $\mathbb{C}$, autrement dit, $(\mathbb{R}, +, \times)$ est un sous-anneau de $(\mathbb{C}, +, \times)$.

2. Le nombre i vérifie $\mathrm{i}^2 = -1$.

Comme aucun élément de $\mathbb{R}$ n'a pour carré -1, on voit que i n'est pas réel. En fait, les éléments $a+b\mathrm{i}$ de $\mathbb{C}$ sont deux à deux distincts lorsque a et b parcourent $\mathbb{R}$: soient en effet a, b, a', b' des réels. On a alors les implications suivantes :

$$a + b\mathrm{i} = a' + b'\mathrm{i} \quad \Longrightarrow \quad a - a' = (b' - b)\mathrm{i} \quad \Longrightarrow \quad (a - a')^2 = -(b' - b)^2,$$

d'où $a = a'$ et $b = b'$ puisque la dernière égalité porte sur des réels. L'application $(a, b) \mapsto a + b\mathrm{i}$ de $\mathbb{R} \times \mathbb{R}$ dans $\mathbb{C}$ est donc injective. Notons temporairement $\mathbb{C}' = \{a + b\mathrm{i} \mid a, b \in \mathbb{R}\} \subset \mathbb{C}$ son image. Dans l'anneau $\mathbb{C}$, on a $0 \times \mathrm{i} = 0$ et l'on voit, en prenant $b = 0$ et $a \in \mathbb{R}$ quelconque, que $\mathbb{C}'$ contient $\mathbb{R}$. Les calculs qui suivent montrent que $\mathbb{C}'$ est stable pour la soustraction et la multiplication :

$$\begin{aligned}(a + b\mathrm{i}) - (a' + b'\mathrm{i}) &= (a - a') + (b - b')\mathrm{i}\\ (a + b\mathrm{i}) \times (a' + b'\mathrm{i}) &= (aa' - bb') + (ab' + ba')\mathrm{i},\end{aligned}$$

cette dernière égalité découlant des propriétés des anneaux $\mathbb{R}$ et $\mathbb{C}$ et de l'égalité $\mathrm{i}^2 = -1$. Comme $\mathbb{C}'$ contient 1 (il contient tous les réels), c'est en fait un sous-anneau de $\mathbb{C}$. De même, $\mathbb{R}$ est un sous-anneau de $\mathbb{C}'$, ce que l'on peut préciser en disant que l'anneau $\mathbb{C}'$ est *engendré par son élément* i *sur le sous-anneau* $\mathbb{R}$. En effet, tout sous-anneau de $\mathbb{C}'$ qui contient $\mathbb{R}$ et i contient également tous les éléments $a + b\mathrm{i}$ lorsque a et b parcourent $\mathbb{R}$.

Parmi toutes les solutions possibles à notre problème de départ, nous en voulons une économique ; nous choisirons donc $\mathbb{C}$ aussi petit que possible : on voit qu'il faut prendre $\mathbb{C} = \mathbb{C}'$. Ce qui précède va nous suffire pour procéder à la construction effective du corps $\mathbb{C}$, dont nous n'avons fait que postuler l'existence. Voici d'abord deux déductions simples. Tout d'abord, il ne peut y avoir d'équivalent de la règle des signes dans $\mathbb{C}$. En effet, lorsqu'une telle règle est valide, tous les carrés sont positifs (que x soit positif ou négatif, la règle des signes entraîne que x^2 est positif). Or, dans $\mathbb{C}$, 1 et -1 sont des carrés et ils ne peuvent être tous deux positifs. *Il n'y a dans* $\mathbb{C}$ *aucune relation d'ordre total compatible avec la structure d'anneau.*

Du côté positif, bien que nous n'ayons pas postulé la possibilité de diviser dans $\mathbb{C}$, on peut vérifier dès à présent que c'est un corps : soit en effet $a + b\mathrm{i}$ un élément non nul de $\mathbb{C}$, c'est-à-dire que $(a, b) \in \mathbb{R}^2 \setminus \{(0, 0)\}$ (d'après l'injectivité prouvée plus haut).

On peut donc poser $a' := \dfrac{a}{a^2 + b^2}$ et $b' := \dfrac{-b}{a^2 + b^2}$.

On trouve sans peine que $aa' - bb' = 1$ et $ab' + ba' = 0$. La formule de multiplication déjà vue donne alors : $(a + b\mathrm{i}) \times (a' + b'\mathrm{i}) = (aa' - bb') + (ab' + ba')\mathrm{i} = 1$.

1.2 Construction effective de $\mathbb{C}$

Nous *définissons* $\mathbb{C}$ comme l'ensemble produit $\mathbb{R} \times \mathbb{R}$ muni des deux lois de composition internes suivantes ; quels que soient $(x_1, y_1), (x_2, y_2) \in \mathbb{R} \times \mathbb{R}$, on pose :

$$(x_1, y_1) + (x_2, y_2) := (x_1 + x_2, y_1 + y_2),$$

$$(x_1, y_1) \times (x_2, y_2) := (x_1x_2 - y_1y_2, x_1y_2 + y_1x_2).$$

La notation $(x_1, y_1) \times (x_2, y_2)$ pourra être allégée en $(x_1, y_1)(x_2, y_2)$.

Proposition 1. On obtient ainsi un corps commutatif $(\mathbb{C}, +, \times)$.

Démonstration. Tout d'abord, $(\mathbb{C}, +)$ est un groupe commutatif : on reconnaît en effet le produit du groupe commutatif $(\mathbb{R}, +)$ par lui-même. En particulier, son élément neutre est $(0, 0)$. La loi $\times$ est associative :

$$(x_1, y_1) \times [(x_2, y_2) \times (x_3, y_3)] = [(x_1, y_1) \times (x_2, y_2)] \times (x_3, y_3).$$

Ce calcul figure dans n'importe quel manuel de terminale et peut d'ailleurs être vérifié à l'aide d'un logiciel de calcul formel ; cela s'applique également à toutes les égalités qui suivant. La commutativité de la multiplication se voit immédiatement sur la définition, dans laquelle les indices 1 et 2 peuvent être permutés. L'élément $(1, 0)$ est neutre à gauche, (donc à droite par commutativité) : $(1, 0) \times (x, y) = (x, y)$. Nous avons enfin la distributivité à gauche de la multiplication par rapport à l'addition (et la distributivité à droite en découle par commutativité) :

$$[(x_1, y_1) + (x_1', y_1')] \times (x_2, y_2) = [(x_1, y_1) \times (x_2, y_2)] + [(x_1', y_1') \times (x_2, y_2)].$$

Il reste à voir que tout élément non nul est inversible. Soit donc $(x, y) \in \mathbb{R} \times \mathbb{R} \setminus \{(0, 0)\}$, de sorte que $x^2 + y^2 > 0$. Posons $x' := \dfrac{x}{x^2 + y^2}$ et $y' := \dfrac{-y}{x^2 + y^2}$. Alors $xx' - yy' = 1$ et $xy' + yx' = 0$, d'où l'on tire : $(x, y) \times (x', y') = (xx' - yy', xy' + yx') = (1, 0)$. On a bien un inverse à gauche (donc à droite), ce qui achève la preuve. ∎

La construction qui précède vise à représenter le nombre complexe $a + b\mathrm{i}$ par le couple (a, b). Ce dernier s'écrit également : $(a, b) = (a, 0) + (0, b) = (a, 0) + (b, 0)(0, 1)$, car

$$(b, 0) \times (0, 1) = (b.0 - 0.1, b.1 + 0.0) = (b, 0).$$

Nous comptons donc identifier le nombre imaginaire i au couple $(0, 1)$ ainsi que les réels a et b aux couples $(a, 0)$ et $(b, 0)$.

Proposition 2. L'application $a \mapsto (a, 0)$ de $\mathbb{R}$ dans $\mathbb{C}$ est un morphisme injectif d'anneaux.

Démonstration. Notons φ cette application.
Les égalités $(a, 0) + (b, 0) = (a + b, 0)$ et $(a, 0)(b, 0) = (ab, 0)$ signifient respectivement que $\varphi(a + b) = \varphi(a) + \varphi(b)$ et que $\varphi(ab) = \varphi(a)\varphi(b)$. Par ailleurs, $\varphi(1) = (1, 0)$ est l'élément neutre de la multiplication dans $\mathbb{C}$. On a donc bien un morphisme d'anneaux. Son noyau est formé des $a \in \mathbb{R}$ tels que $(a, 0) = (0, 0)$, il est donc réduit à $\{0\}$, et φ est bien injectif. ∎

Cette application réalise donc un isomorphisme de l'anneau $\mathbb{R}$ sur son image, qui est le sous-anneau $\mathbb{R} \times \{0\}$ de $\mathbb{C}$. *On convient d'identifier $\mathbb{R}$ à ce sous-anneau et chaque réel $a \in \mathbb{R}$ à son image $(a, 0) \in \mathbb{C}$.* C'est un procédé courant en algèbre (*cf.* l'identification de $\mathbb{N}$ avec son image dans $\mathbb{Z}$, celle de $\mathbb{Z}$ avec son image dans $\mathbb{Q}$, ou bien celle de $\mathbb{Q}$ avec son image dans $\mathbb{R}$).

Définition 1. On note i l'élément $(0,1)$ de $\mathbb{C}$.

Cette écriture est justifiée par les égalités $\mathrm{i}^2 = (0,1)\times(0,1) = (-1,0)$, soit, avec l'identification ci-dessus, $\mathrm{i}^2 = -1$.

Proposition 3. Tout élément de $\mathbb{C}$ s'écrit d'une façon et d'une seule sous la forme $a + b\mathrm{i}$, avec $a, b \in \mathbb{R}$. On a les égalités :

$$(a_1 + b_1\mathrm{i}) + (a_2 + b_2\mathrm{i}) = (a_1 + a_2) + (b_1 + b_2)\mathrm{i},$$
$$(a_1 + b_1\mathrm{i})(a_2 + b_2\mathrm{i}) = (a_1a_2 - b_1b_2) + (b_1a_2 + a_1b_2)\mathrm{i}.$$

Démonstration. Puisque $a+b\mathrm{i} = (a,0)+(b,0)(0,1) = (a,b)$, l'existence et l'unicité de l'écriture indiquée sont claires. Les deux formules qui suivent ne font que plagier la définition de l'addition et de la multiplication dans $\mathbb{C}$. ■

Nous écrirons donc sous la forme $a + b\mathrm{i}$ l'élément (a,b) de $\mathbb{C}$.
Le cas où $b = 0$ correspond à $a + 0\mathrm{i}$, soit le couple $(a,0)$, qui a été identifié au nombre réel a : on l'écrira donc a. Le cas où $a = 0$ correspond au couple $(0,b)$ que l'on écrira $0 + b\mathrm{i} = b\mathrm{i}$; on dira que ce nombre est *un imaginaire pur*. Les éléments de $\mathbb{C}$ seront appelés *nombres complexes*.

Théorème 4. Soit $\tilde{\mathbb{C}}$ un un anneau commutatif dont $\mathbb{R}$ est un sous-anneau et engendré sur $\mathbb{R}$ par un élément $\tilde{\imath}$ de carré -1. Il existe alors un unique isomorphisme de $\mathbb{C}$ dans $\tilde{\mathbb{C}}$ dont la restriction à $\mathbb{R}$ est l'identité et par lequel l'image de i est $\tilde{\imath}$.

Démonstration. Le lecteur vérifiera en effet que l'application $a + b\mathrm{i} \mapsto a + b\tilde{\imath}$ de $\mathbb{C}$ dans $\tilde{\mathbb{C}}$ est la seule qui convienne. ■

Exercice 1.
Déterminer, sans référence au théorème, tous les automorphismes de $\mathbb{C}$ dont la restriction à $\mathbb{R}$ est l'identité. Ceux-ci forment un groupe (pour la composition) noté $\mathrm{Aut}(\mathbb{C} \mid \mathbb{R})$ dont l'identité et la conjugaison sont évidemment éléments.
Solution. Soit σ un tel automorphisme. Pour $x, y \in \mathbb{R}$, on a :

$$\begin{aligned}
\sigma(x + y\mathrm{i}) &= \sigma(x) + \sigma(y)\sigma(\mathrm{i}) && \text{(car } \sigma \text{ est un automorphisme de corps)}\\
&= x + y\sigma(\mathrm{i}) && \text{(car la restriction de } \sigma \text{ à } \mathbb{R} \text{ est l'identité)}\\
\text{et}\quad \sigma(\mathrm{i})^2 &= \sigma(\mathrm{i}^2) && \text{(car } \sigma \text{ est un automorphisme de corps)}\\
&= \sigma(-1) = -1 && \text{(car la restriction de } \sigma \text{ à } \mathbb{R} \text{ est l'identité).}
\end{aligned}$$

On en déduit que $\sigma(\mathrm{i})^2 = \mathrm{i}^2$, d'où $\sigma(\mathrm{i}) = \pm\mathrm{i}$. En effet, dans tout anneau commutatif intègre (donc dans $\mathbb{C}$), on a l'implication $a^2 = b^2 \Rightarrow a = \pm b$. On distingue alors deux cas : si $\sigma(\mathrm{i}) = \mathrm{i}$, on a $\sigma(x + y\mathrm{i}) = x + y\mathrm{i}$ pour tous $x, y \in \mathbb{R}$ et l'automorphisme σ est l'identité. Sinon, $\sigma(\mathrm{i}) = -\mathrm{i}$, et l'on a $\sigma(x+y\mathrm{i}) = x - y\mathrm{i}$ pour tous $x, y \in \mathbb{R}$: c'est l'automorphisme de conjugaison (voir 2.3). On retrouve cet automorphisme en appliquant le théorème ci-dessus à $\tilde{\mathbb{C}} = \mathbb{C}$ et $\tilde{\imath} = -\mathrm{i}$. En conclusion, on a l'égalité $\mathrm{Aut}(\mathbb{C} \mid \mathbb{R}) = \{\mathrm{Id}_{\mathbb{C}}, \sigma\}$.

La construction de $\mathbb{C}$ est achevée. Dans la pratique, le complexe $x + y\mathrm{i}$ est plutôt noté $x + \mathrm{i}y$, et nous nous conformerons à cet usage dès la prochaine section.

2 Règles élémentaires de calcul

Le lecteur qui le désire peut entamer ici l'étude des nombres complexes, quitte à admettre certains résultats élémentaires.

2.1 Représentation cartésienne

Un *nombre complexe* est un nombre $z = x + \mathrm{i}y$, où x et y sont des réels et où i désigne un « nombre imaginaire » de carré -1. On parle aussi plus brièvement du *complexe* z. L'ensemble des complexes est noté :

$$\mathbb{C} := \{x + \mathrm{i}y \mid x, y \in \mathbb{R}\}.$$

Tout complexe $z \in \mathbb{C}$ s'écrit de manière unique $z = x + \mathrm{i}y$, avec $x, y \in \mathbb{R}$. Autrement dit, l'application $(x, y) \mapsto x + \mathrm{i}y$ de $\mathbb{R}^2$ dans $\mathbb{C}$ est bijective.

Définition 2. L'écriture $z = x + \mathrm{i}y$ est appelée *représentation cartésienne* ou *algébrique* du nombre complexe $z \in \mathbb{C}$. On appelle *partie réelle de* z le réel $\operatorname{Re} z := x$ et *partie imaginaire de* z le réel $\operatorname{Im} z := y$.

En principe, il ne devrait pas y avoir de confusion avec la notation $\operatorname{Im} f$ pour l'image d'une application f. On prendra garde que *la partie imaginaire d'un nombre complexe est un nombre réel*.

Pour vérifier l'égalité de deux nombres complexes, il suffit donc de vérifier séparément l'égalité de leurs parties réelles et l'égalité de leurs parties imaginaires. Ainsi, une égalité dans $\mathbb{C}$ correspond, de manière générale, à deux égalités dans $\mathbb{R}$. Comme on le verra, cela exprime le fait que $\mathbb{C}$ est un espace vectoriel de dimension 2 sur $\mathbb{R}$. On convient d'identifier le nombre réel $x \in \mathbb{R}$ au nombre complexe $x + 0\mathrm{i} \in \mathbb{C}$. De même, on notera $\mathrm{i}y$ le *nombre imaginaire pur* $0 + \mathrm{i}y \in \mathbb{C}$. Ainsi, pour tout complexe $z \in \mathbb{C}$:

$$z \text{ est réel } \iff \operatorname{Im} z = 0 \qquad \text{et} \qquad z \text{ est imaginaire pur } \iff \operatorname{Re} z = 0.$$

Définition 3. La somme des complexes $z = x + \mathrm{i}y$ et $z' = x' + \mathrm{i}y'$ est le complexe :

$$z + z' := (x + x') + \mathrm{i}(y + y') \in \mathbb{C}.$$

Le produit des complexes $z = x + \mathrm{i}y$ et $z' = x' + \mathrm{i}y'$ est le complexe :

$$z \times z' := (xx' - yy') + \mathrm{i}(xy' + yx') \in \mathbb{C}.$$

On le notera également $z.z'$, ou zz'.

En particulier, le produit du complexe $y = y + \mathrm{i}0$ et du complexe $\mathrm{i} = 0 + \mathrm{i}1$ est $\mathrm{i}y$; de même, la somme du complexe $x = x + \mathrm{i}0$ et du complexe $\mathrm{i}y = 0 + \mathrm{i}y$ est le complexe $x + \mathrm{i}y$. Nos notations sont donc cohérentes !

Quand nous parlerons du complexe $z = x + \mathrm{i}y$, nous sous-entendrons en général qu'il s'agit là de sa représentation cartésienne ; x et y sont donc implicitement supposés réels : $x = \operatorname{Re} z$ et $y = \operatorname{Im} z$. En l'absence de cette convention, beaucoup de formules deviendraient fausses. Par exemple, le lecteur est invité à examiner la validité de l'égalité $\operatorname{Re}(x + \mathrm{i}y) = x$ lorsque $x = 1$ et $y = \mathrm{i}$.

De la section 1, on retiendra que $(\mathbb{C}, +)$ est un groupe commutatif d'élément neutre $0 = 0 + 0\mathrm{i}$; l'opposé de $z = x + \mathrm{i}y$ y est donné par la formule : $-z = (-x) + \mathrm{i}(-y)$. Le lien entre l'addition de $\mathbb{R}$ et celle de $\mathbb{C}$ s'exprime par les règles suivante :

$$\mathrm{Re}(z + z') = \mathrm{Re}\, z + \mathrm{Re}\, z', \qquad \text{et} \qquad \mathrm{Im}(z + z') = \mathrm{Im}\, z + \mathrm{Im}\, z'. \tag{1}$$

Pour tout réel λ et tout complexe $z = x + \mathrm{i}y$, on a posé $\lambda.z := (\lambda x) + \mathrm{i}(\lambda y)$. Cette multiplication, qui est un cas particulier de la multiplication interne de $\mathbb{C}$ (cas où le premier facteur est réel), fait du groupe additif $\mathbb{C}$ un espace vectoriel sur le corps $\mathbb{R}$. Avec ces notations, on a les règles :

$$\mathrm{Re}(\lambda z) = \lambda\, \mathrm{Re}\, z \qquad \text{et} \qquad \mathrm{Im}(\lambda z) = \lambda\, \mathrm{Im}\, z, \tag{2}$$

qui ne sont, bien sûr, valables, *que si λ est réel*.

Proposition 5. (i) Les applications Re et Im sont des applications linéaires surjectives du $\mathbb{R}$-espace vectoriel $\mathbb{C}$ sur le $\mathbb{R}$-espace vectoriel $\mathbb{R}$. Le noyau de Im est le sous-espace vectoriel $\mathbb{R}$ de $\mathbb{C}$. Le noyau de Re est le sous-espace vectoriel $\mathbb{R}\mathrm{i} = \{\mathrm{i}y \mid y \in \mathbb{R}\}$ de $\mathbb{C}$ formé des imaginaires purs.
(ii) L'application $(x, y) \mapsto x + \mathrm{i}y$ est un isomorphisme de l'espace vectoriel $\mathbb{R}^2$ sur l'espace vectoriel $\mathbb{C}$. L'isomorphisme réciproque est $z \mapsto (\mathrm{Re}\, z, \mathrm{Im}\, z)$.

Démonstration. Elle découle des règles précédentes et sera omise. ■

Corollaire 6. L'image de la base canonique de $\mathbb{R}^2$ par l'isomorphisme ci-dessus est la base $(1, \mathrm{i})$ du $\mathbb{R}$-espace vectoriel $\mathbb{C}$, qui est donc de dimension 2.

Démonstration. Elle est immédiate en vertu du module II.3. ■

Tout corps est naturellement muni d'une structure d'espace vectoriel sur lui-même, et cet espace vectoriel est de dimension 1, car il admet la base (1) (famille réduite à un seul élément). Ainsi, $\mathbb{C}$ est un $\mathbb{C}$-espace vectoriel de dimension 1, tout en étant un $\mathbb{R}$-espace vectoriel de dimension 2, qui admet la base $(1, \mathrm{i})$ d'après le corollaire. De même, les applications Re et Im sont-elles $\mathbb{R}$-linéaires (règles (1) et (2) ci-dessus), mais elles ne sont pas $\mathbb{C}$-linéaires : calculer $\mathrm{Re}(\mathrm{i}z)$ et $\mathrm{Im}(\mathrm{i}z)$ lorsque $z = 1$ ou i.

Nous nous intéressons maintenant à la multiplication. De la section 1, on retiendra que $(\mathbb{C}, +, \times)$ est un anneau commutatif. L'élément neutre de la multiplication (l'unité) est le complexe $1 = 1 + 0\mathrm{i}$. C'est même un corps, car tout élément non nul admet un inverse. En effet, le complexe $z = x + \mathrm{i}y$ est non nul si, et seulement si, $(x, y) \neq (0, 0)$, ce qui équivaut à $x^2 + y^2 > 0$ (il s'agit de réels). On peut alors calculer : $z^{-1} = \dfrac{x}{x^2 + y^2} + \mathrm{i}\dfrac{-y}{x^2 + y^2}$, et vérifier que $zz^{-1} = z^{-1}z = 1$. Notant selon l'usage $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$, on a donc bien un groupe multiplicatif $(\mathbb{C}^*, \times)$.

Théorème 7. L'addition et la multiplication font de $\mathbb{C}$ un corps commutatif.

2.2 Le plan d'Argand-Cauchy

L'identification du corps $\mathbb{R}$ avec une droite permet de prouver algébriquement des théorèmes de géométrie et, réciproquement, d'employer un langage géométrique (points, segments, distances ...) pour parler de nombres. Nous allons pareillement donner une image géométrique du corps $\mathbb{C}$. Nous n'indiquons ici que les rudiments de cette représentation. Les applications géométriques seront décrites en section 4 à la page 267.

Considérons un plan Π, appelé le *plan d'Argand-Cauchy*, muni d'un repère orthonormé $(O, \vec{u}, \vec{v})$. Ainsi, le point $M(x, y)$ du plan Π de coordonnées $(x, y) \in \mathbb{R}^2$, est caractérisé par la relation : $\overrightarrow{OM} = x\vec{u} + y\vec{v}$.

Définition 4. (i) On dit que le complexe $z = x + \mathrm{i}y$ est *l'affixe* du point $M(x, y)$ et on le note $\mathrm{Aff}(M)$ (le mot *affixe* est un nom féminin).
Réciproquement, le point M est appelé *(point) image* du complexe z et noté $[z]$.
(ii) Soit $\overrightarrow{V}$ un vecteur de coordonnées (ξ, η). On appelle *affixe* de $\overrightarrow{V}$ le nombre complexe $\zeta = \xi + \mathrm{i}\eta$. On le note $\mathrm{Aff}(\overrightarrow{V})$.
Réciproquement, $\overrightarrow{V}$ est appelé *(vecteur) image* du complexe ζ et noté $[\vec{\zeta}]$.

Ainsi, si $z = \mathrm{Aff}(M)$, l'abscisse de M est $\mathrm{Re}\, z$ et son ordonnée est $\mathrm{Im}\, z$.

Soient $M(x, y)$ et $M'(x', y')$ deux points de Π. Le vecteur $\overrightarrow{V} = \overrightarrow{MM'}$ a pour coordonnées $(\xi, \eta) = (x' - x, y' - y)$: $\overrightarrow{V} = \xi\vec{u} + \eta\vec{v}$. On déduit immédiatement de ces définitions les règles suivantes :

$$\overrightarrow{[z][z']} = \overrightarrow{[z' - z]}, \quad \mathrm{Aff}(\overrightarrow{MM'}) = \mathrm{Aff}(M') - \mathrm{Aff}(M).$$

Comme première série d'applications, décrivons le lien entre certaines opérations sur les complexes et certaines transformations simples du plan.

Exemples.

1. Soit $\overrightarrow{V}$ un vecteur de coordonnées (a, b). La *translation de vecteur* $\overrightarrow{V}$ transforme le point $M(x, y)$ en le point $M'(x + a, y + b)$. Ainsi :
$$\mathrm{Aff}(M') = \mathrm{Aff}(M) + \mathrm{Aff}(\overrightarrow{V}).$$
2. Soit λ un réel. L'*homothétie de rapport λ et de centre O* transforme le point $M(x, y)$ en le point $M'(\lambda x, \lambda y)$. Ainsi :
$$\mathrm{Aff}(M') = \lambda\ \mathrm{Aff}(M).$$
3. En particulier, la *symétrie centrale de centre O*, qui n'est autre que l'homothétie de rapport -1, transforme $M(x, y)$ en $M'(-x, -y)$:
$$\mathrm{Aff}(M') = -\ \mathrm{Aff}(M).$$
4. La *rotation de centre O et d'angle $\pi/2$* transforme $\vec{u}$ en $\vec{v}$ et $\vec{v}$ en $-\vec{u}$, donc $M(x, y)$ en $M'(x', y')$, avec $x'\vec{u} + y'\vec{v} = x\vec{v} + y(-\vec{u})$, d'où $(x', y') = (-y, x)$.
Alors $x' + \mathrm{i}y' = -y + \mathrm{i}x = \mathrm{i}\,(x + \mathrm{i}y)$:
$$\mathrm{Aff}(M') = \mathrm{i}\ \mathrm{Aff}(M).$$

Exercice 2.
Décrire l'homothétie de rapport λ et de centre A et la similitude directe de centre A, de rapport λ et d'angle $k\pi/2$ ($k \in \mathbb{Z}$).
Solution. Soient M' et M'' les images de M par ces deux transformations.
On a $\overrightarrow{AM'} = \lambda\,\overrightarrow{AM}$, d'où $\mathrm{Aff}(M') - \mathrm{Aff}(A) = \lambda\big(\mathrm{Aff}(M) - \mathrm{Aff}(A)\big)$.
On en conclut : $\mathrm{Aff}(M') = \lambda\,\mathrm{Aff}(M) + (1-\lambda)\,\mathrm{Aff}(A)$.
Le vecteur $\overrightarrow{AM''}$ s'obtient à partir de $\overrightarrow{AM}$ en multipliant par λ et en tournant k fois de $\pi/2$ (ou $-k$ fois de $-\pi/2$). Ainsi : $\mathrm{Aff}(M'') - \mathrm{Aff}(A) = \lambda\,\mathrm{i}^k\big(\mathrm{Aff}(M) - \mathrm{Aff}(A)\big)$.
On laisse au lecteur le calcul de $\mathrm{Aff}(M'')$; il y a une discussion sur la valeur de i^k, qui ne dépend que de $k \mod 4$, car $\mathrm{i}^4 = (-1)^2 = 1$.

Voici enfin une détermination de *lieu géométrique* (partie du plan).
Exemple. La droite reliant les points $M_0 = [z_0]$ et $M_1 = [z_1]$ (supposés distincts) est formée des points M tels que l'on puisse écrire $\overrightarrow{M_0M} = t\,\overrightarrow{M_0M_1}$ pour un réel t. Ce sont donc les points d'affixes $z = tz_1 + (1-t)z_0$, $t \in \mathbb{R}$.
Les points du segment reliant M_0 et M_1 sont caractérisés par la condition $0 \leqslant t \leqslant 1$. Ce sont donc les points d'affixes $z = tz_1 + (1-t)z_0$, $t \in [0,1]$.

2.3 Conjugaison

Définition 5. Le *conjugué* du complexe $z \in \mathbb{C}$ est le complexe $\overline{z} := \mathrm{Re}\,z - \mathrm{i}\,\mathrm{Im}\,z$; en représentation cartésienne :

$$\overline{x + \mathrm{i}y} = x - \mathrm{i}y.$$

Nous donnerons sans preuve quelques règles de calcul et de raisonnement concernant le conjugué ; le lecteur est invité à les vérifier.

$$\overline{\overline{z}} = z \quad \text{(C1)} \qquad \overline{z + z'} = \overline{z} + \overline{z'} \quad \text{(C2)}$$

$$\overline{zz'} = \overline{z}\,\overline{z'} \quad \text{(C3)} \qquad \forall \lambda \in \mathbb{R}\ ,\ \overline{\lambda z} = \lambda\,\overline{z} \quad \text{(C4)}$$

$$\mathrm{Re}\,z = \frac{1}{2}(z + \overline{z}) \quad \text{(C5)} \qquad \mathrm{Im}\,z = \frac{1}{2\mathrm{i}}(z - \overline{z}) = \frac{\mathrm{i}}{2}(\overline{z} - z) \quad \text{(C6)}$$

$$z \in \mathbb{R} \Longleftrightarrow \overline{z} = z \quad \text{(C7)} \qquad z \in \mathbb{R}\mathrm{i} \Longleftrightarrow \overline{z} = -z \quad \text{(C8)}$$

Dans la règle (C6), on prendra garde au dénominateur.

Proposition 8. La conjugaison est un automorphisme involutif du corps commutatif $\mathbb{C}$. C'est également un automorphisme du $\mathbb{R}$-espace vectoriel $\mathbb{C}$.

Démonstration. C'est une traduction des règles (C1) à (C8). ∎

Exemple. La *symétrie par rapport à l'axe des abscisses* transforme $M(x,y)$ en $M'(x,-y)$, d'affixe $\mathrm{Aff}(M') = \overline{\mathrm{Aff}(M)}$. La *symétrie par rapport à l'axe des ordonnées* transforme $M(x,y)$ en $M''(-x,y)$, d'affixe $\mathrm{Aff}(M'') = -\overline{\mathrm{Aff}(M)}$.

La conjugaison va nous fournir un moyen commode de tester si deux vecteurs sont parallèles, voire de même sens. On dira que les complexes z et z' sont *réellement liés* s'il existe $\lambda, \lambda' \in \mathbb{R}$ non tous deux nuls tels que $\lambda z = \lambda' z'$. On dira que z et z' sont *positivement liés* si les réels λ, λ' peuvent être de plus choisis de même signe (au sens large). Les complexes $z, z' \in \mathbb{C}$ sont donc réellement liés (resp. positivement liés) si, et seulement si, les vecteurs (x, y) et (x', y') de $\mathbb{R}^2$ sont proportionnels (resp. proportionnels avec coefficient de proportionnalité positif). Si l'un des complexes z, z' est nul, ils sont réellement liés et positivement liés. Si $z \neq 0$, ils sont réellement liés (resp. positivement liés) si, et seulement si, $\dfrac{z'}{z} \in \mathbb{R}$ (resp. $\dfrac{z'}{z} \in \mathbb{R}_+$).

Lemme 9. Les complexes z et z' sont réellement liés (resp. positivement liés) si, et seulement si, $\overline{z}z' \in \mathbb{R}$ (resp. $\overline{z}z' \in \mathbb{R}_+$).

Démonstration. Si $z = 0$, c'est clair. Sinon, on remarque que $\dfrac{z'}{z} = \dfrac{1}{A}\overline{z}z'$, où $A = z\overline{z} \in \mathbb{R}_+^*$, ce qui rend évidentes les équivalences énoncées. ■

Exemples.

1. Soit à déterminer la droite D passant par le point $M_0(x_0, y_0)$ et parallèle à la direction de vecteur directeur $\overrightarrow{V} = a\vec{u} + b\vec{v} \neq \vec{0}$. Notant $w = \mathrm{Aff}(\overrightarrow{V})$, $z_0 = \mathrm{Aff}(M_0)$ et $z = \mathrm{Aff}(M)$, on voit que $M \in D$ si, et seulement si, les vecteurs $\overrightarrow{AM}$ et $\overrightarrow{V}$ sont liés. Cela équivaut à $\overline{w}(z - z_0) \in \mathbb{R}$, donc à $\mathrm{Im}\big(\overline{w}(z - z_0)\big) = 0$. On retrouve l'équation $b(x - x_0) - a(y - y_0) = 0$.
2. Soit à déterminer la droite D' passant par le point $M_0(x_0, y_0)$ et perpendiculaire à la direction de vecteur directeur $\overrightarrow{V} = a\vec{u} + b\vec{v} \neq \vec{0}$. La rotation d'angle $\pi/2$ multiplie l'affixe de ce vecteur par i. Avec les mêmes notations que précédemment, on voit donc que $M \in D'$ si, et seulement si, $\overline{\mathrm{i}w}(z - z_0) \in \mathbb{R}$, c'est-à-dire $\overline{w}(z - z_0) \in \mathbb{R}\mathrm{i}$, ou encore $\mathrm{Re}\big(\overline{w}(z - z_0)\big) = 0$. On retrouve l'équation $a(x - x_0) + b(y - y_0) = 0$.

Au passage, nous avons trouvé un critère d'orthogonalité.
Soient $z = \mathrm{Aff}(\overrightarrow{V})$ et $z' = \mathrm{Aff}(\overrightarrow{V'})$, alors :

$$\overrightarrow{V} \perp \overrightarrow{V'} \Longleftrightarrow \mathrm{Re}(\overline{z}z') = 0. \tag{3}$$

2.4 Module

Définition 6. On appelle *norme algébrique*[2] du complexe z le réel $N(z) := z\overline{z}$.

[2] Il faut prendre garde que cet usage du mot « norme », traditionnel en théorie des nombres, n'est pas conforme à celui du cours d'analyse, dans lequel on dirait plutôt que c'est l'application $z \mapsto |z|$ (définie un peu plus loin) qui est une norme. On prendra donc garde à ne pas séparer les mots « norme » et « algébrique » dans cette expression.

En représentation cartésienne, si $z = x + \mathrm{i}y$:

$$N(z) = z\overline{z} = (x + \mathrm{i}y)(x - \mathrm{i}y) = x^2 - y^2\mathrm{i}^2 = x^2 + y^2,$$

qui est bien un réel. On tire immédiatement de cette formule :

$$N(z) \geqslant 0, \qquad N(z) = 0 \Longleftrightarrow z = 0 \qquad \text{et} \qquad \forall x \in \mathbb{R}\ ,\ N(x) = x^2.$$

Exercice 3.
Utiliser la norme algébrique de $z \in \mathbb{C}^*$ pour diviser simplement par z dans $\mathbb{C}$ (penser à « la quantité conjuguée »).

Solution. L'inverse de la norme de z est le réel strictement positif $\dfrac{1}{N(z)}$ · Soit $z' = x' + \mathrm{i}y'$ un autre complexe. Pour diviser, on « multiplie en haut et en bas par la quantité conjuguée » :

$$\begin{aligned}
\frac{z'}{z} &= \frac{z'\overline{z}}{z\overline{z}} = \frac{z'\overline{z}}{N(z)} = \frac{1}{N(z)}z'\overline{z} = \frac{1}{N(z)}(x' + \mathrm{i}y')(x - \mathrm{i}y) \\
&= \frac{1}{N(z)}\big((xx' + yy') + \mathrm{i}(xy' - yx')\big) = \frac{xx' + yy'}{N(z)} + \mathrm{i}\frac{xy' - yx'}{N(z)} \\
&= \frac{xx' + yy'}{x^2 + y^2} + \mathrm{i}\frac{xy' - yx'}{x^2 + y^2}\ \cdot
\end{aligned}$$

Proposition 10. On a les égalités :

$$\begin{aligned}
N(\overline{z}) &= N(z) \\
\forall \lambda \in \mathbb{R}\ ,\ N(\lambda z) &= \lambda^2 N(z) \\
N(zz') &= N(z)N(z'), \\
N(z + z') &= N(z) + N(z') + 2\,\mathrm{Re}(\overline{z}z') \qquad (4) \\
N(z - z') &= N(z) + N(z') - 2\,\mathrm{Re}(\overline{z}z') \\
\mathrm{Re}(\overline{z}z') &= \frac{1}{4}\big(N(z + z') - N(z - z')\big)\ \cdot
\end{aligned}$$

Démonstration. Les deux premières égalités sont immédiates, la suivante est conséquence de la règle (C3) de la page 248. Pour prouver la quatrième , on écrit :

$$N(z + z') = (z + z')(\overline{z} + \overline{z'}) = z\overline{z} + z\overline{z'} + z'\overline{z} + z'\overline{z'} = N(z) + N(z') + z\overline{z'} + z'\overline{z},$$

et l'on conclut en remarquant que $z\overline{z'} + z'\overline{z} = z\overline{z'} + \overline{z\overline{z'}} = 2\,\mathrm{Re}(\overline{z}z')$.
Les deux dernières règles s'en déduisent facilement. ■

Définition 7. Le *module* du nombre complexe z est le réel positif $|z| := \sqrt{N(z)}$.

Puisque $N(z) \geqslant 0$, cette définition est justifiée. La racine carrée d'un réel positif est, par convention, un réel positif. En représentation cartésienne :

$$|x + \mathrm{i}y| = \sqrt{x^2 + y^2}.$$

Cette formule est identique à celle qui donne la longueur en géométrie euclidienne. On tire des propriétés de la norme :

$$|z| \geqslant 0 \qquad \text{et} \qquad |z| = 0 \Longleftrightarrow z = 0.$$

De plus, pour un réel x, le module de x est égal à sa valeur absolue. Nous allons traduire en termes de modules la proposition 10 de la page précédente. Pour cela, il nous faudra une traduction algébrique de *l'inégalité de Cauchy-Schwarz* :

Lemme 11. Soient $z, z' \in \mathbb{C}$.
Alors $\mathrm{Re}(\overline{z}z') \leqslant |z|\,|z'|$, avec égalité si, et seulement si, z et z' sont positivement liés.

Démonstration. Si $z = 0$, l'inégalité est vérifiée et les complexes sont positivement liés. Dans le cas contraire, on peut diviser les deux membres de l'inégalité par le réel strictement positif $N(z) = z\overline{z}$, ce qui la transforme en : $\mathrm{Re}\,\dfrac{z'}{z} \leqslant \left|\dfrac{z'}{z}\right|$ · On est donc ramené à prouver, pour tout complexe w, que $\mathrm{Re}(w) \leqslant |w|$ avec égalité si, et seulement si, $w \in \mathbb{R}_+$. Pour cela, on écrit $w = a + \mathrm{i}b$, $\mathrm{Re}(w) = a$ et $|w| = \sqrt{a^2 + b^2}$: c'est alors immédiat. ■

Plus généralement, on a l'inégalité : $\left|\mathrm{Re}(\overline{z}z')\right| \leqslant |z|\,|z'|$, avec égalité si, et seulement si, z et z' sont réellement liés. En effet, $|z|\,|z'|^2 = N(\overline{z}z') = \left(\mathrm{Re}(\overline{z}z')\right)^2 + \left(\mathrm{Im}(\overline{z}z')\right)^2$, et il suffit d'appliquer le lemme 9 de la page 249.

Proposition 12. Soient $z, z' \in \mathbb{C}$. Alors :

$$|\overline{z}| = |z|$$

$$\left|z + z'\right| \leqslant |z| + \left|z'\right|, \tag{5}$$

$$\left|zz'\right| = |z|\,\left|z'\right|. \tag{6}$$

Pour que l'inégalité (5) soit une égalité, il faut, et il suffit, que z et z' soient positivement liés.

Démonstration. Les égalités et l'inégalité portant sur des réels positifs, on peut les élever au carré. Elles découlent alors de la proposition 10 de la page précédente et (pour l'inégalité) du lemme 11. ■

L'inégalité (5) est appelée *(première) inégalité du triangle*, ce que l'on comprend aisément en interprétant le module comme une longueur. En l'appliquant à z' et $z - z'$, on obtient $|z - z'| \geqslant |z| - |z'|$; on prouve de même $|z - z'| \geqslant |z'| - |z|$ d'où la *deuxième inégalité du triangle* :

$$\left|z - z'\right| \geqslant \left||z| - |z'|\right|.$$

La norme algébrique n'est pas une norme au sens du cours d'analyse de L2, par exemple parce que l'on n'a pas $N(\lambda z) = |\lambda|\, N(z)$. En revanche, en vertu des règles ci-dessus, le module l'est. Nous allons maintenant passer à son interprétation géométrique. Pour un vecteur $\overrightarrow{V}$ d'affixe w, la longueur (ou norme euclidienne) vaut $|w|$; de même, pour deux points M, M' d'affixes respectifs z, z', la distance vaut $|z' - z|$, d'où les formules :

$$\|\overrightarrow{V}\| = \left|\mathrm{Aff}(\overrightarrow{V})\right| \qquad \text{et} \qquad MM' = \left|\mathrm{Aff}(M') - \mathrm{Aff}(M)\right|.$$

L'inégalité du triangle démontrée plus haut s'exprime en disant que, pour trois points A, B, C quelconques, on a $AC \leqslant AB + BC$, avec égalité si, et seulement si, $\overrightarrow{AB}$ et $\overrightarrow{BC}$ sont positivement liés, c'est-à-dire si les points sont alignés et B entre A et C. On le prouve en appliquant la règle (5) aux complexes $\mathrm{Aff}(B) - \mathrm{Aff}(A)$ et $\mathrm{Aff}(C) - \mathrm{Aff}(B)$.

Exemple. Le théorème de Pythagore dit que $AC^2 = AB^2 + BC^2$ si, et seulement si, $\overrightarrow{AB} \perp \overrightarrow{BC}$. En termes de complexes, selon la caractérisation (3) de la page 249 de l'orthogonalité, il s'agit de démontrer que $N(z + z') = N(z) + N(z')$ si, et seulement si, $\mathrm{Re}(\overline{z}z') = 0$. Pour cela, on applique la règle (4) de la page 250.

Exercice 4.

Soit I le milieu des points A' et A''.

Montrer que M est équidistant de A' et A'' si, et seulement si, $\overrightarrow{IM} \perp \overrightarrow{A'A''}$.

Solution. Notons z, z', z'', Z et W les affixes respectives de M, A', A'', $\overrightarrow{IM}$ et $\overrightarrow{A'A''}$. On a $z - z' = Z + W/2$ et $z - z'' = Z - W/2$, d'où :

$$\begin{aligned} MA' = MA'' &\iff |Z + W/2| = |Z - W/2| \\ &\iff N(Z + W/2) - N(Z - W/2) = 0 \\ &\iff \mathrm{Re}\,\overline{Z}W = 0, \end{aligned}$$

et l'on applique encore (3). Le même calcul permet d'ailleurs de caractériser les points strictement plus proches de A' que de A'' (par exemple) : ce sont ceux tels que $\mathrm{Re}\,\overline{Z}W < 0$. Ils forment donc un demi-plan ouvert de frontière la médiatrice, celui qui contient A', bien sûr.

Définition 8. Soient $z_0 \in \mathbb{C}$ et $r \in \mathbb{R}$. En vertu de l'identification de $\mathbb{C}$ avec le plan d'Argand-Cauchy, on appelle respectivement *disque ouvert, disque fermé et cercle de centre* z_0 *et de rayon* r les ensembles $\overset{\circ}{D}(z_0, r) := \{z \in \mathbb{C} \mid |z - z_0| < r\}$, $\overline{D}(z_0, r) := \{z \in \mathbb{C} \mid |z - z_0| \leqslant r\}$ et $\partial D(z_0, r) := \{z \in \mathbb{C} \mid |z - z_0| = r\}$.

Si $r < 0$, ces trois ensembles sont vides.
Si $r = 0$, le premier est vide et les deux autres réduits à $\{z_0\}$.

Exercice 5.

À quelle condition $\overset{\circ}{D}(z_0, r)$ et $\overset{\circ}{D}(z_0', r')$ se rencontrent-ils ?

Solution. Un dessin suggère la condition $|z_0' - z_0| < r + r'$. On vérifie d'abord qu'elle est nécessaire : si $z \in \overset{\circ}{D}(z_0, r) \cap \overset{\circ}{D}(z_0', r')$, il suffit d'appliquer l'inégalité du triangle à $z - z_0$ et $z_0' - z$.

Supposons la condition vérifiée. On cherchera un point commun sur le segment qui relie les centres, donc sous la forme $tz_0 + (1-t)z_0'$, $t \in [0,1]$ (exemple de la page 248). Pour que z soit dans les deux disques, il faut, et il suffit, que $t < \dfrac{r'}{|z_0' - z_0|}$ et $1 - t < \dfrac{r}{|z_0' - z_0|}$.

En vertu de l'hypothèse, le réel $t := \dfrac{r'}{r + r'}$ convient.

2.5 Racines carrées

On sait que les carrés des nombres réels sont exactement les nombres positifs, que 0 est le carré d'un seul réel (lui-même) et que tout réel strictement positif x est le carré d'exactement deux réels opposés. Ceux-ci sont appelés *les* racines carrées de x et celui qui est positif est appelé *la* racine carrée de x. Dans le corps des complexes, la situation est un peu plus compliquée : nous prouverons que tout complexe non nul admet deux racines carrées ; malheureusement, il n'y a pas de moyen naturel d'attacher à chaque complexe une racine carrée privilégiée.

Définition 9. On dit que $w \in \mathbb{C}$ est *une racine carrée* de $z \in \mathbb{C}$ si $w^2 = z$.

La seule racine carrée de 0 est 0 lui-même. Si un complexe non nul z admet une racine carrée w, il en a exactement deux, qui sont w et $-w$. En effet, z et w étant donnés tels que $w^2 = z$, on a les équivalences : $w'^2 = z \Leftrightarrow w'^2 = w^2 \Leftrightarrow w' = \pm w$, en vertu de l'identité remarquable $w'^2 - w^2 = (w' - w)(w' + w)$ et de l'intégrité de l'anneau $\mathbb{C}$. Un réel x strictement positif a donc toujours deux racines carrées réelles opposées dans $\mathbb{C}$, et nous conserverons la notation $\sqrt{x}$ pour désigner celle qui est positive. Si x est un réel négatif, ses racines carrées sont $\pm \mathrm{i}\sqrt{-x}$. En effet, cette notation a un sens puisque $-x > 0$; et, par ailleurs, $\left(\pm \mathrm{i}\sqrt{-x}\right)^2 = (-1)(-x) = x$.

Proposition 13. Tout complexe admet des racines carrées.

Démonstration. Si $z = x + \mathrm{i}y$ est un réel, cela résulte des arguments précédents. Nous supposons donc $y \neq 0$ et nous cherchons à résoudre l'équation $w^2 = z$ avec $w = u + v\mathrm{i}$. De l'égalité $(u + \mathrm{i}v)^2 = (u^2 - v^2) + 2\mathrm{i}uv$, on déduit l'équivalence :

$$w^2 = z \Longleftrightarrow \left\{u^2 - v^2 = x \qquad \text{et} \qquad 2uv = y\right\}.$$

Pour faciliter le calcul, remarquons que, si $w^2 = z$, on a $|w| = \sqrt{|z|}$, d'où l'égalité $u^2 + v^2 = \sqrt{x^2 + y^2}$. On en déduit l'équivalence légèrement améliorée :

$$w^2 = z \Longleftrightarrow \begin{cases} u^2 - v^2 = x \\ u^2 + v^2 = \sqrt{x^2 + y^2} \\ 2uv = y \end{cases}$$

Le système formé des deux premières égalités équivaut à :

$$u^2 = \frac{\sqrt{x^2 + y^2} + x}{2} \qquad \text{et} \qquad v^2 = \frac{\sqrt{x^2 + y^2} - x}{2}.$$

Les membres droits de ces égalités sont visiblement positifs. On en tire :

$$w^2 = z \Longleftrightarrow \begin{cases} u = \pm\sqrt{\dfrac{\sqrt{x^2+y^2}+x}{2}} \\ v = \pm\sqrt{\dfrac{\sqrt{x^2+y^2}-x}{2}} \\ 2uv = y \end{cases}$$

Les deux premières égalités entraînent $(uv)^2 = \dfrac{y^2}{4}$, d'où $2uv = \pm y$. Il s'agit donc de choisir les signes de u et v de manière compatible avec celui de y :

1. Si $y > 0$, on prend $u = \sqrt{\dfrac{\sqrt{x^2+y^2}+x}{2}}$ et $v = \sqrt{\dfrac{\sqrt{x^2+y^2}-x}{2}}$.
2. Si $y < 0$, on prend $u = \sqrt{\dfrac{\sqrt{x^2+y^2}+x}{2}}$ et $v = -\sqrt{\dfrac{\sqrt{x^2+y^2}-x}{2}}$.

Dans les deux cas, les racines carrées de z sont $\pm(u + \mathrm{i}v)$. ■

Exercice 6.

Calculer les racines carrées de i.

Solution. Les notations sont celles de la preuve ci-dessus, avec $x = 0$ et $y = 1$. Il faut résoudre le système : $\left\{u^2 - v^2 = 0 \text{ et } 2uv = 1\right\}$. De plus, $u^2+v^2 = 1$, d'où $u^2 = v^2 = \dfrac{1}{2}$ et u, v sont de même signe. Finalement, on trouve $\dfrac{\sqrt{2}}{2} + \mathrm{i}\dfrac{\sqrt{2}}{2}$ et $-\dfrac{\sqrt{2}}{2} - \mathrm{i}\dfrac{\sqrt{2}}{2}$.

Exercice 7.

Calculer les racines quatrièmes de i, puis les racines carrées de $-\mathrm{i}$.

Solution. Il s'agit donc des racines carrées de $z = \dfrac{\sqrt{2}}{2} + \mathrm{i}\dfrac{\sqrt{2}}{2}$ et de celles de $-z$. Remarquons que, si w est une racine carrée de z, alors $(\mathrm{i}.w)^2 = -w^2 = -z$: on obtiendra donc les racines carrées de $-z$ en multipliant par i celles de z. Cherchons ces dernières. Nous notons : $x = y = \dfrac{\sqrt{2}}{2}$ et posons : $\left\{u^2 - v^2 = \dfrac{\sqrt{2}}{2} \text{ et } 2uv = \dfrac{\sqrt{2}}{2}\right\}$. On sait de plus que $u^2 + v^2 = 1$, d'où $u = \pm\sqrt{\dfrac{1}{2}\left(1 + \dfrac{\sqrt{2}}{2}\right)}$, $v = \pm\sqrt{\dfrac{1}{2}\left(1 - \dfrac{\sqrt{2}}{2}\right)}$ et u, v doivent être de même signe.

En conclusion, les racines carrées de z sont $\pm\left(\sqrt{\dfrac{1}{2}\left(1 + \dfrac{\sqrt{2}}{2}\right)} + \mathrm{i}\sqrt{\dfrac{1}{2}\left(1 - \dfrac{\sqrt{2}}{2}\right)}\right)$.

En les multipliant par i, on trouve que les racines carrées de $-z$ sont :

$$\pm\left(\sqrt{\frac{1}{2}\left(1+\frac{\sqrt{2}}{2}\right)}-\mathrm{i}\sqrt{\frac{1}{2}\left(1-\frac{\sqrt{2}}{2}\right)}\right)\cdot$$

Pour déterminer les racines, carrées de $-\mathrm{i}$ le plus simple est de multiplier par i les racines carrées de i. On trouve ainsi $\pm\left(\frac{\sqrt{2}}{2}-\mathrm{i}\frac{\sqrt{2}}{2}\right)\cdot$

Y a-t-il un moyen raisonnable de choisir une racine carrée particulière pour tout nombre complexe ? Un critère serait la continuité, mais nous prouverons plus loin (page 258) que c'est impossible. On peut s'en convaincre intuitivement en cherchant, parmi les racines possibles, à en choisir « de proche en proche ». *Le lecteur est vivement invité à représenter sur un dessin l'argumentation qui va suivre.* Tous les complexes dont nous allons parler sont de module 1, donc situés sur le cercle trigonométrique. On part de $\sqrt{1}=1$. Les deux racines de -1 sont $\pm\mathrm{i}$ et elles sont équidistantes de 1, ce qui ne permet pas de trancher. Mais, pour les racines de i, la plus proche est $\frac{\sqrt{2}}{2}+\mathrm{i}\frac{\sqrt{2}}{2}$ que nous sélectionnons donc. Il est alors naturel de choisir i comme racine de -1, puis, toujours de proche en proche, $-\frac{\sqrt{2}}{2}+\mathrm{i}\frac{\sqrt{2}}{2}$ comme racine de $-\mathrm{i}$. Pourtant, si l'on était passé de 1 à $-\mathrm{i}$, le critère de proximité aurait plutôt conduit à sélectionner l'autre racine : $+\frac{\sqrt{2}}{2}-\mathrm{i}\frac{\sqrt{2}}{2}\cdot$
Il y a donc une contradiction dans notre processus de choix. C'est ce phénomène qui sous-tend la démonstration que nous donnerons page 258.
Voici un autre critère raisonnable possible pour choisir une racine carrée : il serait bon de disposer de la règle algébrique $\sqrt{xx'}=\sqrt{x}\sqrt{x'}$, dont on sait qu'elle est valide sur $\mathbb{R}_+$. Cela revient à postuler que l'application $x\mapsto\sqrt{z}$ est un morphisme du groupe $\mathbb{C}^*$ dans lui-même. En fait, *il n'existe pas de morphisme f du groupe multiplicatif $\mathbb{C}^*$ dans lui-même tel que l'on ait* $\forall z\neq 0,\ (f(z))^2=z$. Supposons en effet que f est un tel morphisme. On a $f(1)=1$, puisque f est un morphisme pour la multiplication. Donc $\big(f(-1)\big)^2=f\big((-1)^2\big)=f(1)=1$ (car f est un morphisme), d'où $f(-1)=\pm 1$. Mais, par hypothèse, $f(-1)$ doit être une racine carrée de -1, et $\big(f(-1)\big)^2=-1$, d'où $f(-1)=\pm\mathrm{i}$: contradiction.

2.5.1 Application à l'équation du second degré

Soient $a,b,c\in\mathbb{C}$, avec $a\neq 0$. La mise sous forme canonique du trinôme du second degré donne l'équivalence :

$$az^2+bz+c=0\iff\left(z+\frac{b}{2a}\right)^2=\frac{\Delta}{4a^2},$$

où l'on a noté $\Delta=b^2-4ac$ le discriminant du trinôme. Il n'est pas question ici de discuter le signe du discriminant, puisque c'est en général un complexe. Il n'y a que deux cas à considérer :

1. Si $\Delta=0$, l'équation admet la racine double $-\frac{b}{2a}\cdot$

2. Si le discriminant Δ n'est pas nul, il admet deux racines carrées complexes opposées δ_1 et δ_2. L'équation admet alors les deux racines :

$$z_1 = \frac{-b+\delta_1}{2a} \qquad \text{et} \qquad z_2 = \frac{-b+\delta_2}{2a}\ .$$

Dans tous les cas, on voit que la somme des racines est $-\dfrac{b}{a}$ et que leur produit est $\dfrac{c}{a}$.
Réciproquement, pour trouver deux nombres complexes de somme s et de produit p, on prend les racines de l'équation $z^2-sz+p=0$. Le lecteur retrouvera facilement les propriétés des racines lorsque a, b, c sont réels : elles sont toutes deux réelles ou bien complexes conjuguées.

3 Représentation trigonométrique

Nous étudierons en 5.2 la fonction exponentielle complexe $\exp : \mathbb{C} \to \mathbb{C}^*$. En principe, c'est cette étude qui fonde rigoureusement celle des fonctions trigonométriques. Pour simplifier l'exposé, nous allons commencer par définir et étudier l'exponentielle complexe à partir des connaissances acquises en terminale, que nous rappelons pour la commodité du lecteur :

1. La fonction exponentielle réelle est dérivable sur $\mathbb{R}$ et égale à sa dérivée. On a, pour x, y réels, $\mathrm{e}^{x+y} = \mathrm{e}^x\mathrm{e}^y$.
2. Les fonctions $\cos$ et $\sin$ sont 2π-périodiques et dérivables sur $\mathbb{R}$, et $\cos' = -\sin$, $\sin' = \cos$. Pour tout réel t, on a $\cos^2 t + \sin^2 t = 1$.

   Réciproquement, pour tout couple (x, y) de réels tels que $x^2 + y^2 = 1$, il existe $t \in \mathbb{R}$ tel que $\cos t = x$, $\sin t = y$ et ce réel est unique modulo $2\pi\mathbb{Z}$ (ou modulo 2π, ce qui revient au même).
3. Les formules d'addition, l'étude du signe de $\cos t$ et $\sin t$ à l'aide du cercle trigonométrique, et les valeurs spéciales sont également supposées acquises : consulter là-dessus un manuel de terminale.

Définition 10. La *fonction exponentielle complexe* $z \mapsto \exp z = \mathrm{e}^z$ de $\mathbb{C}$ dans $\mathbb{C}^*$ est définie par la formule $\exp(x+\mathrm{i}y) := \mathrm{e}^x\ (\cos y + \mathrm{i}\sin y)$.

Proposition 14. La fonction $\exp$ est un morphisme surjectif du groupe additif $\mathbb{C}$ sur le groupe multiplicatif $\mathbb{C}^*$, de noyau $2\mathrm{i}\pi\mathbb{Z}$.

Cela signifie que, pour deux complexes z_1, z_2 quelconques :

$$\exp(z_1+z_2) = \exp(z_1)\ \exp(z_2),$$

$$\exp z_1 = \exp z_2 \Longleftrightarrow z_1 \equiv z_2 \pmod{2\mathrm{i}\pi}.$$

Par définition, la notation $z_1 \equiv z_2 \pmod{2\mathrm{i}\pi}$ (congruence modulo $2\mathrm{i}\pi$) signifie que $z_1 - z_2$ est élément du groupe additif $2\mathrm{i}\pi\mathbb{Z}$ des nombres de la forme $2k\mathrm{i}\pi$, $k \in \mathbb{Z}$. Ces propriétés sont une conséquence facile des rappels précédents.

3.1 Le groupe des nombres complexes de module 1

Définition 11. On appelle *cercle unité* ou *cercle trigonométrique* de $\mathbb{C}$ le cercle $\mathbb{U} = \partial D(0,1)$ de centre 0 et de rayon 1.
C'est donc l'ensemble des nombres complexes de module 1 : $\mathbb{U} = \{z \in \mathbb{C} \mid |z| = 1\}$.

Théorème 15. (i) Le cercle unité $\mathbb{U}$ est un sous-groupe du groupe multiplicatif $\mathbb{C}^*$.
(ii) L'application $t \mapsto \mathrm{e}^{\mathrm{i}t}$ est un morphisme surjectif du groupe additif $\mathbb{R}$ sur le groupe multiplicatif $\mathbb{U}$, de noyau $2\pi\mathbb{Z}$.

Démonstration. (i) Il résulte de la propriété (6) de la page 251 que l'application $z \mapsto |z|$ de $\mathbb{C}^*$ est un morphisme du groupe multiplicatif $\mathbb{C}^*$ dans le groupe multiplicatif $\mathbb{R}_+^*$. L'ensemble $\mathbb{U}$ est le noyau de ce morphisme, donc un sous-groupe.
(ii) Notons $\mathbf{e}(t) = \mathrm{e}^{\mathrm{i}t}$. Il faut démontrer que $\mathbf{e}(t+t') = \mathbf{e}(t)\,\mathbf{e}(t')$, ce qui découle de la proposition 14 de la page ci-contre. Soit $z = x + \mathrm{i}y \in \mathbb{U}$: donc $x^2 + y^2 = 1$ et l'on sait qu'il existe $t \in \mathbb{R}$ tel que $\cos t = x$, $\sin t = y$, c'est-à-dire tel que $\mathbf{e}(t) = z$. Le morphisme $\mathbf{e}$ est donc surjectif. Enfin, $t \in \operatorname{Ker} \mathbf{e}$ si, et seulement si, $\cos t = 1$ et $\sin t = 0$, c'est-à-dire si $t \in 2\pi\mathbb{Z}$. ∎

Pratiquement, cela signifie que tout complexe u de module 1 peut s'écrire $\mathrm{e}^{\mathrm{i}t}$ pour un réel t bien défini modulo 2π ; autrement dit, c'est la classe d'équivalence du réel t pour la relation de congruence modulo $2\pi\mathbb{Z}$ (module I.1, section 5) qui est bien définie pour u donné. Nous noterons cette classe $t + 2\pi\mathbb{Z}$, ou encore $t \pmod{2\pi}$, ou même, pour alléger l'écriture, $t + 2k\pi$ (étant sous-entendu que k parcourt $\mathbb{Z}$). L'ensemble de ces classes est noté $\mathbb{R}/2\pi\mathbb{Z}$. En vertu du théorème, si l'on associe à $u \in \mathbb{U}$ la classe $t + 2\pi\mathbb{Z} \in \mathbb{R}/2\pi\mathbb{Z}$ telle que $u = \mathrm{e}^{\mathrm{i}t}$, et, de même, à $u' \in \mathbb{U}$ la classe $t' + 2\pi\mathbb{Z} \in \mathbb{R}/2\pi\mathbb{Z}$ telle que $u' = \mathrm{e}^{\mathrm{i}t'}$, alors la classe associée à uu' est $t + t' + 2\pi\mathbb{Z} \in \mathbb{R}/2\pi\mathbb{Z}$. Il est donc naturel de munir $\mathbb{R}/2\pi\mathbb{Z}$ de la loi de composition interne définie par :

$$(t + 2\pi\mathbb{Z}) + (t' + 2\pi\mathbb{Z}) := t + t' + 2\pi\mathbb{Z}.$$

On fait ainsi de $\mathbb{R}/2\pi\mathbb{Z}$ un groupe et le théorème nous dit que ce groupe est isomorphe à $\mathbb{U}$. Nous verrons dans le cours de L2 que ce groupe est le *quotient* du groupe $\mathbb{R}$ par son sous-groupe $2\pi\mathbb{Z}$.

Exemple. Les antécédents de 1 forment la classe $2\pi\mathbb{Z} = 0 + 2\pi\mathbb{Z} = 0 \pmod{2\pi}$. Ceux de -1 forment la classe $\pi + 2\pi\mathbb{Z} = \pi \pmod{2\pi}$.
Ceux de i sont les $\dfrac{\pi}{2} + 2k\pi$. Ceux de $-\mathrm{i}$ sont les $3\dfrac{\pi}{2} + 2k\pi$, ou encore les $-\dfrac{\pi}{2} + 2k\pi$.

Exercice 8.

Préciser les antécédents de $\pm\dfrac{\sqrt{2}}{2} \pm \mathrm{i}\dfrac{\sqrt{2}}{2}$, $\pm\dfrac{1}{2} \pm \mathrm{i}\dfrac{\sqrt{3}}{2}$ et $\pm\dfrac{\sqrt{3}}{2} \pm \mathrm{i}\dfrac{1}{2}$ ·

Solution. Il s'agit des classes modulo 2π de $\pi/4$, $3\pi/4$, $5\pi/4$, $7\pi/4$, $\pi/3$, $2\pi/3$, $4\pi/3$, $5\pi/3$, $\pi/6$, $5\pi/6$, $7\pi/6$, $11\pi/6$. Le lecteur consciencieux attribuera soigneusement les signes des parties réelles et imaginaires.

Exercice 9.
Quelles sont les racines carrées de $\mathrm{e}^{\mathrm{i}t} \in \mathbb{U}$?
Solution. Puisque $\left|u'^2\right| = 1 \Rightarrow |u'| = 1$, elles sont de la forme $u' = \mathrm{e}^{\mathrm{i}t'}$ avec $\mathrm{e}^{2\mathrm{i}t'} = \mathrm{e}^{\mathrm{i}t}$, c'est-à-dire $2t' \equiv t \pmod{2\pi}$.
Elles correspondent donc aux deux classes $\dfrac{t}{2} + 2\pi\mathbb{Z}$ et $\dfrac{t}{2} + \pi + 2\pi\mathbb{Z}$.

Théorème 16. (i) L'application $t \mapsto \mathrm{e}^{\mathrm{i}t}$ de $\mathbb{R}$ dans $\mathbb{C}$ est continue et son image est $\mathbb{U}$.
(ii) Soit f une fonction continue de $\mathbb{U}$ dans $\mathbb{R}$. La fonction f admet alors un minimum $\min f$ et un maximum $\max f$ atteints, et son image est le segment : $f(\mathbb{U}) = [\min f, \max f]$.

Démonstration. (i) Notons temporairement $\mathbf{e}(t) = \mathrm{e}^{\mathrm{i}t}$. Puisque $\mathrm{e}^{\mathrm{i}t} = \cos t + \mathrm{i}\sin t$, la continuité de $\mathbf{e}$ est évidente (module IV.6, section 2). De plus, $\left|\mathrm{e}^{\mathrm{i}t}\right| = (\cos t)^2 + (\sin t)^2 = 1$, et l'image de $\mathbf{e}$ est bien incluse dans $\mathbb{U}$. La surjectivité a déjà été démontrée.
(ii) La fonction $g = f \circ \mathbf{e}$ est continue sur le segment $[0, 2\pi]$, à valeurs dans $\mathbb{R}$; elle admet donc des extrema atteints (module IV.2, théorème 32), et $g([0, 2\pi]) = [\min g, \max g]$. Comme l'application $\mathbf{e}$ est surjective de $[0, 2\pi]$ sur $\mathbb{U}$, on conclut en remarquant que $\min f = \min g$, $\max f = \max g$ et $f(\mathbb{U}) = g([0, 2\pi])$. ■

On verra dans le module IV.7, définition 5 que $\mathbb{U}$ est une partie compacte du plan, et, dans le cours d'analyse de L2, qu'elle est connexe par arcs (comme image d'une courbe paramétrée de classe $\mathcal{C}^1$), donc connexe. L'assertion (ii) sera alors un cas particulier de théorèmes généraux. À titre d'application, nous allons démontrer qu'il n'est pas possible de définir de manière continue une fonction « racine carrée » sur $\mathbb{U}$ (et donc, *a fortiori*, sur $\mathbb{C}^*$ ou sur $\mathbb{C}$). Soit en effet $f : \mathbb{U} \to \mathbb{U}$ telle que $\forall u \in \mathbb{U}, (f(u))^2 = u$. Le carré de $\varepsilon(t) = f\left(\mathrm{e}^{\mathrm{i}t}\right)\mathrm{e}^{-\mathrm{i}t/2}$ est $\mathrm{e}^{\mathrm{i}t}\mathrm{e}^{-\mathrm{i}t} = 1$, c'est donc une application de $\mathbb{R}$ dans $\{+1, -1\}$. Comme $f\left(\mathrm{e}^{\mathrm{i}(t+2\pi)}\right) = f\left(\mathrm{e}^{\mathrm{i}t}\right)$ et $\mathrm{e}^{\mathrm{i}(t+2\pi)/2} = -\mathrm{e}^{\mathrm{i}t/2}$, on a $\varepsilon(t + 2\pi) = -\varepsilon(t)$. La fonction ε change donc de signe. Si elle était continue, elle s'annulerait d'après le théorème des valeurs intermédiaires, ce qui n'est pas le cas : elle n'est donc pas continue, et f ne peut pas l'être non plus.

Les mathématiciens du XIXe siècle parlaient de « la fonction multiforme $\sqrt{z}$ ». C'est pour comprendre de telles fonctions que Riemann a édifié la théorie des *surfaces de Riemann*, dont l'importance est fondamentale dans les mathématiques du vingtième siècle.

3.2 Racines de l'unité

Définition 12. Soit $n \in \mathbb{N}^*$. Une *racine $n^{\text{ème}}$ de l'unité* est un complexe z tel que $z^n = 1$. L'ensemble des racines $n^{\text{èmes}}$ de l'unité est noté μ_n. Une *racine de l'unité* est un complexe z qui est une racine $n^{\text{ème}}$ de l'unité pour un $n \in \mathbb{N}^*$ au moins. L'ensemble des racines de l'unité est noté μ_∞ (certains manuels le notent U_n).

Ainsi, $\mu_n := \{z \in \mathbb{C} \mid z^n = 1\}$ et $\mu_\infty = \bigcup_{n \in \mathbb{N}^*} \mu_n$. Si $p = nm$ est un multiple de n, l'implication $z^n = 1 \Rightarrow z^p = (z^n)^m = 1$ montre que $\mu_n \subset \mu_p$, ce qui justifie le terme « au moins » dans la définition.

Exemple. Il est immédiat que $\mu_1 = \{1\}$, $\mu_2 = \{+1, -1\}$.
On trouve facilement que $\mu_4 = \{+1, -1, +\mathrm{i}, -\mathrm{i}\}$ en remarquant que $z^4 = 1 \Leftrightarrow z^2 = \pm 1$.
Les exercices d'application sur les racines carrées permettent de trouver μ_8 :

$$\mu_8 = \left\{+1, -1, +\mathrm{i}, -\mathrm{i}, -\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\mathrm{i}, \frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}\mathrm{i}, \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\mathrm{i}, -\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}\mathrm{i}\right\} \cdot$$

Les classes dans $\mathbb{R}/2\pi\mathbb{Z}$ de ces nombres sont les $k\pi/4 + 2\pi\mathbb{Z}$ pour $k = 0, \ldots, 7$. Déterminons μ_3.
Comme $z^3 - 1 = (z-1)(z^2 + z + 1)$, μ_3 contient 1 et les racines de l'équation $z^2 + z + 1 = 0$, qui sont $j := -\dfrac{1}{2} + \mathrm{i}\dfrac{\sqrt{3}}{2} = \mathrm{e}^{2\mathrm{i}\pi/3}$ et son carré $j^2 = -\dfrac{1}{2} - \mathrm{i}\dfrac{\sqrt{3}}{2} = \mathrm{e}^{4\mathrm{i}\pi/3}$, d'où $\mu_3 = \{1, j, j^2\}$.

Exercice 10.

Déterminer μ_5. On ramènera une équation de degré 4 à une équation de degré 2 grâce à l'inconnue auxiliaire $Z = z + z^{-1}$. Reconnaître $\mathrm{e}^{2\mathrm{i}\pi/5}$ dans la liste obtenue.
Solution. Comme $z^5 - 1 = (z-1)(z^4 + z^3 + z^2 + z + 1)$, μ_5 contient 1 et les racines de l'équation $z^4 + z^3 + z^2 + z + 1 = 0$. Comme 0 n'en est pas solution, on la divise par z^2, ce qui la ramène à $z^2 + z + 1 + z^{-1} + z^{-2} = 0$. Posant $Z = z + z^{-1}$, ce qui équivaut à $z^2 - Zz + 1 = 0$, on a $Z^2 = z^2 + z^{-2} + 2$, d'où l'équation : $Z^2 + Z - 1 = 0$. Celle ci admet deux racines réelles : $Z' = \dfrac{-1+\sqrt{5}}{2}$ et $Z'' = \dfrac{-1-\sqrt{5}}{2} \cdot$ Il faut maintenant résoudre les équations du second degré $z^2 - Z'z + 1 = 0$ et $z^2 - Z''z + 1 = 0$. Leurs discriminants respectifs sont $\Delta' = \dfrac{-5-\sqrt{5}}{2}$ et $\Delta'' = \dfrac{-5+\sqrt{5}}{2}$: ils sont tous deux négatifs.
Outre $z_0 = 1$, on trouve finalement $\mu_5 = \{z_0, z_1, z_2, z_3, z_4\}$, avec :

$$\begin{aligned}
z_1 &= \frac{Z'}{2} + \mathrm{i}\frac{\sqrt{-\Delta'}}{2} = \frac{-1+\sqrt{5}}{4} + \mathrm{i}\frac{\sqrt{10+2\sqrt{5}}}{4},\\
z_2 &= \frac{Z'}{2} - \mathrm{i}\frac{\sqrt{-\Delta'}}{2} = \frac{-1+\sqrt{5}}{4} - \mathrm{i}\frac{\sqrt{10+2\sqrt{5}}}{4},\\
z_3 &= \frac{Z''}{2} + \mathrm{i}\frac{\sqrt{-\Delta''}}{2} = \frac{-1-\sqrt{5}}{4} + \mathrm{i}\frac{\sqrt{10-2\sqrt{5}}}{4},\\
z_4 &= \frac{Z''}{2} - \mathrm{i}\frac{\sqrt{-\Delta''}}{2} = \frac{-1-\sqrt{5}}{4} - \mathrm{i}\frac{\sqrt{10-2\sqrt{5}}}{4} \cdot
\end{aligned}$$

Le complexe $\mathrm{e}^{2\mathrm{i}\pi/5}$ est une racine 5ème de l'unité car $\left(\mathrm{e}^{2\mathrm{i}\pi/5}\right)^5 = \mathrm{e}^{2\mathrm{i}\pi} = 1$.
D'après la liste trouvée, le seul élément de μ_5 d'abscisse et d'ordonnée strictement positives est $z_1 = \dfrac{-1+\sqrt{5}}{4} + \mathrm{i}\dfrac{\sqrt{10+2\sqrt{5}}}{4}$: comme $\cos 2\pi/5$ et $\sin 2\pi/5$ sont strictement positifs, on a $\mathrm{e}^{2\mathrm{i}\pi/5} = z_1$.

Proposition 17. Soit $n \in \mathbb{N}^*$. L'ensemble μ_n des racines $n^{\text{èmes}}$ de l'unité est un sous-groupe cyclique de $\mathbb{U}$. Ses éléments sont les $\mathrm{e}^{2k\mathrm{i}\pi/n}$, où k parcourt les entiers modulo n ; par exemple : $\mu_n = \left\{\mathrm{e}^{2k\mathrm{i}\pi/n} \mid k \in [\![0, n-1]\!]\right\}$.

Démonstration. On a les implications : $z^n = 1 \Rightarrow |z|^n = 1 \Rightarrow |z| = 1$, d'où l'inclusion $\mu_n \subset \mathbb{U}$. De plus, l'application $z \mapsto z^n$ est un morphisme du groupe $\mathbb{C}^*$ dans lui-même de noyau μ_n, qui est donc bien un sous-groupe de $\mathbb{U}$. Puisque $\mu_n \subset \mathbb{U}$, ses éléments sont de la forme $z = \mathrm{e}^{\mathrm{i}t}$, avec $\mathrm{e}^{\mathrm{i}nt} = 1$, ce qui équivaut à $nt = 2k\pi$, avec $k \in \mathbb{Z}$, d'où $z = z_k = \mathrm{e}^{2k\mathrm{i}\pi/n}$.

Par ailleurs, $z_k = z_{k'} \Leftrightarrow \mathrm{e}^{2k\mathrm{i}\pi/n} = \mathrm{e}^{2k'\mathrm{i}\pi/n} \Leftrightarrow 2k\pi/n \equiv 2k'\pi/n \pmod{2\pi}$, ce qui équivaut à $k \equiv k' \pmod{n}$, d'où la description des éléments de μ_n. Le groupe μ_n a n éléments et il formé des puissances de $\mathrm{e}^{2\mathrm{i}\pi/n}$, donc cyclique. ■

Corollaire 18. Les seuls sous-groupes finis de $\mathbb{C}^*$ sont les μ_n.

Démonstration. Soit en effet G un sous-groupe fini de $\mathbb{C}^*$, et soit n l'ordre de G. D'après le théorème de Lagrange (module II.2), tout élément x de G vérifie $x^n = 1$, d'où l'inclusion $G \subset \mu_n$.

Mais, ces deux groupes ayant même nombre d'éléments, ils sont égaux : $G = \mu_n$. ■

Exemple. Le groupe μ_6 est engendré par $\mathrm{e}^{2\mathrm{i}\pi/6} = \mathrm{e}^{\mathrm{i}\pi/3} = \dfrac{1}{2} + \mathrm{i}\dfrac{\sqrt{3}}{2}$.

Soit K un corps commutatif. On démontre que tout sous-groupe fini de K^* est cyclique et qu'il y en a au plus un d'ordre n donné. Si K est algébriquement clos de caractéristique nulle, il y a exactement un sous-groupe pour tout ordre $n \in \mathbb{N}^*$. Si K est de caractéristique $p > 0$, cela ne vaut que pour les n non divisibles par p. La spécificité de $\mathbb{C}$ est que l'on peut décrire explicitement les racines de l'unité à l'aide de fonctions.

Définition 13. Une *racine primitive $n^{\text{ème}}$ de l'unité* est un générateur de μ_n.

Il résulte de l'exercice 5 de la page 132 (module II.2) que les générateurs de μ_n sont les $\left(\mathrm{e}^{2\mathrm{i}\pi/n}\right)^k = \mathrm{e}^{2k\mathrm{i}\pi/n}$, où $0 \leqslant k \leqslant n-1$ et où k est premier avec n.

Il y a donc $\varphi(n)$ tels générateurs (φ est l'indicatrice d'Euler).

Exercice 11.

Quelles sont les racines primitives de μ_2, μ_3, μ_4, μ_5 et μ_6 ?

Solution. Dans μ_2, la seule racine primitive est -1. Dans μ_3 et μ_5 toutes les racines autres que 1 sont primitives ; il en est de même dans tout μ_p, p étant premier.

Dans μ_4, il y a i et $-\mathrm{i}$. Dans μ_6, il y a $\mathrm{e}^{\mathrm{i}\pi/3}$ et $\mathrm{e}^{-\mathrm{i}\pi/3}$.

Soit $z = \mathrm{e}^{2k\mathrm{i}\pi/n} \neq 1$ une racine $n^{ème}$ non triviale de l'unité. Si l'on écrit k/n sous forme de fraction irréductible ℓ/p, on a $z = \mathrm{e}^{2\ell\mathrm{i}\pi/p}$ et ℓ et p sont premiers entre eux ; z est donc une racine primitive $p^{ème}$ de l'unité et p est le plus petit naturel non nul tel que $z^p = 1$: c'est *l'ordre* de z.

Exercice 12.

Soit $n \geqslant 3$. On note $z_k = \mathrm{e}^{2k\mathrm{i}\pi/n}$ et M_k le point du cercle unité d'affixe z_k. On verra plus loin que $M_0, \ldots, M_{n-1}$ sont les sommets d'un polygone régulier.
Calculer le périmètre de ce polygone et l'aire du domaine qu'il entoure, puis les limites de ces nombres quand $n \to +\infty$.

Solution. La longueur M_kM_{k+1} vaut $|z_{k+1} - z_k| = |z_1^{k+1} - z_1^k|$, soit $|z_1 - 1| = |\mathrm{e}^{2\mathrm{i}\pi/n} - 1|$ (car $|z_1^k| = 1$). Les n côtés ont donc pour longueur :

$$\sqrt{(\cos 2\pi/n - 1)^2 + (\sin 2\pi/n)^2} = \sqrt{2 - 2\cos 2\pi/n} = \sqrt{4\sin^2 \pi/n},$$

c'est-à-dire $2\sin \pi/n$ car ce dernier est visiblement positif. Le périmètre est donc $2n \sin \pi/n$ et sa limite lorsque $n \to +\infty$ est 2π, la circonférence du cercle, ce qui est bien normal. Par ailleurs, les triangles OM_kM_{k+1} sont tous égaux (trois côtés égaux) donc ont même aire. On calcule celle de OM_0M_1 en prenant la base OM_0 (longueur 1) et la hauteur menée de M_1 (sa longueur est $\sin \frac{2\pi}{n}$, l'ordonnée de M_1) : son aire est donc $\frac{1}{2} \sin \frac{2\pi}{n}$ et celle enclose par le polygone est $\frac{n}{2} \sin \frac{2\pi}{n}$.
La limite lorsque n tend vers $+\infty$ est π, l'aire du disque unité, ce qui est encore normal.

Exercice 13.

Soit $z = \mathrm{e}^{4\mathrm{i}\pi/5}$: c'est une racine primitive 5ème de l'unité, qui engendre donc μ_5. Quelle figure obtient-on en reliant dans cet ordre les points du cercle unité d'affixes $z^0, z^1, z^2, z^3, z^4, z^5$?

Solution. Il s'agit d'un parcours cyclique des 5 sommets d'un pentagone régulier, lequel est un cas particulier du polygone de l'exercice précédent. Toutefois, ici, les sommets ne sont pas parcourus consécutivement mais de 2 en 2. Les segments tracés sont donc les 5 diagonales du pentagone. On obtient un pentagone étoilé[3].

Pour être précis, en examinant les signes des parties réelles et imaginaires des z_k déterminés précédemment, on trouve $z = \mathrm{e}^{2\mathrm{i}\pi/5} = z_1$, $z^2 = \mathrm{e}^{4\mathrm{i}\pi/5} = z_3$, $z^3 = \mathrm{e}^{6\mathrm{i}\pi/5} = z_4$ et $z^4 = \mathrm{e}^{8\mathrm{i}\pi/5} = z_2$. Le lecteur est invité à placer ces points sur le cercle $\mathbb{U}$. On obtient de même un polygone régulier étoilé chaque fois que l'on parcourt les puissances d'une racine primitive $n^{ème}$ autre que $z = \mathrm{e}^{2\mathrm{i}\pi/n}$ ou son inverse $z^{-1} = \mathrm{e}^{-2\mathrm{i}\pi/n}$.

---

[3] Un tel pentagone apparaît sur le drapeau du Maroc (c'est *l'étoile Chérifienne*). Si l'on veut obtenir *l'étoile de David*, composée de deux triangles équilatéraux (elle figure sur le drapeau d'Israël), il faut utiliser les puissances de $\mathrm{e}^{4\mathrm{i}\pi/6}$; mais on ne peut pas le faire en reliant en une seule passe des puissances successives, il y aura nécessairement deux séries de trois points : chacun des deux seuls générateurs de μ_6 engendreraient un hexagone convexe et non étoilé.

3.3 Arguments d'un nombre complexe

Théorème 19.
(i) Les applications : $(\rho, u) \mapsto \rho\, u$, de $\mathbb{R}_+^* \times \mathbb{U}$ dans $\mathbb{C}^*$, et $z \mapsto \left(|z|, \dfrac{z}{|z|}\right)$, de $\mathbb{C}^*$ dans $\mathbb{R}_+^* \times \mathbb{U}$, sont des isomorphismes de groupes multiplicatifs. Ces isomorphismes sont réciproques l'un de l'autre.
(ii) Tout nombre complexe non nul z peut s'écrire : $z = \rho\, \mathrm{e}^{\mathrm{i}\theta}$, où $\rho \in \mathbb{R}_+^*$ est unique et où $\theta \in \mathbb{R}$ est déterminé modulo 2π.

Démonstration. L'égalité $z = \rho\, u$ avec $\rho > 0$ et $|u| = 1$ équivaut à $\rho = |z|$ et $u = \dfrac{z}{|z|}$; on a donc bien deux bijections réciproques l'une de l'autre. Ce sont des morphismes en vertu de la formule (6) de la page 251. La seconde assertion est conséquence triviale de la première. ■

Définition 14. Soit $z \in \mathbb{C}^*$. Toute écriture $z = \rho\, \mathrm{e}^{\mathrm{i}\theta},\ \rho \in \mathbb{R}_+^*,\ \theta \in \mathbb{R}$ est appelée *représentation trigonométrique*, ou encore *écriture trigonométrique* de z.

Voici quelques règles de calcul sur les représentations trigonométriques :

$$\left(\rho\, \mathrm{e}^{\mathrm{i}\theta}\right)\left(\rho' \mathrm{e}^{\mathrm{i}\theta'}\right) = (\rho\rho')\, \mathrm{e}^{\mathrm{i}(\theta+\theta')}, \qquad \left(\rho\, \mathrm{e}^{\mathrm{i}\theta}\right)^{-1} = \rho^{-1}\, \mathrm{e}^{-\mathrm{i}\theta},$$

$$\left(\rho\, \mathrm{e}^{\mathrm{i}\theta}\right)^{n} = \rho^{n}\, \mathrm{e}^{n\mathrm{i}\theta}, \qquad -\rho\, \mathrm{e}^{\mathrm{i}\theta} = \rho\, \mathrm{e}^{\mathrm{i}(\theta+\pi)}.$$

Il est d'usage, dans l'étude des courbes planes, de repérer un point du plan par ses *coordonnées polaires* (module III.2). Celles-ci sont presque identiques au module et à l'argument, à cela près que l'on s'autorise des « modules négatifs » : le point de coordonnées polaires $(\rho, \theta) \in \mathbb{R}^2$ est le point d'affixe $\rho \mathrm{e}^{\mathrm{i}\theta}$. Si $\rho > 0$, c'est bien une représentation trigonométrique. Si $\rho < 0$, l'affixe du point correspondant admet pour représentation trigonométrique $|\rho|\, \mathrm{e}^{\mathrm{i}(\theta+\pi)}$. Si $\rho = 0$, l'angle θ n'a aucune signification.

Exercice 14.
Calculer $(1+\mathrm{i})^n$ et préciser sa position dans le plan, c'est-à-dire les signes de ses parties réelle et imaginaire.
Solution. Le complexe $1+\mathrm{i}$ peut s'écrire $\sqrt{2}\, \mathrm{e}^{\mathrm{i}\pi/4}$. Ses puissances sont donc de la forme : $(1+\mathrm{i})^n = 2^{n/2}\, \mathrm{e}^{n\mathrm{i}\pi/4}$. Leur position dépend seulement de la classe de n modulo 8 :

$$\begin{aligned}
n \equiv 0 \pmod 8 &\implies \mathrm{Re}\big((1+\mathrm{i})^n\big) > 0 \text{ et } \mathrm{Im}\big((1+\mathrm{i})^n\big) = 0,\\
n \equiv 1 \pmod 8 &\implies \mathrm{Re}\big((1+\mathrm{i})^n\big) > 0 \text{ et } \mathrm{Im}\big((1+\mathrm{i})^n\big) > 0,\\
n \equiv 2 \pmod 8 &\implies \mathrm{Re}\big((1+\mathrm{i})^n\big) = 0 \text{ et } \mathrm{Im}\big((1+\mathrm{i})^n\big) > 0,\\
n \equiv 3 \pmod 8 &\implies \mathrm{Re}\big((1+\mathrm{i})^n\big) < 0 \text{ et } \mathrm{Im}\big((1+\mathrm{i})^n\big) > 0,\\
n \equiv 4 \pmod 8 &\implies \mathrm{Re}\big((1+\mathrm{i})^n\big) < 0 \text{ et } \mathrm{Im}\big((1+\mathrm{i})^n\big) = 0,\\
n \equiv 5 \pmod 8 &\implies \mathrm{Re}\big((1+\mathrm{i})^n\big) < 0 \text{ et } \mathrm{Im}\big((1+\mathrm{i})^n\big) < 0,\\
n \equiv 6 \pmod 8 &\implies \mathrm{Re}\big((1+\mathrm{i})^n\big) = 0 \text{ et } \mathrm{Im}\big((1+\mathrm{i})^n\big) < 0,\\
n \equiv 7 \pmod 8 &\implies \mathrm{Re}\big((1+\mathrm{i})^n\big) > 0 \text{ et } \mathrm{Im}\big((1+\mathrm{i})^n\big) < 0.
\end{aligned}$$

La prochaine définition pourra sembler compliquée, mais il n'y a pas moyen de faire mieux. Le problème est qu'il y a une infinité de réels θ tels que $\dfrac{z}{|z|} = \mathrm{e}^{\mathrm{i}\theta}$, et que l'on ne sait pas en choisir un par une méthode qui soit satisfaisante dans tous les cas. Ce qui transparaît ici, c'est l'impossibilité d'une définition simple du logarithme complexe : il n'existe pas de fonction continue $\log : \mathbb{C}^* \to \mathbb{C}$ telle que, pour tout complexe $z \neq 0$, $\exp(\log z) = z$. De même, il n'existe pas de fonction continue $\theta : \mathbb{C}^* \to \mathbb{R}$ telle que $\mathrm{e}^{\mathrm{i}\theta(z)} = z$. En effet, s'il existait une telle fonction, en prenant $f(z) = \mathrm{e}^{\mathrm{i}\theta(z)/2}$, on obtiendrait une fonction « racine carrée » qui soit continue, ce qui est impossible, comme on l'a remarqué à la page 258.

Définition 15.
Soit $z = \rho\,\mathrm{e}^{\mathrm{i}\theta}$, $\rho \in \mathbb{R}_+^*$, $\theta \in \mathbb{R}$ un complexe non nul écrit sous forme trigonométrique.
(i) La classe $\theta + 2\pi\mathbb{Z} \in \mathbb{R}/2\pi\mathbb{Z}$ (qui est bien déterminée d'après le théorème 19 de la page ci-contre) est appelée *(l') argument* de z et notée $\arg z$.
(ii) Tout élément de cette classe, c'est-à-dire tout $\theta \in \mathbb{R}$ tel que $\dfrac{z}{|z|} = \mathrm{e}^{\mathrm{i}\theta}$ est appelé *(un) argument* de z : on note alors $\theta \in \arg z$.
(iii) L'unique représentant de $\arg(z)$ dans l'intervalle $]-\pi, \pi]$, c'est-à-dire l'unique $\theta \in \arg z$ tel que $\theta \in\ ]-\pi, \pi]$, est appelé *argument principal de* z et on le note $\operatorname{Arg} z$.
La fonction $z \mapsto \operatorname{Arg} z$ est appelée *détermination principale de l'argument*.

En pratique, on écrira souvent $\theta = \arg z$ pour dire que θ est un argument de z. C'est toutefois une notation ambiguë, donc dangereuse : si par exemple, en un point d'un calcul, on écrit $\theta = \arg z$ et en un autre point $\theta' = \arg z$, on peut en déduire que $\theta' \equiv \theta \pmod{2\pi}$ mais pas nécessairement que $\theta' = \theta$.

Les cas de base suivants sont à connaître par cœur. Les réels strictement positifs sont les complexes d'argument $0 + 2\pi\mathbb{Z}$. Les réels strictement négatifs sont les complexes d'argument $\pi + 2\pi\mathbb{Z}$. Les éléments de $\mathbb{R}_+^*\,\mathrm{i}$ sont les complexes d'argument $\pi/2 \pmod{2\pi}$. Les éléments de $\mathbb{R}_-^*\,\mathrm{i}$ sont les complexes d'argument $-\pi/2 \pmod{2\pi}$. Soit $z \in \mathbb{C}^*$.
Alors $\operatorname{Arg} z \geqslant 0$ si, et seulement si, $\operatorname{Im} z \geqslant 0$. De même, $\operatorname{Arg} z \in [-\pi/2, \pi/2]$ si, et seulement si, $\operatorname{Re} z \geqslant 0$.

Exercice 15.
Soit $x \in\ ]-\pi, \pi]$. Déterminer le module et l'argument de $\mathrm{e}^{\mathrm{i}x} + 1$.
Solution. Si $x = \pi$, $\mathrm{e}^{\mathrm{i}x} + 1 = 0$. Supposons donc $x \in\ ]-\pi, \pi[$. Alors :

$$\left|\mathrm{e}^{\mathrm{i}x} + 1\right|^2 = \left(\mathrm{e}^{\mathrm{i}x} + 1\right)\left(\mathrm{e}^{-\mathrm{i}x} + 1\right) = 2 + 2\cos x = 4\cos^2(x/2),$$

d'où le module : $\left|\mathrm{e}^{\mathrm{i}x} + 1\right| = 2\cos(x/2)$ (qui est bien positif car $x/2 \in\ ]-\pi/2, \pi/2[$). Les formules de duplication du sinus et du cosinus donnent : $\mathrm{e}^{\mathrm{i}x} + 1 = 2\cos(x/2)\ \mathrm{e}^{\mathrm{i}x/2}$. L'argument de $\mathrm{e}^{\mathrm{i}x} + 1$ est donc $x/2 \pmod{2\pi}$.

Exercice 16.
Soit $x \in]-\pi, \pi]$. Calculer le module et l'argument de $\mathrm{e}^{\mathrm{i}x} - 1$ en commençant par mettre en facteur $\mathrm{e}^{\mathrm{i}x/2}$.
Solution. On a : $\mathrm{e}^{\mathrm{i}x} - 1 = \mathrm{e}^{\mathrm{i}x/2}\left(\mathrm{e}^{\mathrm{i}x/2} - \mathrm{e}^{-\mathrm{i}x/2}\right) = 2\mathrm{i}\sin(x/2)\,\mathrm{e}^{\mathrm{i}x/2}$.
Si $x = 0$, $\mathrm{e}^{\mathrm{i}x} - 1 = 0$. Si $x \in]0, \pi]$, $\sin(x/2) > 0$, le module de $\mathrm{e}^{\mathrm{i}x} - 1$ est $2\sin(x/2)$ et son argument est $x/2 + \pi/2 \pmod{2\pi}$. Si $x \in]-\pi, 0[$, $\sin(x/2) < 0$, le module de $\mathrm{e}^{\mathrm{i}x} - 1$ est $-2\sin(x/2)$ et son argument est $x/2 - \pi/2 \pmod{2\pi}$, ou encore $x/2 + 3\pi/2 \pmod{2\pi}$.

Voici quelques règles de calcul sur l'argument principal. Leur complication est une difficulté inhérente au problème : *le choix d'une détermination de l'argument dans une formule requiert presque toujours une discussion et toujours beaucoup de soin*. Les complexes z et z' sont supposés non nuls.

$$\begin{aligned}
\operatorname{Arg}(zz') &= \operatorname{Arg} z + \operatorname{Arg} z' \\
&\quad \text{ou } \operatorname{Arg} z + \operatorname{Arg} z' - 2\pi \\
&\quad \text{ou } \operatorname{Arg} z + \operatorname{Arg} z' + 2\pi, \\
\operatorname{Arg}\left(\frac{1}{z}\right) &= -\operatorname{Arg} z \\
&\quad \text{ou } \operatorname{Arg} z \text{ dans le cas où } z \in \mathbb{R}_-^*, \\
\operatorname{Arg}(-z) &= \operatorname{Arg} z + \pi \\
&\quad \text{ou } \operatorname{Arg} z - \pi.
\end{aligned}$$

3.4 Racines $n^{\text{èmes}}$ des nombres complexes

Définition 16. Soit $n \geqslant 2$ un entier naturel.
On appelle *racine $n^{\text{ème}}$ du complexe z* tout complexe w tel que $w^n = z$.

La seule racine $n^{\text{ème}}$ de 0 est donc 0.

Proposition 20. Soit $z \in \mathbb{C}^*$ de module ρ et d'argument $\theta + 2\pi\mathbb{Z}$. Alors z admet n racines $n^{\text{èmes}}$ distinctes $z_1, \ldots, z_n$, toutes de module $\rho^{1/n}$, et d'arguments respectifs $\theta_1 + 2\pi\mathbb{Z}, \ldots, \theta_n + 2\pi\mathbb{Z}$, où $\theta_k = \theta/n + 2k\pi/n$, $k = 1, \ldots, n$. Toutes ces racines s'obtiennent à partir de l'une quelconque d'entre elles successivement multipliée par les racines $n^{\text{èmes}}$ de l'unité.

Démonstration. On cherche ces racines sous la forme $z' = \rho'\,\mathrm{e}^{\mathrm{i}\theta'}$, avec les contraintes $\rho'^n = \rho$ et $n\theta' \equiv \theta \pmod{n}$. La première condition équivaut à $\rho' = \rho^{1/n}$. La seconde se réécrit $n\theta' = \theta + 2k\pi$, $k \in \mathbb{Z}$, soit $\theta' = \theta + 2k\pi/n$, $k \in \mathbb{Z}$. La valeur de $\mathrm{e}^{\mathrm{i}\theta'}$ ne dépend alors que de la classe de congruence $k \pmod{n}$ d'où la première assertion.
Fixant $k \in [\![1, n]\!]$, on voit alors que $z_\ell = z_k \mathrm{e}^{2(\ell-k)\mathrm{i}\pi/n}$, où $\ell - k = 1 - k, \ldots, n - k$ parcourt toutes les classes d'entiers modulo n, d'où la deuxième assertion. ∎

Exemple. Les racines cubiques du réel non nul a sont le réel $a^{1/3}$ (c'est-à-dire $\sqrt[3]{a}$) et les complexes conjugués $a^{1/3}j$ et $a^{1/3}j^2$, avec la notation usuelle $j = \mathrm{e}^{2\mathrm{i}\pi/3} = \dfrac{-1}{2} + \mathrm{i}\dfrac{\sqrt{3}}{2}$.

3.4.1 Application à l'équation du troisième degré

Soient $a, b, c, d \in \mathbb{C}$, $a \neq 0$. Alors l'équation du troisième degré $aw^3 + bw^2 + cw + d = 0$ se ramène à l'équation $z^3 + pz + q = 0$ par le changement de variable $z = w + \dfrac{b}{3a}$: le lecteur patient calculera p et q en fonction de a, b, c, d. Si $p = 0$, on est ramené à calculer les racines cubiques de $-q$, ce que permet la proposition 20 de la page précédente. Nous supposerons donc $p \neq 0$. Quel que soit z, on peut choisir des complexes u et v tels que $u + v = z$ et $uv = -p/3$ (somme et produit : cela revient à résoudre une équation du second degré, ce qui est possible dans $\mathbb{C}$). On ne perd donc pas en généralité en cherchant nos solutions z sous la forme $z = u + v$, avec la condition $uv = -p/3$. Les inconnues auxiliaires u et v jouent d'ailleurs un rôle symétrique. L'intérêt de cette forme réside dans l'égalité $z^3 + pz + q = u^3 + v^3 + (3uv + p)(u + v) + q$, qui nous conduit au système : $uv = -p/3$, $u^3 + v^3 = -q$.
On pose $U = u^3, V = v^3$, d'où $U + V = -q$, $UV = -p^3/27$. Soit $\Delta = q^2 + 4p^3/27$ et soient $\pm\delta$ ses racines carrées. On prendra $U, V = \dfrac{-q \pm \delta}{2}$, l'ordre étant indifférent. Soient u_1, u_2, u_3 les racines cubiques de U et v_1, v_2, v_3 celles de V. Comme $q \neq 0$, U et V sont non nuls et chacun a trois racines distinctes. Pour chacune des racines $u = u_i$ de U, les complexes uv_1, uv_2 et uv_3 sont distincts et leur cube est $UV = -p^3/27$. L'un d'entre eux est égal à $-p/3$. Quitte à renuméroter les v_i, on peut donc supposer que $u_1v_1 = u_2v_2 = u_3v_3 = -p/3$. Les racines de notre équation sont alors $z_1 = u_1 + v_1$, $z_2 = u_2 + v_2$ et $z_3 = u_3 + v_3$. Cette discussion est résumée par la *formule de Cardan* (ou *de Scipion del Ferro*) :

$$z = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} .$$

Si, par exemple, on applique cette formule à l'équation $z^3 - 15z - 4 = 0$, on obtient :

$$z = \sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}.$$

Il faut donc, pour commencer, trouver une racine cubique de $2 + 11\mathrm{i}$. Avec un peu de chance, on trouve $u = 2 + \mathrm{i}$. Les racines cubiques de $2 + 11\mathrm{i}$ (resp. de $2 - 11\mathrm{i}$) sont donc u, uj, uj^2 (resp. v, vj, vj^2, où $v = \overline{u} = 2 - \mathrm{i}$). On constate que u et v s'apparient : $u + v = 4$ est une racine. Elle est réelle, bien que l'on ait dû passer par les complexes pour la calculer. On termine en factorisant $z^3 - 15z - 4 = (z-4)(z^2 + az + b)$. On trouve $a = 4$ et $b = 1$, d'où les racines $-2 \pm \sqrt{3}$.

3.5 Applications à la trigonométrie

Le conjugué de $\mathrm{e}^{\mathrm{i}x} = \cos x + \mathrm{i}\sin x$ est $\mathrm{e}^{-\mathrm{i}x} = \cos x - \mathrm{i}\sin x$.
On en tire les *formules d'Euler* :

$$\cos x = \frac{\mathrm{e}^{\mathrm{i}x} + \mathrm{e}^{-\mathrm{i}x}}{2}, \qquad \sin x = \frac{\mathrm{e}^{\mathrm{i}x} - \mathrm{e}^{-\mathrm{i}x}}{2\mathrm{i}} .$$

Ces formules peuvent servir à *linéariser* les expressions polynomiales en $\cos x$ et $\sin x$.

Exemples.

1. On n'est pas obligé d'apprendre par cœur les formules donnant les produits. Par exemple, pour calculer $\cos a \cos b$, on écrit :

$$\frac{\mathrm{e}^{\mathrm{i}a}+\mathrm{e}^{-\mathrm{i}a}}{2}\times\frac{\mathrm{e}^{\mathrm{i}b}+\mathrm{e}^{-\mathrm{i}b}}{2}=\frac{\mathrm{e}^{\mathrm{i}(a+b)}+\mathrm{e}^{\mathrm{i}(a-b)}+\mathrm{e}^{\mathrm{i}(-a+b)}+\mathrm{e}^{\mathrm{i}(-a-b)}}{4},$$

d'où l'on tire : $\cos a \cos b = \dfrac{\cos(a+b)+\cos(a-b)}{2}$.

De même, le calcul de $\cos a \sin b$:

$$\frac{\mathrm{e}^{\mathrm{i}a}+\mathrm{e}^{-\mathrm{i}a}}{2}\times\frac{\mathrm{e}^{\mathrm{i}b}-\mathrm{e}^{-\mathrm{i}b}}{2\mathrm{i}}=\frac{\mathrm{e}^{\mathrm{i}(a+b)}-\mathrm{e}^{\mathrm{i}(a-b)}+\mathrm{e}^{\mathrm{i}(-a+b)}-\mathrm{e}^{\mathrm{i}(-a-b)}}{4\mathrm{i}},$$

d'où l'on tire : $\cos a \sin b = \dfrac{\sin(a+b)-\sin(a-b)}{2}$.

De même, le calcul de $\sin a \sin b$:

$$\frac{\mathrm{e}^{\mathrm{i}a}-\mathrm{e}^{-\mathrm{i}a}}{2\mathrm{i}}\times\frac{\mathrm{e}^{\mathrm{i}b}-\mathrm{e}^{-\mathrm{i}b}}{2\mathrm{i}}=\frac{\mathrm{e}^{\mathrm{i}(a+b)}-\mathrm{e}^{\mathrm{i}(a-b)}-\mathrm{e}^{\mathrm{i}(-a+b)}+\mathrm{e}^{\mathrm{i}(-a-b)}}{-4},$$

d'où l'on tire : $\sin a \sin b = \dfrac{\cos(a-b)-\cos(a+b)}{2}$.

2. On obtient de la même manière des expressions pour les puissances du cosinus et du sinus :

$$\begin{aligned}\cos^n x=\left(\frac{\mathrm{e}^{\mathrm{i}x}+\mathrm{e}^{-\mathrm{i}x}}{2}\right)^n &= 2^{-n}\sum_{k=0}^{n}\binom{n}{k}\left(\mathrm{e}^{\mathrm{i}x}\right)^k\left(\mathrm{e}^{-\mathrm{i}x}\right)^{n-k}\\ &= 2^{-n}\sum_{k=0}^{n}\binom{n}{k}\mathrm{e}^{\mathrm{i}(2k-n)x}.\end{aligned}$$

Plutôt que développer $\mathrm{e}^{\mathrm{i}(2k-n)x}$ en cosinus et sinus, le plus simple est de remarquer que $\cos^n x = \mathrm{Re}(\cos^n x)$, d'où :

$$\cos^n x = 2^{-n}\sum_{k=0}^{n}\binom{n}{k}\mathrm{Re}\left(\mathrm{e}^{\mathrm{i}(2k-n)x}\right)=2^{-n}\sum_{k=0}^{n}\binom{n}{k}\cos(2k-n)x.$$

On peut améliorer cette formule en tenant compte de la parité de la fonction cosinus, mais il y a discussion selon la parité de n.

Si $n=2m$, par le changement d'indice $\ell=m-k$, on obtient :

$$\cos^{2m} x = 2^{-2m}\left(\binom{2m}{m}+2\sum_{\ell=1}^{m}\binom{2m}{\ell+m}\cos 2\ell x\right).$$

Pour $n=2m+1$, le même changement d'indice donne :

$$\cos^{2m+1} x = 2^{-2m}\sum_{\ell=0}^{m}\binom{2m+1}{\ell+m+1}\cos(2\ell+1)x.$$

De l'égalité $\left(\mathrm{e}^{\mathrm{i}x}\right)^n = \mathrm{e}^{\mathrm{i}nx}$, on tire la *formule de Moivre* :

$$\cos nx + \mathrm{i}\sin nx = (\cos x + \mathrm{i}\sin x)^n .$$

Exemple. Écrivons que $\cos nx$ et $\sin nx$ sont respectivement la partie réelle et la partie imaginaire de $(\cos x + \mathrm{i}\sin x)^n$:

$$\cos nx = \sum_{k=0}^{n} \binom{n}{k} \operatorname{Re}\left((\cos x)^{n-k}(\mathrm{i}\sin x)^k\right) = \sum_{\ell=0}^{\lfloor n/2 \rfloor} \binom{n}{2\ell} (\cos x)^{n-2\ell}(\mathrm{i}\sin x)^{2\ell},$$

$$\sin nx = \sum_{k=0}^{n} \binom{n}{k} \operatorname{Im}\left((\cos x)^{n-k}(\mathrm{i}\sin x)^k\right) = \frac{1}{\mathrm{i}} \sum_{\ell=0}^{\lfloor (n-1)/2 \rfloor} \binom{n}{2\ell+1} (\cos x)^{n-2\ell-1}(\mathrm{i}\sin x)^{2\ell+1},$$

car les puissances paires de i sont réelles et ses puissances impaires sont imaginaires pures ; on a noté $\lfloor r \rfloor$ la partie entière de r. Le calcul se poursuit en remplaçant $\mathrm{i}^{2\ell}$ par $(-1)^\ell$ et $\sin^2 x$ par $1 - \cos^2 x$:

$$\cos nx = \sum_{\ell=0}^{\lfloor n/2 \rfloor} \binom{n}{2\ell} (-1)^\ell (\cos x)^{n-2\ell}(1 - \cos^2 x)^\ell,$$

$$\sin nx = \sin x \sum_{\ell=0}^{\lfloor (n-1)/2 \rfloor} \binom{n}{2\ell+1} (-1)^\ell (\cos x)^{n-2\ell-1}(1 - \cos^2 x)^\ell.$$

On constate que $\cos nx$ et $\dfrac{\sin nx}{\sin x}$ sont des polynômes en $\cos x$: ce sont les *polynômes de Tchebychef*, que nous retrouverons par une autre voie au module II.6 (section 2.3, page 310).

4 Quelques applications géométriques

4.1 Similitudes planes

Nous avons vu que l'application $z \mapsto \mathrm{i}z$ s'interprète géométriquement comme une rotation de centre 0 et d'angle $\pi/2$. Cela permet de caractériser les bases orthonormées directes du plan : ce sont les couples (u, v) de complexes de module 1 tels que $v = \mathrm{i}u$. Un tel couple est l'image de la base canonique $(1, \mathrm{i})$ par l'application $R : z \mapsto uz$. Posons $u = \mathrm{e}^{\mathrm{i}\theta}$. L'application R transforme le point (x, y) en le point $u(x + \mathrm{i}y)$ de coordonnées $(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$. C'est la *rotation* $R(O, \theta)$ *de centre* O *et d'angle* θ. Elle ne dépend que de la classe $\theta + 2\pi\mathbb{Z}$.

Exemple. Soit $\omega \in \mu_n$ une racine de l'unité. L'application $z \mapsto \omega z$ est la rotation $R(O, \arg\omega)$. L'ensemble des rotations $R(O, \arg\omega)$ lorsque ω parcourt μ_n est le groupe des rotations d'angles $2k\pi/n$, $k \in \mathbb{Z}$. L'application $\omega \mapsto R(O, \arg\omega)$ est un isomorphisme de μ_n sur ce groupe, qui est donc cyclique. Supposons $n \geqslant 3$. Pour tout $\omega \in \mu_n$, l'application $z \mapsto \omega z$ permute les éléments de μ_n. Le polygone dont les sommets sont les éléments de μ_n parcourus dans l'ordre « naturel » est transformé en lui même par $R(O, \arg\omega)$. Il est donc invariant par le groupe des rotations d'angles $2k\pi/n$: c'est en ce sens qu'on le considère comme *régulier*.

Rappelons que la similitude directe $S(A,\rho,\theta)$ de centre A, de rapport $\rho \neq 0$ et d'angle θ est l'application obtenue en composant l'homothétie $H(A,\rho)$ de centre A et de rapport ρ avec la rotation $R(A,\theta)$ de centre A et d'angle θ, l'ordre de composition étant indifférent :

$$S(A,\rho,\theta) = H(A,\rho) \circ R(A,\theta) = R(A,\theta) \circ H(A,\rho).$$

La transformation identité est une similitude $S(A,1,0)$ de rapport 1, d'angle $0 \pmod{2\pi}$ et de centre arbitraire (on peut aussi lui attribuer le rapport -1 et l'angle π). Pour tout autre similitude, le centre est l'unique point fixe. Avec ces notations, on a l'égalité : $S(A,\rho,\theta) = S(A,-\rho,\theta+\pi)$; on pourra donc, sans restreindre la généralité, supposer que $\rho > 0$. D'autre part, pour tout $k \in \mathbb{Z}$, on a : $S(A,\rho,\theta) = S(A,\rho,\theta+2k\pi)$. Ce sont donc le rapport (supposé positif) et l'angle (connu à $2\pi\mathbb{Z}$ près) qui sont, avec le centre, caractéristiques de la similitude. Composant rotations et homothéties, on voit que la similitude directe de centre O, de rapport $\rho > 0$ et d'angle θ est l'application $z \mapsto az$, où $a = \rho\,\mathrm{e}^{\mathrm{i}\theta}$. Les paramètres qui définissent cette similitude sont donc le module et l'argument de a.

Les similitudes directes autres que l'identité sont les applications de la forme :

$$z \mapsto z' = az + b \qquad \text{où} \qquad a, b \in \mathbb{C} \quad \text{et} \quad a \neq 0, 1.$$

Pour le voir, on exprime que le centre est z_0, en écrivant que $z' - z_0 = a(z - z_0)$, soit $z' = az + (1-a)z_0$. Réciproquement, si $a \neq 1$, la relation $z' = az + b$ équivaut à $z' - z_0 = a(z - z_0)$, où $z_0 = \dfrac{b}{1-a}$ ·

Exemple. Nous allons décrire le *groupe G engendré par les similitudes directes*. On calcule d'abord la composée des deux similitudes $z' = az+b$ et $z'' = a'z'+b'$: c'est $z'' = a''z+b''$, avec $a'' = aa'$ et $b'' = b' + ba'$. Si $aa' = 1$ (similitudes directes de rapports inverses et angles opposés), la transformation $z \mapsto z''$ est la translation de vecteur $b' + ba'$; si de plus $b' + ba' = 0$, c'est même à la fois une translation et une similitude directe. Si $aa' \neq 1$, on obtient une similitude directe de rapport le produit des rapports et d'angle la somme des angles. Son centre n'a pas de caractérisation simple, sauf dans le cas où les deux similitudes de départ ont même centre : c'est alors aussi celui de leur composée. Le groupe G contient donc les similitudes directes et les translations, donc toutes les transformations $\sigma_{a,b} : z \mapsto az + b$, $a \in \mathbb{C}^*$, $b \in \mathbb{C}$. D'après le calcul ci-dessus, l'ensemble de ces transformations est stable pour la composition, et il contient l'identité (c'est $\sigma_{1,0}$). De plus, la réciproque de la transformation $\sigma_{a,b}$ est la transformation $\sigma_{a',b'}$, où $a' = a^{-1}$ et $b' = -a^{-1}b$. Finalement : $G = \{\sigma_{a,b} \mid (a,b) \in \mathbb{C}^* \times \mathbb{C}\}$.

4.2 Angles de vecteurs et angles de droites

Soit $z \mapsto z' = az + b$ une similitude directe. L'image du vecteur $\overrightarrow{U} = \overrightarrow{AB}$ est le vecteur $\overrightarrow{U'} = \overrightarrow{A'B'}$, avec $\mathrm{Aff}(A') = a\ \mathrm{Aff}(A) + b$, $\mathrm{Aff}(B') = a\ \mathrm{Aff}(B) + b$, de sorte que $\mathrm{Aff}(\overrightarrow{U'}) = a\ \mathrm{Aff}(\overrightarrow{U})$. En ce qui concerne les vecteurs (ou les complexes considérés comme des vecteurs), on ne s'intéresse donc qu'aux similitudes directes de la forme $z \mapsto az$. Si l'on se donne réciproquement deux vecteurs non nuls $\overrightarrow{U}$ et $\overrightarrow{U'}$ du plan, il existe une et une seule similitude directe $z \mapsto az$ qui transforme $\overrightarrow{U}$ en $\overrightarrow{U'}$: c'est celle qui vérifie $a = \dfrac{\mathrm{Aff}(\overrightarrow{U'})}{\mathrm{Aff}(\overrightarrow{U})}$ ·

Définition 17. L'*angle orienté* $(\widehat{\overrightarrow{U},\overrightarrow{U'}})$ est l'angle de l'unique similitude directe qui transforme $\overrightarrow{U}$ en $\overrightarrow{U'}$. On a donc l'égalité :

$$(\widehat{\overrightarrow{U},\overrightarrow{U'}}) = \arg\left(\frac{\mathrm{Aff}(\overrightarrow{U'})}{\mathrm{Aff}(\overrightarrow{U})}\right) .$$

L'angle orienté de deux vecteurs est donc un élément de $\mathbb{R}/2\pi\mathbb{Z}$, c'est-à-dire un réel défini modulo 2π. Les règles qui suivent mettent donc en jeu l'addition modulo 2π. Les deux premières sont les *relations de Chasles* pour les angles et se déduisent des règles de calcul sur les arguments.

$$\begin{aligned}
(\widehat{\overrightarrow{U},\overrightarrow{U''}}) &= (\widehat{\overrightarrow{U},\overrightarrow{U'}}) + (\widehat{\overrightarrow{U'},\overrightarrow{U''}}),\\
(\widehat{\overrightarrow{U},\overrightarrow{U'}}) &= -(\widehat{\overrightarrow{U'},\overrightarrow{U}}),\\
(\widehat{\overrightarrow{U},\overrightarrow{U'}}) - (\widehat{\overrightarrow{V},\overrightarrow{V'}}) &= (\widehat{\overrightarrow{U},\overrightarrow{V}}) - (\widehat{\overrightarrow{U'},\overrightarrow{V'}}),\\
\overrightarrow{U},\overrightarrow{V} \text{ positivement liés} &\iff (\widehat{\overrightarrow{U},\overrightarrow{V}}) = 0 \pmod{2\pi},\\
\overrightarrow{U},\overrightarrow{V} \text{ liés} &\iff (\widehat{\overrightarrow{U},\overrightarrow{V}}) = 0 \text{ ou } \pi \pmod{2\pi}.
\end{aligned}$$

La troisième règle se démontre à l'aide des relations de Chasles et d'un calcul dans le groupe $\mathbb{R}/2\pi\mathbb{Z}$: la preuve en est laissée en exercice. Dans les deux dernières règles, le complexe $\dfrac{\mathrm{Aff}(\overrightarrow{V})}{\mathrm{Aff}(\overrightarrow{U})}$ est respectivement réel strictement positif (argument nul) et réel non nul (argument nul ou égal à π).

Supposons les vecteurs $\overrightarrow{U}$, $\overrightarrow{V}$ positivement liés, de même que les vecteurs $\overrightarrow{U'}$, $\overrightarrow{V'}$. D'après les règles ci-dessus, $(\widehat{\overrightarrow{U},\overrightarrow{U'}}) = (\widehat{\overrightarrow{V},\overrightarrow{V'}})$. Comme tous les vecteurs directeurs d'une droite orientée (c'est-à-dire munie d'un sens positif) sont positivement liés entre eux, la définition qui suit a bien un sens. Elle garderait d'ailleurs un sens pour des demi-droites.

Définition 18. Soient $\overrightarrow{D}$ et $\overrightarrow{D'}$ deux droites orientées. L'*angle orienté* $(\widehat{\overrightarrow{D},\overrightarrow{D'}})$ est l'angle orienté $(\widehat{\overrightarrow{U},\overrightarrow{U'}})$ d'un vecteur directeur $\overrightarrow{U}$ quelconque de $\overrightarrow{D}$ et d'un vecteur directeur $\overrightarrow{U'}$ quelconque de $\overrightarrow{D'}$.

Si l'on suppose seulement les vecteurs $\overrightarrow{U}$, $\overrightarrow{V}$ liés, de même que les vecteurs $\overrightarrow{U'}$, $\overrightarrow{V'}$, on a $(\widehat{\overrightarrow{U},\overrightarrow{U'}}) = (\widehat{\overrightarrow{V},\overrightarrow{V'}})$ ou $(\widehat{\overrightarrow{V},\overrightarrow{V'}}) + \pi$. L'angle orienté de deux droites ne peut être défini que modulo π.

Exemple. Fixons deux points A et B distincts. On va déterminer l'ensemble Γ des points M (distincts de A, B) tels que l'angle des droites (MA) et (MB) ait une valeur donnée. D'après la remarque ci-dessus, cette valeur n'est définie que modulo π. Notant a, b, z les affixes de A, B, M, on a la condition : $z \in \Gamma \Leftrightarrow \mathrm{Arg}\,\dfrac{z-b}{z-a} \equiv \alpha \pmod{\pi}$.

Cette condition équivaut à $\mathrm{e}^{-\mathrm{i}\alpha}\dfrac{z-b}{z-a} \in \mathbb{R}^*$. Comme on sait que $\dfrac{z-b}{z-a} \in \mathbb{C}^*$, il suffit d'écrire que ce nombre est égal à son conjugué : $\mathrm{e}^{-\mathrm{i}\alpha}\dfrac{z-b}{z-a} = \mathrm{e}^{\mathrm{i}\alpha}\dfrac{\overline{z}-\overline{b}}{\overline{z}-\overline{a}}$, ce qui équivaut à la relation : $(\mathrm{e}^{2\mathrm{i}\alpha}-1)z\overline{z} - -(\mathrm{e}^{2\mathrm{i}\alpha}\overline{b}-\overline{a})z - (\mathrm{e}^{2\mathrm{i}\alpha}a-b)\overline{z}(\mathrm{e}^{2\mathrm{i}\alpha}a\overline{b}-b\overline{a}) = 0$. Le lecteur vérifiera que, si $\alpha \equiv 0 \pmod{\pi}$, c'est l'équation d'une droite ; sinon, c'est l'équation d'un cercle (auxquels il faut toutefois ôter A et B).

4.3 Constructions à la règle et au compas

Une élégante théorie *algébrique* permet de dire quelles constructions géométriques planes peuvent être effectuées à l'aide d'une règle et d'un compas. On démontre ainsi que trois problèmes antiques n'admettent pas de solution : la duplication du cube (construire le côté d'un cube de volume 2 : autrement dit, construire le réel $\sqrt[3]{2}$) ; la trisection de l'angle (construire les trisectrices d'un angle, ou encore le diviser en 3) ; la quadrature du cercle (construire le côté d'un carré ayant même aire qu'un disque de rayon 1 : autrement dit, construire le réel $\sqrt{\pi}$). À l'inverse, Gauß a montré à l'âge de 19 ans comment construire un polygone régulier à 17 côtés. Nous effleurons ici le sujet. Nous admettons les constructions élémentaires suivantes, qui relèvent de la géométrie de terminale :

1) Possibilité d'additionner, de soustraire, de multiplier et de diviser des longueurs ; ces deux dernières reposent sur le théorème de Thalès et sur la construction de la parallèle à une droite donnée passant par un point donné.

2) Possibilité d'additionner et de soustraire des angles.

3) Possibilité d'extraire la racine carrée d'une longueur (qui repose sur la géométrie des triangles rectangles) et de construire la moitié d'un angle (ou ses bissectrices).

Nous dirons qu'un complexe z est connu si le point M d'affixe z est donné et que l'on peut s'en servir pour de nouvelles constructions. Bien sûr, si z est connu, les réels $\operatorname{Re} z$, $\operatorname{Im} z$ et $|z|$ ainsi que l'angle $\arg z$ le sont. Réciproquement, si les réels $\operatorname{Re} z$ et $\operatorname{Im} z$ sont connus, ou bien si la longueur $|z|$ et l'angle $\arg z$ le sont, alors z l'est. Les complexes 0, 1 et i sont évidemment supposés connus dès le départ. On voit alors que les constructions suivantes sont possibles :

1) Les complexes a et b étant connus, construire $a+b$ et $a-b$. Il suffit en effet d'additionner ou de soustraire les parties réelles et imaginaires.

2) Les complexes a et b étant connus, construire le produit ab et, si $a \neq 0$, le quotient b/a. Il suffit en effet de multiplier ou de diviser les longueurs $|a|$ et $|b|$ et d'additionner ou de soustraire les angles $\arg a$ et $\arg b$.

3) Le complexe a étant connu, construire ses racines carrées. Il suffit en effet de construire la racine carrée de la longueur $|a|$ et la moitié de l'angle $\arg a$.

On déduit de ces préliminaires que, les complexes a, b, c étant supposés connus, *on peut construire les racines de l'équation* az^2+bz+c. Si $a=b=0$, ces racines sont tous les complexes (si $c=0$) ou aucun (si $c \neq 0$) et c'est évident. Si $a=0$ et $b \neq 0$, l'unique racine est $-b/a$, que l'on sait construire. Si $a \neq 0$, on construit successivement $\Delta = b^2-4ac$, ses racines carrées $\pm\delta$ et les racines de l'équation $\dfrac{-b\pm\delta}{2a}$. Pour construire un pentagone régulier, il suffit de construire $\mathrm{e}^{2\imath\pi/5}$. C'est une racine de l'équation $z^2 - zZ + 1 = 0$, où Z est une racine de l'équation $Z^2+Z-1=0$. Pour un algorithme pratique, consulter les vieux manuels de géométrie plane. En revanche, on ne peut pas résoudre en général les équations de degré 3 par à la règle et au compas : sinon, on pourrait construire $\sqrt[3]{2}$, ou encore $z = \cos\theta$ à partir de $a = \cos 3\theta$ (et donc trisecter les angles) ; voyez-vous pourquoi ?

5 Topologie de $\mathbb{C}$

5.1 Rappels sur la convergence dans $\mathbb{C}$

Rappelons (module IV.1, section 2.2.2) qu'une suite $(z_n)_{n\in\mathbb{N}}$ de nombres complexes converge vers le complexe ℓ si $\lim\limits_{n\to+\infty} |z_n - \ell| = 0$. On écrit alors $\lim\limits_{n\to+\infty} z_n = \ell$. Rappelons également (module IV.4) qu'une série $\sum\limits_{n\geqslant 0} z_n$ de nombres complexes converge, si, et seulement si, les séries $\sum\limits_{n\geqslant 0} \operatorname{Re} z_n$ et $\sum\limits_{n\geqslant 0} \operatorname{Im} z_n$ convergent. On a les égalités :

$$\lim_{n\to+\infty}(x_n + \mathrm{i}y_n) = \lim_{n\to+\infty} x_n + \mathrm{i}\lim_{n\to+\infty} y_n \qquad \text{et} \qquad \sum_{n\geqslant 0}(x_n + \mathrm{i}y_n) = \sum_{n\geqslant 0} x_n + \mathrm{i}\sum_{n\geqslant 0} y_n,$$

ainsi que leurs conséquences triviales : $\overline{\lim\limits_{n\to+\infty} z_n} = \lim\limits_{n\to+\infty} \overline{z_n}$ et $\overline{\sum\limits_{n\geqslant 0} z_n} = \sum\limits_{n\geqslant 0} \overline{z_n}$.

On en déduit les règles concernant l'addition :

$$\lim_{n\to+\infty}(z_n + z'_n) = \lim_{n\to+\infty} z_n + \lim_{n\to+\infty} z'_n \qquad \text{et} \qquad \sum_{n\geqslant 0}(z_n + z'_n) = \sum_{n\geqslant 0} z_n + \sum_{n\geqslant 0} z'_n$$

ainsi que celle concernant la limite d'un produit :

$$\lim_{n\to+\infty}(z_n z'_n) = \big(\lim_{n\to+\infty} z_n\big)\big(\lim_{n\to+\infty} z'_n\big).$$

En particulier, pour $w \in \mathbb{C}$ quelconque :

$$\lim_{n\to+\infty} w\, z_n = w \lim_{n\to+\infty} z_n \quad \text{et} \quad \sum_{n\geqslant 0} w\, z_n = w \sum_{n\geqslant 0} z_n.$$

Notons $z_n = x_n + \mathrm{i}y_n$ La série de complexes $\sum\limits_{n\geqslant 0} z_n$ est dite *absolument convergente* si la série à termes positifs $\sum\limits_{n\geqslant 0} |z_n|$ converge, ce qui a lieu si, et seulement si, les séries de réels $\sum\limits_{n\geqslant 0} x_n$ et $\sum\limits_{n\geqslant 0} y_n$ sont absolument convergentes. Toute série absolument convergente de complexes est convergente (module IV.4, théorème 20 de la page 730). La somme de deux séries absolument convergentes est absolument convergente. Pour toute série absolument convergente $\sum\limits_{n\geqslant 0} z_n$, on a : $\big|\sum\limits_{n\geqslant 0} z_n\big| \leqslant \sum\limits_{n\geqslant 0} |z_n|$.

Le *produit de Cauchy* des séries absolument convergentes $\sum\limits_{n\geqslant 0} z_n$ et $\sum\limits_{n\geqslant 0} z'_n$ est la série $\sum\limits_{n\geqslant 0} w_n$ de terme général $w_n = \sum\limits_{j+k=n} z_j z'_k$.

Cette série converge absolument et l'on a l'égalité : $\big(\sum\limits_{n\geqslant 0} z_n\big)\big(\sum\limits_{n\geqslant 0} z'_n\big) = \sum\limits_{n\geqslant 0} w_n$ (module IV.4, théorème et définition 21 de la page 731).

5.2 L'exponentielle complexe

Théorème et définition 21. Pour tout complexe z, la série $\sum_{n\geqslant 0} \frac{z^n}{n!}$ converge absolument. Sa somme est notée $\exp z$ ou e^z et la fonction $\exp : z \mapsto \exp z$ de $\mathbb{C}$ dans $\mathbb{C}$ est appelée fonction *exponentielle (complexe)*. Elle est à valeurs dans $\mathbb{C}^*$ et vérifie les relations :

$$\begin{gathered} \exp(0) = 1, \\ \forall a, b \in \mathbb{C}\ ,\ \exp(a+b) = (\exp a)\,(\exp b)\,, \\ \forall a \in \mathbb{C}\ ,\ \exp(-a) = (\exp a)^{-1}\,, \\ \forall a \in \mathbb{C}\ ,\ \overline{\exp a} = \exp \overline{a}. \end{gathered} \tag{7}$$

En particulier, la fonction $\exp$ envoie $\mathbb{R}$ dans $\mathbb{R}_+^*$ et $\mathrm{i}\,\mathbb{R}$ dans $\mathbb{U}$.

Démonstration. On peut supposer $z \neq 0$. Soit n_0 un entier tel que $m = \frac{|z|}{n_0} < 1$. Notant $u_n = \left|\frac{z^n}{n!}\right|$, on voit que, pour $n \geqslant n_0$, $\frac{u_{n+1}}{u_n} < m$. La série $\sum_{n\geqslant 0} u_n$ est donc convergente et la série $\sum_{n\geqslant 0} \frac{z^n}{n!}$ absolument convergente. La première égalité à démontrer est évidente. La seconde résulte du calcul du terme général du produit de Cauchy des séries $\sum_{n\geqslant 0} \frac{a^n}{n!}$ et $\sum_{n\geqslant 0} \frac{b^n}{n!}$: $\sum_{j+k=n} \frac{a^j}{j!}\frac{b^k}{k!} = \frac{(a+b)^n}{n!}$, qui n'est autre que la formule du binôme. La troisième égalité résulte des deux premières en posant $b = -a$. Elle implique que $\exp$ est à valeurs dans $\mathbb{C}^*$. La dernière égalité résulte des règles générales de 5.1. Elle entraîne que, pour $t \in \mathbb{R}$, $\exp t \in \mathbb{R}^*$, puis que $\exp t = (\exp t/2)^2 > 0$; et que l'inverse $\exp(-\mathrm{i}t)$ de $\exp(\mathrm{i}t)$ est aussi son conjugué, donc que $\exp(\mathrm{i}t) \in \mathbb{U}$. ∎

La preuve (facile) de la majoration suivante, est laissée en exercice au lecteur :

$$\forall z \in \mathbb{C}^*\ ,\ 0 < |z| \leqslant M \Longrightarrow \left|\frac{\exp z - 1 - z}{z^2}\right| \leqslant \exp M. \tag{8}$$

Avec les notations du module IV.8, on en déduit que $\exp z = 1 + z + \mathrm{O}(z^2)$ lorsque $z \to 0$.

Théorème 22.
La restriction de la fonction $\exp$ à $\mathbb{R}$ est la fonction exponentielle réelle.

Notons $f(x) = \exp x$ pour $x \in \mathbb{R}$.
On a vu que f est à valeurs réelles et que $f(0) = 1$. Des propriétés (7) et (8) on déduit que, pour $a \in \mathbb{R}$ fixé, $f(a+h) = f(a) + hf(a) + \mathrm{O}(h^2)$ lorsque $h \to 0$. Ainsi f est dérivable sur $\mathbb{R}$ et $f' = f$. L'exponentielle réelle est l'unique fonction dérivable sur $\mathbb{R}$ égale à sa dérivée et prenant la valeur 1 en 0 : *cf.* le corollaire 6 de la page 681.

5.2.1 La fonction $\exp$ et le cercle unité $\mathbb{U}$

Notons, pour tout $y \in \mathbb{R}$, $C(y) = \mathrm{Re}(\exp \mathrm{i}y)$ et $S(y) = \mathrm{Im}(\exp \mathrm{i}y)$. Les fonctions C et S sont définies sur $\mathbb{R}$ et à valeurs dans $\mathbb{R}$, et l'on a, d'après l'équation 7 et le théorème 22 de la page précédente, l'égalité : $\forall x, y \in \mathbb{R}$, $\exp(x+\mathrm{i}y) = \mathrm{e}^x\big((C(y)+\mathrm{i}S(y)\big)$. D'après le théorème 21 de la page ci-contre, $\mathrm{e}^x > 0$ et $C(y) + \mathrm{i}S(y) \in \mathbb{U}$, c'est à dire $C^2 + S^2 = 1$. Par exemple, $C(0) = 1$ et $S(0) = 0$. Du même théorème, on déduit que C est paire et S impaire. Les théorèmes qui suivent font alors le lien avec la présentation « élémentaire » de l'exponentielle au début de la section 3.

Théorème et définition 23.
Les fonctions C et S sont dérivables et $C' = -S$, $S' = C$. Il existe un unique réel $a > 0$ tel que $C(a) = 0, S(a) = 1$ avec $C > 0$ et $S > 0$ sur $]0, a[$. On notera π le réel $2a$. Les fonctions C et S sont 2π-périodiques.
L'application $t \mapsto (C(t), S(t))$ est une bijection de $[0, 2\pi[$ sur le cercle trigonométrique $\mathbb{U} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\} \subset \mathbb{R}^2$; de manière équivalente, l'application $t \mapsto \exp \mathrm{i}t$ est une bijection de $[0, 2\pi[$ sur le cercle unité $\mathbb{U} \subset \mathbb{C}$.

Démonstration. Des propriétés (7) et (8) ci-dessus, on déduit que, pour $a \in \mathbb{R}$ fixé, $C(a+h) = C(a) - hS(a) + \mathrm{O}(h^2)$ et $S(a+h) = S(a) + hC(a) + \mathrm{O}(h^2)$ lorsque $h \to 0$, d'où $C' = -S$, $S' = C$. Les fonctions C et S sont donc indéfiniment dérivables. Comme $S(0) = 0$ et $C(0) = S'(0) = 1$, il existe $\alpha > 0$ tel que $C, S > 0$ sur $]0, \alpha]$. Supposons que C reste > 0 sur $\mathbb{R}_+^*$. Alors S croît sur $\mathbb{R}_+^*$, donc $S \geqslant S(\alpha)$ sur $[\alpha, +\infty[$, donc $C(x) \leqslant C(\alpha) - (x - \alpha)S(\alpha)$ sur $[\alpha, +\infty[$ (inégalité des accroissements finis), ce qui contredit l'hypothèse $C > 0$. Il existe donc un plus petit réel $a > 0$ tel que $C(a) = 0$. Alors $S(a) = \pm 1$. Sur $]0, a[$, $C > 0$, donc S croît strictement, donc $S > 0$ et $S(a) = 1$. C'est le a de l'énoncé du théorème et son unicité est évidente ; nous l'écrirons $a = \pi/2$. On a donc $\exp \mathrm{i}\pi/2 = \mathrm{i}$, d'où, par application répétée de (7), $\exp \mathrm{i}(a + 2\pi) = \exp \mathrm{i}a$: cela entraîne la 2π-périodicité de C et de S. L'étude précédente (avec le théorème des valeurs intermédiaires) montre que l'application $t \mapsto \big(C(t), S(t)\big)$ est une bijection de $[0, \pi/2]$ sur $\mathbb{U} \cap (\mathbb{R}_+ \times \mathbb{R}_+)$. La relation $\exp \mathrm{i}(t + \pi/2) = \mathrm{i} \exp \mathrm{i}t$ entraîne alors que l'on a des bijections de $[\pi/2, \pi[$, $[\pi, 3\pi/2[$ et $[3\pi/2, 2\pi[$ sur les intersections de $\mathbb{U}$ avec les trois autres quadrants. ∎

Proposition 24. La fonction $t \mapsto (C(t), S(t))$ est un paramétrage du cercle trigonométrique par l'abscisse curviligne (voir le module III.2).

Démonstration. Cela résulte de l'égalité $C'^2 + S'^2 = (-S)^2 + C^2 = 1$. ∎

La « condition initiale » $\big(C'(0), S'(0)\big) = \big(-S(0), C(0)\big) = (0, 1)$, garantit que le sens du paramétrage est le sens trigonométrique habituel.
Autrement dit, pour $t \in [0, 2\pi[$, $(C(t), S(t))$ sont les coordonnées du point M tel que l'arc de cercle (parcouru dans le sens positif) qui joint $(1, 0)$ à M a pour longueur t : au sens du programme de terminale, on a bien $C(t) = \cos t$ et $S(t) = \sin t$.

5.3 Le théorème de d'Alembert-Gauß

Nous allons démontrer que le corps $\mathbb{C}$ est algébriquement clos. Cette propriété purement algébrique de $\mathbb{C}$ admet de nombreuses preuves substantiellement différentes, qui reposent *toutes* sur la topologie ou sur l'analyse. Celle que nous proposons ici est un cas particulier du *principe du maximum* pour les fonctions de variables complexes, qui seront étudiées en L2.

Soit $P = \sum_{n \geqslant 0} a_n X^n \in \mathbb{C}[X]$ un polynôme à coefficients complexes et soit $(z_n)_{n\in\mathbb{N}}$ une suite de complexes de limite $\lim_{n\to+\infty} z_n = z$.

Il découle des rappels de 5.1 que $\lim_{n\to+\infty} P(z_n) = P(z)$. On dit que *l'application polynomiale P est continue*.

Proposition 25. Supposons $P = a_0 + \cdots + a_n X^n$ de degré $n \geqslant 1$ (en particulier $a_n \neq 0$). Alors :

$$\forall M > 0\ ,\ \exists R > 0\ ,\ \forall z \in \mathbb{C}\ :\ |z| > R \Longrightarrow |P(z)| > M. \tag{9}$$

On dit que la fonction $|P|$ sur $\mathbb{C}$ *tend vers l'infini à l'infini*.

Démonstration. Supposons $z \neq 0$. Alors :

$$\left|\frac{P(z)}{a_n z^n}\right| = \left|1 + \frac{a_{n-1}}{a_n z} + \cdots + \frac{a_0}{a_n z^n}\right| \geqslant 1 - \left|\frac{a_{n-1}}{a_n z}\right| - \cdots - \left|\frac{a_0}{a_n z^n}\right|.$$

Si de plus, $|z| \geqslant R \geqslant 1$, on en tire l'inégalité :

$$\left|\frac{P(z)}{a_n z^n}\right| \geqslant 1 - \frac{1}{R}\sum_{i=0}^{n-1}\left|\frac{a_i}{a_n}\right|. \tag{10}$$

Il suffit donc de prendre $R \geqslant \max\left(1, 2\sum_{i=0}^{n-1}\left|\frac{a_i}{a_n}\right|\right)$ et tel que $|a_n|\, R^n \geqslant 2M$. ■

Rappelons (module IV.7, définition 5 de la page 822) qu'une partie de $\mathbb{C}$ qui est *fermée* (c'est-à-dire stable par passage à la limite) et *bornée* (c'est-à-dire incluse dans un disque) est dite *compacte*.

Proposition 26. Soit A une partie compacte non vide de $\mathbb{C}$. La fonction $|P|$ admet un maximum et un minimum atteints sur A.

Démonstration. C'est évident si P est constant.

Nous supposerons donc que $P = a_0 + \cdots + a_n X^n$, avec $a_n \neq 0$ et $n \geqslant 1$. Si $M > 0$ est tel que $\forall z \in A, |z| \leqslant M$, alors, sur A : $|P(z)| \leqslant |a_0| + \cdots + |a_n|\, M^n$. L'ensemble non vide $\{|P(z)| \mid z \in A\}$ est donc majoré et minoré (par 0), donc il admet une borne supérieure et une borne inférieure. Nous noterons m cette dernière et prouverons qu'elle est atteinte : le raisonnement est identique en ce qui concerne la borne supérieure. Par définition de la borne inférieure, il existe une suite $(z_n)_{n\in\mathbb{N}}$ d'éléments de A telle que $\lim_{n\to+\infty} |P(z_n)| = m$. La partie A étant bornée, cette suite est bornée. D'après le

théorème de Bolzano-Weierstraß (module IV.6, théorème 11 de la page 784), elle admet une sous-suite convergente ; quitte à remplacer notre suite par cette sous-suite, nous pouvons donc admettre d'emblée que la suite $(z_n)_{n\in\mathbb{N}}$ admet une limite z. La partie A étant fermée, $z \in A$. L'application polynomiale P étant continue, $|P(z)| = \lim\limits_{n\to+\infty} |P(z_n)| = m$. La borne inférieure m est bien atteinte, c'est donc le minimum de $|P|$ sur A. ■

Corollaire 27. La fonction $|P|$ admet un minimum atteint sur $\mathbb{C}$.

Démonstration. Soit $M = |P(0)|$. Si $M = 0$, c'est le minimum ; supposons donc $M > 0$. Comme $|P|$ tend vers l'infini à l'infini, il existe $R > 0$ tel que $|P| > M$ hors du disque fermé $A = \overline{D}(0, R)$. Celui-ci étant fermé borné non vide, $|P|$ y admet un minimum atteint, qui est nécessairement inférieur ou égal à $|P(0)| = M$: c'est donc le minimum sur $\mathbb{C}$. ■

Lemme 28. Soit $k \geqslant 1$ et soit $Q = 1 + X^k +$ des termes de degrés supérieurs un polynôme sur $\mathbb{C}$. Il existe $z_0 \in \mathbb{C}$ tel que $|Q(z_0)| < 1$.

Démonstration. Écrivons $Q = 1 + X^k R$, où R est un polynôme de terme constant 1. Nous prendrons $z_0 = \dfrac{1}{p}\mathrm{e}^{\mathrm{i}\pi/k}$, avec $p \in \mathbb{N}^*$ à déterminer. L'application R étant continue, $\lim\limits_{p\to+\infty} R\left(\dfrac{1}{p}\mathrm{e}^{\mathrm{i}\pi/k}\right) = 1$. On choisit p tel que $R\left(\dfrac{1}{p}\mathrm{e}^{\mathrm{i}\pi/k}\right) = 1 + u$ avec $|u| < 1$. On a alors : $Q(z_0) = 1 + z_0^k R(z_0) = 1 - p^{-k}(1 + u)$, d'où :

$$|Q(z_0)| \leqslant |1 - p^{-k}| + |p^{-k}u| < 1 - p^{-k} + p^{-k} = 1.$$ ■

Théorème 29 (de d'Alembert-Gauß).
Tout polynôme non constant de $\mathbb{C}[X]$ admet au moins une racine.

Démonstration. Soit $z_1 \in \mathbb{C}$ un point où $|P|$ atteint son minimum. On va démontrer par l'absurde que $P(z_1) = 0$. Sinon, on considère le polynôme $\dfrac{P(z_1 + X)}{P(z_1)}$.

Il est non constant (sinon P serait constant) et son terme constant vaut 1 ; il est donc de la forme $1 + aX^k$ plus des termes de degrés supérieurs, avec $k \geqslant 1$ et $a \neq 0$. Soit $\lambda \in \mathbb{C}$ une racine $k^{ème}$ de a^{-1}. On peut alors appliquer le lemme au polynôme $Q = \dfrac{P(z_1 + \lambda X)}{P(z_1)}$: il existe z_0 tel que $|Q(z_0)| < 1$, c'est-à-dire $|P(z_1 + \lambda z_0)| < |P(z_1)|$, contredisant la minimalité de $|P(z_1)|$. ■

Exercice type corrigé

II.5.0 1) Soient A, B, C, D les images dans le plan d'Argand-Cauchy (section 2.2) de $a, b, c, d \in \mathbb{U}$ (définition 12 de la page 258). On suppose $a \neq b$ et $c \neq d$. Donner une condition nécessaire et suffisante simple de parallélisme des droites (AB) et (CD).
2) Lorsque $a = b$, on convient que (AB) désigne la tangente au cercle unité en A (*i.e.* la perpendiculaire au rayon OA passant par A). Généraliser la relation précédente aux cas $a = b$ ou $c = d$.
3) Construire géométriquement ab.
4) Construire géométriquement a^{-1}.
5) Pour tout $u \in \mathbb{U}$, résoudre géométriquement l'équation $x^2 = u$.

Solution. 1) La condition de parallélisme se traduit par la relation $\dfrac{d-c}{b-a} \in \mathbb{R}$, qui équivaut à :

$$\frac{d-c}{b-a} = \frac{\overline{d}-\overline{c}}{\overline{b}-\overline{a}} = \frac{d^{-1}-c^{-1}}{b^{-1}-a^{-1}} = \frac{ab(c-d)}{cd(a-b)},$$

car les éléments $u \in \mathbb{U}$ vérifient $\overline{u} = u^{-1}$.
Un petit calcul donne alors la condition nécessaire et suffisante $ab = cd$.
2) Puisque (AB) est perpendiculaire à OA, on peut remplacer $b-a$ par $\mathrm{i}a$ dans le calcul précédent, ce qui donne $\dfrac{d-c}{\mathrm{i}a} = \dfrac{d^{-1}-c^{-1}}{-\mathrm{i}a^{-1}}$, soit la condition $a^2 = cd$. Si de plus $c = d$, on remplace $d - c$ par $\mathrm{i}c$ et l'on trouve de manière analogue $a^2 = c^2$.
Dans tous les cas, on retrouve la condition $ab = cd$.
3) On prend $d := 1$ et $D := [d]$. On mène la parallèle à (AB) passant par D. Si elle est tangente au cercle, on pose $C := D$; sinon, elle recoupe le cercle en un point C.
Dans tous les cas, le point C a pour affixe $c = ab$.
4) On prend $c := d := 1$, donc pour (CD) la tangente au cercle au point d'affixe 1. La parallèle à (CD) menée par A recoupe le cercle en B (qui est A lui-même si c'est une tangente).
Le point B a pour affixe $b = a^{-1}$.
5) On prend $c := u$ et $d := 1$. On cherche A tel que la tangente en A soit parallèle à (CD), autrement dit, tel que le rayon OA soit perpendiculaire à CD. Il suffit donc de mener la perpendiculaire à (CD) passant par le centre O. Comme c'est un diamètre, elle coupe le cercle en deux points A et B d'affixes a et $b = -a$, qui sont les solutions de $x^2 = u$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

II.5.1 Démontrer qu'il n'existe sur $\mathbb{C}$ aucune relation d'ordre total qui soit compatible avec l'addition et la multiplication. Pour le voir, on discutera du signe de i et du signe de son carré.

II.5.2 Cet exercice propose une construction de $\mathbb{C}$ à l'aide des matrices réelles. Soit $\tilde{\mathbb{C}}$ le sous-ensemble $\left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \middle| a, b \in \mathbb{R} \right\}$ de $M_2(\mathbb{R})$.

(i) Montrer que $\tilde{\mathbb{C}}$ est un sous-anneau de $M_2(\mathbb{R})$. Montrer que l'application $a \mapsto \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$ est un morphisme injectif de $\mathbb{R}$ dans $\tilde{\mathbb{C}}$, qui permet donc d'identifier $\mathbb{R}$ à son image, un sous-anneau de $\tilde{\mathbb{C}}$; ainsi $1 \in \mathbb{R}$ s'identifie à la matrice identité $I_2 \in M_2(\mathbb{R})$.
(ii) Montrer que la matrice $\tilde{\imath} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \tilde{\mathbb{C}}$ vérifie $\tilde{\imath}^2 = -1$ et que tout élément de $\tilde{\mathbb{C}}$ s'écrit de manière unique $a + b\tilde{\imath}$, avec $a, b \in \mathbb{R}$. En déduire que l'application $a + b\mathrm{i} \mapsto a + b\tilde{\imath}$ définit un isomorphisme d'anneaux de $\mathbb{C}$ dans $\tilde{\mathbb{C}}$ dont la restriction à $\mathbb{R}$ est l'identité.

II.5.3 Soit A un anneau commutatif et soit $\alpha \in A$ tel que $\alpha^2 = -1$. Soit $\varphi : \mathbb{R} \to A$ un morphisme d'anneaux. Montrer qu'il existe un unique morphisme d'anneaux $\psi : \mathbb{C} \to A$ dont la restriction à $\mathbb{R}$ est φ et tel que $\psi(\mathrm{i}) = \alpha$.

II.5.4 $*$ En analogie avec la construction de $\mathbb{C}$, on ajoute ici à $\mathbb{R}$ *trois* nouveaux éléments, notés $\mathbf{i}, \mathbf{j}, \mathbf{k}$, et l'on note : $\mathbf{H} := \{a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \mid a, b, c, d \in \mathbb{R}\}$. Les éléments de $\mathbf{H}$ sont appelés *quaternions de Hamilton*, ou simplement quaternions. Nous voulons faire de $\mathbf{H}$ un anneau (ici, non commutatif) admettant $\mathbb{R}$ comme sous-anneau (éléments tels que $b = c = d = 0$). Nous imposons de plus aux nouveaux éléments les règles suivantes :

$$\mathbf{i}^2 = \mathbf{j}^2 = -1, \quad \mathbf{ij} = \mathbf{k} = -\mathbf{ji} \quad \text{et} \quad \forall a \in \mathbb{R}\ ,\ a\mathbf{i} = \mathbf{i}a \quad \text{et} \quad a\mathbf{j} = \mathbf{j}a.$$

(i) Vérifier que $\mathbb{R}$ est central dans $\mathbf{H}$, que $\mathbf{k}^2 = -1$ et que $\mathbf{jk} = \mathbf{i} = -\mathbf{kj}$ et $\mathbf{ki} = \mathbf{j} = -\mathbf{ik}$. Calculer la somme et le produit de $a_1 + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}$ et de $a_2 + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}$. En examinant le cas spécial où $a_2 = a_1$, $b_2 = -b_1$, $c_2 = -c_1$ et $d_2 = -d_1$, vérifier que l'égalité $a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} = 0$ équivaut à $a = b = c = d = 0$, puis que $\mathbf{H}$ est un corps. Vérifier que l'application $a + b\mathrm{i} \mapsto a + b\mathbf{i}$ définit un morphisme injectif d'anneaux de $\mathbb{C}$ dans $\mathbf{H}$ dont la restriction à $\mathbb{R}$ est l'identité.
(ii) Proposer une construction rigoureuse de $\mathbf{H}$.
(iii) On définit une application σ de $\mathbf{H}$ dans lui-même en posant :

$$\sigma(a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}) = a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}.$$

Vérifier que c'est un automorphisme pour la structure de groupe (et même de $\mathbb{R}$-espace vectoriel) et que : $\forall x, y \in \mathbf{H}\ ,\ \sigma(xy) = \sigma(y)\sigma(x)$. On pose, pour tout $x \in \mathbf{H}$, $N(x) = x\sigma(x)$. Vérifier que :

$$\forall x, y \in \mathbf{H}\ ,\ N(xy) = N(x)N(y).$$

Écrire explicitement la formule correspondante ne mettant en jeu que des nombres réels. En déduire que le produit de deux sommes de quatre carrés de $\mathbb{N}$ est une somme de quatre carrés de $\mathbb{N}$.
Cette formule sert à démontrer le célèbre théorème de Lagrange : tout entier naturel est somme de quatre carrés.

Exercices sur la section 2

II.5.5 $*$ On va ramener la résolution de l'équation du 4ème degré à celle de l'équation du 3ème degré par la *méthode de Ferrari*.
(i) Démontrer que l'on peut se ramener à l'équation $z^4 + az^2 + bz + c = 0$.
(ii) Soit $w \in \mathbb{C}$. Calculer les complexes a', b', c' tels que l'égalité ci-dessus soit équivalente à l'égalité $(z^2 + w)^2 = a'z^2 + b'z + c'$.
(iii) À quelle condition portant sur w le trinôme $a'z^2 + b'z + c'$ est-il un carré parfait ? (On trouvera une équation du 3ème degré en w.) En déduire la résolution de l'équation du 4ème degré.

II.5.6 (i) Soient $p, q \in \mathbb{C}$. Démontrer que l'ensemble E des suites $\underline{u} = (u_n)_{n\in\mathbb{N}}$ telles que : $\forall n \in \mathbb{N}$, $u_{n+2} = pu_{n+1} + qu_n$ est un sous-espace vectoriel de l'espace vectoriel $\mathbb{C}^{\mathbb{N}}$ de toutes les suites de complexes.
(ii) Déterminer les éléments de E de la forme $u_n = x^n$. On démontrera que, si $p^2 + 4q \neq 0$, il y en a deux ; si $p^2 + 4q = 0$, il n'y en a qu'une, mais alors la suite $u_n = nx^n$ est dans E.
(iii) On note $\underline{v}$ et $\underline{w}$ les deux éléments particuliers de E trouvés ci-dessus. Démontrer qu'ils forment une base de E. (Pour que l'égalité $\underline{u} = a\underline{v} + b\underline{w}$, soit vérifiée, il faut, et il suffit, que l'on ait $u_0 = av_0 + bw_0$ et $u_1 = av_1 + bw_1$.)
(iv) On suppose p et q réels et l'on s'intéresse au sous-ensemble F des suites réelles de E. Montrer que c'est un sous-espace vectoriel du $\mathbb{R}$-espace vectoriel E et qu'il est de dimension 2. On en décrira une base dans chacun des trois cas : $p^2 + 4q > 0$, $p^2 + 4q = 0$, $p^2 + 4q < 0$.

II.5.7 Démontrer que l'ensemble $\mathbb{Z}[\mathrm{i}] := \{a + \mathrm{i}b \mid a, b \in \mathbb{Z}\}$ est un sous-anneau de $\mathbb{C}$ stable par conjugaison (ses éléments sont appelés *entiers de Gauß*). Démontrer que les normes algébriques des éléments de $\mathbb{Z}[\mathrm{i}]$ sont exactement les sommes de deux carrés dans $\mathbb{N}$. En déduire que l'ensemble des sommes de deux carrés dans $\mathbb{N}$ est stable par multiplication. Démontrer que les éléments inversibles de cet anneau sont ceux dont la norme algébrique vaut 1. (Si $zz' = 1$, alors $N(z)N(z') = 1$. Si $N(z) = 1$, $\overline{z} = z^{-1}$.) Énumérer ces éléments.

II.5.8 On rappelle que $j = \mathrm{e}^{2\mathrm{i}\pi/3} = \dfrac{-1}{2} + \mathrm{i}\dfrac{\sqrt{3}}{2}$.
On note $\mathbb{Z}[j]$ le sous-ensemble $\{a + bj \mid a, b \in \mathbb{Z}\}$ de $\mathbb{C}$. Démontrer que $\mathbb{Z}[j]$ est un sous-anneau de $\mathbb{C}$ stable par l'automorphisme de conjugaison. Caractériser les normes algébriques des éléments de $\mathbb{Z}[j]$. En déduire que l'ensemble des entiers de la forme $a^2 + ab + b^2$ avec $a, b \in \mathbb{Z}$ est stable par multiplication.

Exercices sur la section 3

II.5.9 Démontrer la formule suivante (*théorème de l'angle au centre*) :

$$\forall z \in \mathbb{C} \setminus \mathbb{R}_- \ , \ \operatorname{Arg} z = 2 \arctan \frac{\operatorname{Im} z}{|z| + \operatorname{Re} z} \ .$$

II.5.10 $*$ (i) Calculer les nombres $\sum\limits_{k=0}^{n} \cos k\theta$ et $\sum\limits_{k=0}^{n} \sin k\theta$. (Les considérer respectivement comme des parties réelle et imaginaire).
(ii) Calculer de même $\sum\limits_{k=0}^{n} k \cos k\theta$ et $\sum\limits_{k=0}^{n} k \sin k\theta$.

II.5.11 Linéariser l'expression $\cos^2 x \sin^3 x$.

II.5.12 Démontrer la proposition 14 de la page 256 à partir de la définition 10 et des rappels du cours de terminale qui la précèdent.

II.5.13 Montrer qu'il n'existe pas de fonction $\theta : \mathbb{C}^* \to \mathbb{R}$ telle que $\mathrm{e}^{\mathrm{i}\theta(z)} = z$ et telle que, de plus, $\theta(zz') = \theta(z) + \theta(z')$.
Indication : s'il existait une fonction θ, en prenant $f(z) = \mathrm{e}^{\mathrm{i}\theta(z)/2}$, on obtiendrait une fonction « racine carrée » qui soit un morphisme.

Exercices sur la section 4

II.5.14 $*$ (i) Démontrer que l'application $z \mapsto \dfrac{z-\mathrm{i}}{z+\mathrm{i}}$ définit une bijection de $\mathbb{C}\setminus\{-\mathrm{i}\}$ sur $\mathbb{C}\setminus\{1\}$, et que la bijection réciproque est l'application $w \mapsto \mathrm{i}\dfrac{1+w}{1-w}$.

(ii) On note $\mathcal{D}$ le disque unité ouvert $\overset{\circ}{D}(0,1)$ et $\mathcal{H}$ le *demi-plan de Poincaré* $\{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$.
Démontrer *géométriquement* que $z \in \mathcal{H}$ si, et seulement si, $\dfrac{z-\mathrm{i}}{z+\mathrm{i}} \in \mathcal{D}$.
En déduire une bijection de $\mathcal{H}$ sur $\mathcal{D}$.

II.5.15 (i) Soit x un complexe, $x \neq -\mathrm{i}$.
Démontrer *géométriquement* que $\left|\dfrac{1+\mathrm{i}x}{1-\mathrm{i}x}\right| = 1$ si, et seulement si, x est réel.

(ii) Soit a un réel. Démontrer que les racines de l'équation : $\left(\dfrac{1+\mathrm{i}x}{1-\mathrm{i}x}\right)^n = \dfrac{1+\mathrm{i}a}{1-\mathrm{i}a}$ sont les $x_k = \tan\dfrac{\alpha+k\pi}{n}$, $k = 0, \ldots, n-1$ *(mettre a sous la forme $\tan\alpha$, avec $\alpha \in]-\pi/2, \pi/2[$).*
On discutera la possibilité que l'un des x_k ne soit pas défini.

II.5.16 Démontrer que le groupe des similitudes directes est isomorphe au groupe obtenu en munissant l'ensemble $\mathbb{C}^* \times \mathbb{C}$ de la loi de composition interne : $(a,b) \star (a',b') = (aa', ab'+b)$.

Exercices sur la section 5

II.5.17 À l'aide de l'inégalité $\left|\displaystyle\sum_{n\geqslant 0} \frac{z^n}{(n+2)!}\right| \leqslant \displaystyle\sum_{n\geqslant 0} \left|\frac{z^n}{(n+2)!}\right|$, démontrer la majoration 8 de la page 272.

II.5.18 Démontrer que la seule fonction $f : \mathbb{R} \to \mathbb{R}$ telle que $f(0) = 1$ et $f' = f$ est l'exponentielle réelle. (Dériver $x \mapsto f(x)\mathrm{e}^{-x}$.)

II.5.19 Déduire la proposition 14 de la page 256 des résultats de la section 5.2.

Polynômes et fractions rationnelles

II.6

On a rencontré dans le secondaire le trinôme du second degré qui est un exemple de *fonction polynomiale* sur $\mathbb{R}$. Une partie des calculs sur les fonctions polynomiales est purement algébrique et les objets correspondants sont les *polynômes*, qui sont manipulés dans les systèmes de calcul formel comme Maple. Nous étudions dans ce module les lois qui les gouvernent. On se placera, en général, sur un corps commutatif quelconque : le cas de $\mathbb{R}$ et de $\mathbb{C}$, qui utilise un peu de topologie et d'analyse, sera détaillé en 2.

1 Polynômes sur un corps quelconque

Comme on l'a fait pour $\mathbb{C}$, on se fixe pour premier objectif de fonder rigoureusement l'algèbre des polynômes. Pratiquement, le lecteur qui le désire peut aborder directement l'étude de la section 1.2.

1.1 Construction et axiomes

Le problème posé est ici le suivant : comment adjoindre à un corps commutatif K un « élément indéterminé » X que l'on pourra, par la suite, remplacer par des valeurs arbitraires. Cependant, une grande partie des constructions et des raisonnements garde un sens et un intérêt sous des hypothèses plus générales. On se donne donc pour commencer un *anneau commutatif* A, auquel on veut adjoindre une *indéterminée* X de manière à obtenir un anneau commutatif $A[X]$.

Soit B un anneau commutatif qui contient A (comme sous-anneau) et X (comme élément). Le sous-anneau de B engendré par A et par X est, par définition (voir module II.2, définition 16 de la page 140) le plus petit sous-anneau de B contenant A et X. Il contient donc toutes les *expressions polynomiales en X à coefficients dans A*, c'est-à-dire les sommes $\sum_{n\geqslant 0} a_n X^n$ dont les coefficients a_n sont des éléments de A presque tous nuls. Rappelons qu'une suite $(a_n)_{n\in\mathbb{N}} \in A^{\mathbb{N}}$ est dite *à support fini* si ses termes sont nuls à partir d'un certain rang, et que le sous-ensemble de $A^{\mathbb{N}}$ formé des suites à support fini est noté $A^{(\mathbb{N})}$. On introduit donc l'ensemble :

$$C := \left\{ \sum_{n\geqslant 0} a_n X^n \;\middle|\; (a_n)_{n\in\mathbb{N}} \in A^{(\mathbb{N})} \right\}.$$

C'est un sous-anneau de B, et c'est donc le sous-anneau engendré par A et X. On le prouve en vérifiant qu'il contient 0 (prendre pour $(a_n)_{n\in\mathbb{N}}$ la suite constante nulle) et 1 (prendre la suite telle que $a_0 = 1$ et $a_i = 0$ pour $i \geqslant 1$. La stabilité pour les lois internes découle des formules suivantes :

$$\sum_{n\geqslant 0} a_n X^n \pm \sum_{n\geqslant 0} b_n X^n = \sum_{n\geqslant 0} (a_n \pm b_n) X^n, \tag{1}$$

$$\sum_{n\geqslant 0} a_n X^n \times \sum_{n\geqslant 0} b_n X^n = \sum_{n\geqslant 0} \left(\sum_{i=0}^{n} a_i b_{n-i} \right) X^n. \tag{2}$$

La dernière formule résulte du calcul suivant :

$$\sum_{n\geqslant 0} a_n X^n \times \sum_{n\geqslant 0} b_n X^n = \sum_{i,j\geqslant 0} a_i b_j X^{i+j} = \sum_{n\geqslant 0} \left(\sum_{i+j=n} a_i b_j \right) X^n.$$

La suite de terme général $c_n = \sum\limits_{i+j=n} a_i b_j = \sum\limits_{i=0}^{n} a_i b_{n-i}$ est bien dans $A^{(\mathbb{N})}$: soient $\alpha, \beta \in \mathbb{N}$ tels que $a_i = 0$ pour $i > \alpha$ et $b_j = 0$ pour $j > \beta$ (il en existe par hypothèse) ; alors $c_n = 0$ pour $n > \alpha + \beta$. La structure de l'anneau C sera complètement élucidée si l'on peut affirmer que l'écriture $P = \sum\limits_{n\geqslant 0} a_n X^n$ est unique.

Théorème et définition 1. Il existe un anneau commutatif contenant A (comme sous-anneau) et X (comme élément) dans lequel tout élément s'écrit $P = \sum\limits_{n\geqslant 0} a_n X^n$ de manière unique et dans lequel les lois de compositions internes sont données par les formules précédentes. Ses éléments sont appelés *polynômes à coefficients dans A en l'indéterminée X*. L'anneau de ces polynômes est noté $A[X]$. Le *polynôme nul* $0 = \sum\limits_{n\geqslant 0} 0X^n$ est le zéro de $A[X]$. Un polynôme $\sum\limits_{n\geqslant 0} a_n X^n$ tel que $a_n = 0$ pour $n \geqslant 1$ est appelé *polynôme constant* et identifié à l'élément a_0 de A.

On peut donc *identifier les coefficients de deux polynômes égaux* :

$$\sum_{n\geqslant 0} a_n X^n = \sum_{n\geqslant 0} b_n X^n \iff \forall n \in \mathbb{N}\ ,\ a_n = b_n, \tag{3}$$

$$\sum_{n\geqslant 0} a_n X^n = 0 \iff \forall n \in \mathbb{N}\ ,\ a_n = 0. \tag{4}$$

Démonstration. L'application $(a_n)_{n\in\mathbb{N}} \mapsto \sum\limits_{n\geqslant 0} a_n X^n$ sera au bout du compte une bijection de $A^{(\mathbb{N})}$ sur $A[X]$. On va donc définir directement une structure d'anneau sur $A^{(\mathbb{N})}$. Ses lois sont données par les formules :

$$(a_n)_{n\in\mathbb{N}} + (b_n)_{n\in\mathbb{N}} := (a_n + b_n)_{n\in\mathbb{N}},$$

$$(a_n)_{n\in\mathbb{N}} \times (b_n)_{n\in\mathbb{N}} := (c_n)_{n\in\mathbb{N}}, \text{ où } c_n := \sum_{i=0}^{n} a_i b_{n-i}.$$

La première loi est celle qui fait de $A^{\mathbb{N}}$ un groupe commutatif, et il est clair que $A^{(\mathbb{N})}$ en est un sous-groupe. On vérifie que la seconde loi est bien interne, c'est-à-dire que la suite $(c_n)_{n\in\mathbb{N}}$ est à support fini, par le même raisonnement que plus haut. Pour la commutativité, on remarque que :

$$\sum_{i=0}^{n} a_i b_{n-i} = \sum_{i+j=n} a_i b_j = \sum_{i+j=n} a_j b_i = \sum_{i=0}^{n} b_i a_{n-i}.$$

On montre ensuite que la suite $(e_n)_{n\in\mathbb{N}}$ de termes $1, 0, 0, 0, \ldots$ est élément neutre pour la multiplication : $\sum\limits_{i=0}^{n} a_i e_{n-i} = 0 + \cdots + 0 + a_n e_0 = a_n$ (cela suffit, en vertu de la commutativité). Pour l'associativité, on utilise la même remarque, d'où l'on tire mécaniquement les égalités :

$$\begin{aligned}\big((a_n)_{n\in\mathbb{N}} \times (b_n)_{n\in\mathbb{N}}\big) \times (c_n)_{n\in\mathbb{N}} &= (a_n)_{n\in\mathbb{N}} \times \big((b_n)_{n\in\mathbb{N}} \times (c_n)_{n\in\mathbb{N}}\big) \\ &= (d_n)_{n\in\mathbb{N}}, \text{ où } d_n = \sum_{i+j+k=n} a_i b_j c_k.\end{aligned}$$

Enfin, la commutativité permet de ne vérifier la distributivité que d'un seul côté :

$$\sum_{i+j=n} a_i(b_j + b'_j) = \sum_{i+j=n} a_i b_j + \sum_{i+j=n} a_i b'_j.$$

L'application $a \mapsto (a, 0, 0, 0 \ldots)$ est un isomorphisme de A sur un sous-anneau de $A^{(\mathbb{N})}$, et l'on identifie donc A à ce sous-anneau. Si l'on *définit* enfin X comme la suite $(0, 1, 0, 0, \ldots) \in A^{(\mathbb{N})}$, on démontre par récurrence que X^p est la suite dont le terme d'indice p vaut 1 et dont tous les autres termes sont nuls. Tout élément $P = (a_n)_{n\in\mathbb{N}}$ de $A^{(\mathbb{N})}$ s'écrit donc $\sum\limits_{n\geqslant 0} a_n X^n$ et cette écriture est unique. ∎

Notons de plus que l'anneau $A[X]$ est muni d'une loi externe : si $P := \sum\limits_{n\geqslant 0} a_n X^n \in A[X]$ et si $a \in A$, on peut poser $aP := \sum\limits_{n\geqslant 0} a a_n X^n$.

Proposition et définition 2. L'anneau $K[X]$ des polynômes sur un corps commutatif K, muni de cette loi externe, est un K-sous–espace vectoriel. La famille $(X^n)_{n\in\mathbb{N}}$ en est une base, appelée *base canonique*. L'espace vectoriel $K[X]$ est donc de dimension infinie.

Démonstration. La loi externe est celle qui fait de $K^{(\mathbb{N})}$ un espace vectoriel, dont $(X^n)_{n\in\mathbb{N}}$ est la base canonique (module II.3, section 3.3). ∎

Si A est un sous-anneau de K, alors $A[X]$ est un sous-anneau de $K[X]$.
Les anneaux $A[X]$ que nous rencontrerons dans ce volume seront presque tous de cette forme ; typiquement : $\mathbb{Z}[X] \subset \mathbb{Q}[X] \subset \mathbb{R}[X] \subset \mathbb{C}[X]$.

1.2 Règles élémentaires de calcul

Résumons la description de l'anneau $K[X]$ des polynômes à coefficients dans le corps commutatif K. C'est un K-espace vectoriel, dont la famille des X^n forme la *base canonique* ; autrement dit, tout polynôme s'écrit de manière unique $P = \sum\limits_{n\geqslant 0} a_n X^n$, les coefficients $a_n \in K$ étant nuls à partir d'un certain rang. En conséquence, l'addition et la soustraction sont donnés par les règles (1) de la page 282 et, si deux polynômes sont

égaux, on peut identifier leurs coefficients (règles (3) et (4) de la page 282). Il est muni de la multiplication donnée par la règle (2) et son neutre est $1 = 1X^0 + 0X^1 + \cdots$
Pour tout polynôme $P = \sum_{n\geqslant 0} a_n X^n$ et pour tout entier naturel k, on appellera respectivement *terme de degré* k et *coefficient de degré* k du polynôme P le *monôme* $a_k X^k \in K[X]$ et le *coefficient* $a_k \in K$ (les a_k sont donc les coordonnées de P dans la base canonique). En particulier, le *terme constant* de P est $\mathrm{tc}(P) := a_0$; c'en est aussi le *coefficient constant*
L'application $P \mapsto \mathrm{tc}(P)$ est une forme linéaire sur $K[X]$, et aussi un morphisme d'anneaux de $K[X]$ sur K :

$$\mathrm{tc}(\lambda P + \mu Q) = \lambda\ \mathrm{tc}(P) + \mu\ \mathrm{tc}(Q),$$
$$\mathrm{tc}(PQ) = \mathrm{tc}(P)\ \mathrm{tc}(Q).$$

Exemple. La formule du binôme de Newton, valide dans tout anneau commutatif (module II.2, théorème 32 de la page 134), entraîne $(1+X)^n = \sum_{i=0}^{n} \binom{n}{i} X^i$. De l'égalité $(1+X)^{n+p} = (1+X)^n(1+X)^p$, on tire, par identification du coefficient de degré k :

$$\forall k \in [\![0, n+p]\!]\ ,\ \binom{n+p}{k} = \sum_{i+j=k} \binom{n}{i}\binom{p}{j}.$$

Cette jolie formule admet également une preuve combinatoire (exercice II.1.0 de la page 105).
Un polynôme $P := \sum_{n\geqslant 0} a_n X^n \in K[X]$ et un élément x de K étant donnés, on peut *calculer la valeur de* P *en* x (on dit encore *évaluer* P *en* x ou *substituer l'élément* x *à l'indéterminée* X) en posant :

$$P(x) := \sum_{n\geqslant 0} a_n x^n \in K.$$

L'application $P \mapsto P(x)$ est K-linéaire et c'est un morphisme d'anneaux.
Par exemple, $P(0)$ est le terme constant $\mathrm{tc}(P)$ de P.

Exemple. En substituant 1 à X dans $1+X$, on trouve 2 ; en substituant 1 à X dans $(1+X)^n$, on trouve donc : $\sum_{i=0}^{n} \binom{n}{i} = 2^n$. En substituant -1 à X dans $1+X$, on trouve 0 ; en substituant -1 à X dans $(1+X)^n$, on trouve donc : $\sum_{i=0}^{n} \binom{n}{i}(-1)^n = 0^n$ qui vaut 1 si $n = 0$, et 0 sinon. Une preuve combinatoire de la première formule a été donnée au module II.1 (après le théorème 29 de la page 78) ; sauriez-vous en trouver une pour la seconde ?

Coût de l'évaluation.. Calculer la valeur de P en x n'est pas gratuit ! Cela *coûte* un certain nombre d'opérations. Nous admettrons ici que les additions et les multiplications d'éléments de K ont un *coût* constant. Pour de très grands nombres, cette hypothèse n'est pas réaliste. Ce point est discuté au module II.8. En général, (c'est-à-dire si aucun terme n'est

nul), l'expression $P(x) = a_0 + a_1 x + \cdots + a_d x^d$ comporte $d+1$ termes et son calcul donc nécessite d additions. Le calcul de $a_i x^i$ nécessite de même i multiplications, ce qui donne au total $0+1+\cdots+d = \dfrac{d(d+1)}{2}$ multiplications. En réalité, cette estimation est très pessimiste car le calcul de x^{i+1} *alors que l'on a déjà calculé* x^i ne nécessite qu'une multiplication. Voici donc un algorithme plus raisonnable et moins coûteux. On calcule $x^2, \ldots, x^d$; coût : $d-1$ multiplications. Puis l'on calcule $a_1 x, \ldots, a_d x^d$; coût : d multiplications. Finalement, il n'y a eu que $2d-1$ multiplications. Mais on peut faire mieux ! C'est le *schéma de Horner*. On pose $v_0 := a_d$, puis, pour $1 \leqslant i \leqslant d$: $v_i := v_{i-1}x + a_{d-i}$. On vérifie facilement que $v_i = a_{d-i} + a_{d-i+1}x + \cdots + a_d x^i$ (par récurrence sur i). En particulier, $v_d = P(x)$. Mais le calcul de chaque v_i pour $1 \leqslant i \leqslant d$ comporte une multiplication et une addition. Le coût du calcul de $P(x)$ est donc ici de d additions et d multiplications.

On peut, plus généralement, *substituer à* X *dans* P *un polynôme* $Q \in K[X]$.
Si $P := \sum\limits_{n \geqslant 0} a_n X^n$, on pose ainsi :

$$P(Q) := \sum_{n \geqslant 0} a_n Q^n.$$

Le polynôme $P(Q)$ est également noté $P \circ Q$. Le polynôme Q étant fixé, l'application $P \mapsto P(Q)$ de $K[X]$ dans $K[X]$ est K-linéaire et c'est un morphisme d'anneaux.

Définition 1.
Le polynôme P est dit *pair* si $P(-X) = P(X)$ et *impair* si $P(-X) = -P(X)$.

Ainsi, P *a la parité de* $k \in \mathbb{N}$ si, et seulement si, $P(-X) = (-1)^k P(X)$.
Pour $P = \sum\limits_{n \geqslant 0} a_n X^n$, on a $P(-X) = \sum\limits_{n \geqslant 0} (-1)^n a_n X^n$. Sur un corps de caractéristique 2, tous les polynômes sont donc pairs et impairs. Supposons le corps K de caractéristique différente de 2. Pour un polynôme pair, tous les coefficients a_{2n+1} sont nuls ; pour un polynôme impair, tous les coefficients a_{2n} sont nuls. Les polynômes pairs sont ceux qui peuvent s'écrire $P(X) = Q(X^2)$, les polynômes impairs ceux qui peuvent s'écrire $P(X) = XQ(X^2)$.

1.2.1 Le degré d'un polynôme

Tout polynôme non nul de $K[X]$ admet une unique écriture :

$$P(X) = a_0 + \cdots + a_d X^d \ , \quad \text{avec } a_d \neq 0.$$

Définition 2. Le *terme dominant* de P est $\mathrm{td}(P) := a_d X^d$ et son *coefficient dominant* est $\mathrm{cd}(P) := a_d$; ils ne sont définis que pour des polynômes non nuls. Le polynôme P est dit *unitaire* si son coefficient dominant est 1.

Proposition 3. L'anneau $K[X]$ est intègre et l'on a les règles suivantes :

$$\mathrm{td}(PQ) = \mathrm{td}(P)\,\mathrm{td}(Q), \qquad \mathrm{cd}(PQ) = \mathrm{cd}(P)\,\mathrm{cd}(Q).$$

Les éléments inversibles de $K[X]$ sont les polynômes constants non nuls.

Démonstration. Écrivons $P(X) := a_0 + \cdots + a_d X^d$ et $Q(X) := b_0 + \cdots + b_e X^e$, avec $a_d, b_e \neq 0$. Alors, de la définition de la multiplication dans $K[X]$, on tire :

$$PQ(X) = a_0 b_0 + \cdots + a_d b_e X^{d+e} \qquad \text{avec} \qquad a_d b_e \neq 0.$$

Le fait que $PQ \neq 0$ et les deux règles s'en déduisent immédiatement.
Il est clair que, si $P \in K^*$, P est inversible dans K donc dans $K[X]$. Réciproquement, si $PQ = 1$, alors P et Q sont non nuls, et, notant $aX^d = \mathrm{td}(P)$, $bX^e = \mathrm{td}(Q)$, on a $1 = \mathrm{td}(P)\,\mathrm{td}(Q) = abX^{d+e}$, d'où $ab = 1$ et $d + e = 0$, d'où $d = e = 0$. ∎

Deux polynômes P et Q sont donc *associés* (module II.2, définition 18 de la page 143) si, et seulement s'il existe $\lambda \in K^*$ tel que $Q = \lambda P$.

Définition 3. On appelle *degré du polynôme* P et l'on note $\deg P$ l'entier d ci-dessus si P est non nul. On convient que le degré du polynôme nul est $-\infty$.
Le symbole $-\infty$ qui apparaît ici est soumis aux axiomes suivants : pour tout $n \in \mathbb{Z}$,

$$-\infty < n \qquad \text{et} \qquad (-\infty) + n = n + (-\infty) = (-\infty) + (-\infty) = -\infty.$$

Par exemple, les polynômes de degré 0 sont les polynômes constants non nuls ; les polynômes de degré 1 (resp. de degré 2) sont les polynômes $aX + b$ (resp. $aX^2 + bX + c$) tels que $a \neq 0$.

Proposition 4. Soient P et Q deux polynômes quelconques. Alors :

$$\deg(PQ) = \deg P + \deg Q,$$

$$\deg(P + Q) \leqslant \max(\deg P, \deg Q),$$

et cette inégalité devient une égalité si $\deg P \neq \deg Q$ ou si $\deg P = \deg Q \geqslant 0$ et $\mathrm{cd}(P) + \mathrm{cd}(Q) \neq 0$, ainsi que pour $P = Q = 0$.

Démonstration. Si P et Q sont non nuls, la première relation est conséquence immédiate de la proposition précédente ; si l'un des deux est nul, elle découle des axiomes auxquels est soumis le symbole $-\infty$. Si P ou Q est nul, la seconde relation est immédiate (on a même égalité) ; si P et Q sont non nuls et opposés, c'est encore vrai (avec inégalité stricte). On peut donc supposer P, Q et $P + Q$ non nuls. Supposons d'abord $\deg P \neq \deg Q$, par exemple $\deg P < \deg Q$ (l'autre cas est symétrique). Alors, avec les notations de la proposition précédente, $(P + Q)(X) = (a_0 + b_0) + \cdots + (a_d + b_d)X^d + \cdots + b_e X^e$, avec $b_e \neq 0$. Ainsi, $\mathrm{td}(P + Q) = \mathrm{td}(Q)$ et $\deg(P + Q) = \deg Q$ dans ce cas. Supposons maintenant $\deg P = \deg Q$. Alors $(P + Q)(X) = (a_0 + b_0) + \cdots + (a_d + b_d)X^d$. Ainsi, si $a_d + b_d \neq 0$, $\mathrm{td}(P + Q) = \mathrm{td}(P) + \mathrm{td}(Q)$ et $\deg(P + Q) = \deg P = \deg Q$ dans ce cas. Si $a_d + b_d = 0$, on a l'inégalité stricte $\deg(P + Q) < \deg P = \deg Q$. ∎

Exemple. Supposons P et Q non nuls. En écrivant $P \circ Q = a_0 + \cdots + a_d Q^d$, on voit que $\deg(P \circ Q) = \deg P \deg Q$, avec une exception : celle où $Q = b_0$ est constant et $P \circ Q = P(b_0) = 0$. Par exemple, le degré de $P(X+a)$ est égal à celui de P ; en fait, ils ont même terme dominant, comme le montre le calcul suivant :

$$P(X+a) = a_0 + a_1(X+a) + \cdots + a_k \sum_{i=0}^{k} \binom{k}{i} a^{k-i} X^i + \cdots + a_d \sum_{i=0}^{d} \binom{d}{i} a^{d-i} X^i,$$

qui se réécrit $\sum_{k=0}^{d} \left(a_k + (k+1)a_{k+1}a + \cdots + \binom{d}{k} a_d a^{d-k} \right) X^k$ (ces coefficients seront « expliqués » plus loin par les formules de Taylor, théorème 35 de la page 304).
Il vient alors $\operatorname{td} P(X+a) = \operatorname{td} P$ et $\operatorname{tc} P(X+a) = P(a)$.

Proposition et définition 5. (i) Soit $n \in \mathbb{N}$. Les polynômes de degré inférieur ou égal à n forment un sous–espace vectoriel de $K[X]$, que l'on note $K_n[X]$. La famille $(1, X, \ldots, X^n)$ en forme la *base canonique*. La dimension de $K_n[X]$ est donc $n+1$.
(ii) Les polynômes P_k $(0 \leqslant k \leqslant n)$ (resp. $k \in \mathbb{N}$) forment une famille *à degrés étagés* si, pour tout k, $\deg P_k = k$. Alors $(P_0, \ldots, P_n)$ est une base de $K_n[X]$ (resp. $(P_n)_{n \in \mathbb{N}}$ est une base de $K[X]$).

Démonstration. (i) Par définition, un polynôme de degré $d \in \mathbb{N}$ est combinaison linéaire de $1, X, \ldots, X^d$. Comme $\deg 0 = -\infty$, on voit que $K_n[X]$ est engendré par la famille $(1, X, \ldots, X^n)$. Celle-ci est libre puisqu'extraite de la base canonique de $K[X]$.
(ii) Notons, pour $0 \leqslant k \leqslant n$, C_k le vecteur colonne des coordonnées de P_k dans la base canonique de $K_n[X]$ et soit $A = [C_0; \ldots; C_n] \in M_{n+1}(K)$. Pour que la famille des P_k soit à degrés étagés, il faut, et il suffit, que la matrice A soit triangulaire supérieure à coefficients diagonaux non nuls. On a vu au module II.4 (proposition 11 de la page 221) que A est alors inversible. Il en découle, d'après le théorème 21 de la page 343 du module II.7, que $(P_0, \ldots, P_n)$ est une base de $K_n[X]$. ∎

Exercice 1.

Sur un corps K de caractéristique nulle, on définit les *polynômes de Newton*

$$\binom{X}{n} := \frac{X(X-1)\cdots(X-n+1)}{n!} \qquad \text{(et donc } \binom{X}{0} = 1\text{)}.$$

Alors $\deg \binom{X}{n} = n$ et l'on a une famille à degrés étagés. Exprimer le polynôme $P = X^3 - 1$ dans la base formée de $\binom{X}{0} = 1$, $\binom{X}{1} = X$, $\binom{X}{2} = \frac{1}{2}X^2 - \frac{1}{2}X$, $\binom{X}{3} = \frac{1}{6}X^3 - \frac{1}{2}X^2 + \frac{1}{3}X$ ·

Solution. On a $Q := P - 6\binom{X}{3} = 3X^2 - 2X - 1$, puis $R := Q - 6\binom{X}{2} = X - 1$, puis $S := R - \binom{X}{1} = -1$, et enfin $T := S + \binom{X}{0} = 0$.
Finalement : $P = 6\binom{X}{3} + 6\binom{X}{2} + \binom{X}{1} - \binom{X}{0}$.

Avec des arguments analogues, on prouve que, si $\deg P_0, \ldots, \deg P_n$ sont deux à deux distincts, alors la famille des P_k $(0 \leqslant k \leqslant n)$ est libre. De même, si $\deg P_0, \ldots, \deg P_n$ prennent toutes les valeurs de 0 à d, alors $P_0, \ldots, P_n$ engendrent $K_d[X]$.

1.2.2 Division euclidienne

On dit que $A \in K[X]$ *divise* $B \in K[X]$, ce qui se note $A \mid B$, s'il existe $Q \in K[X]$ tel que $B = QA$. Par exemple, tout polynôme divise 0 ; et un polynôme inversible (c'est-à-dire une constante non nulle) divise tout polynôme. De plus, l'anneau $K[X]$ étant intègre, si $A \neq 0$ divise B, le polynôme Q est unique : c'est le *quotient* $Q = \dfrac{B}{A}$ ·

Soient par exemple $a, b \in K^*$ et $d, e \in \mathbb{N}$; alors $aX^d \mid bX^e$ si, et seulement si, $d \leqslant e$ et, dans ce cas, $\dfrac{bX^e}{aX^d} = \dfrac{b}{a}X^{e-d}$.

Lemme 6. Si $P \neq 0$ et si $Q \mid P$, alors $\deg Q \leqslant \deg P$ et $\deg \dfrac{P}{Q} = \deg P - \deg Q$.

Démonstration. Le polynôme Q n'est pas nul (sinon P le serait) et le polynôme $R := \dfrac{P}{Q}$ non plus (même raison). On a donc une relation $\deg P = \deg Q + \deg R$ entre éléments de $\mathbb{N}$, d'où les conclusions. ■

Théorème et définition 7. Soient $A, B \in K[X]$; on suppose $A \neq 0$. Il existe alors un unique couple $(Q, R) \in K[X]^2$ tel que :

$$B = QA + R, \qquad \text{avec} \qquad \deg R < \deg A.$$

Les polynômes Q et R sont respectivement appelés *quotient* et *reste* de la *division euclidienne de B par A* et notés $Q = B \operatorname{div} A$ et $R = B \mod A$.

On a de plus $\operatorname{td}(B \operatorname{div} A) = \dfrac{\operatorname{td}(B)}{\operatorname{td}(A)}$ ·

Ainsi, $A \mid B$ si, et seulement si, $B \mod A = 0$, et, dans ce cas, $B \operatorname{div} A = \dfrac{B}{A}$·

Démonstration. Prouvons d'abord l'unicité.

Si $B = QA + R = Q_1A + R_1$ avec $\deg R, \deg R_1 < \deg A$, alors $(Q - Q_1)A = R_1 - R$. Ainsi, $A \mid R_1 - R$ et $\deg A > \deg(R_1 - R)$, ce qui implique $R = R_1$ (lemme 6). On a donc $QA = Q_1A$, donc $Q = Q_1$ puisque $K[X]$ est intègre et A non nul.

Prouvons ensuite l'existence. Si $\deg A = 0$, A est inversible et l'on a $B = QA$ avec $R = 0$, donc $\deg R < \deg A$. On suppose donc $\deg A > 0$ et l'on raisonne par récurrence forte sur $\deg B$. Si $\deg B < \deg A$, le problème est résolu en prenant $Q := 0, R := B$. Si $\deg B \geqslant \deg A$, on va éliminer le terme dominant de B en lui soustrayant un multiple de A. Plus précisément, $\dfrac{\operatorname{td}(B)}{\operatorname{td}(A)}A$ a pour terme dominant $\operatorname{td}(B)$.

On pose donc $B_1 = B - \dfrac{\operatorname{td}(B)}{\operatorname{td}(A)}A$, et l'on a $\deg B_1 < \deg B$.

Par hypothèse de récurrence, $B_1 = Q_1A + R_1$ avec $\deg R_1 < \deg A$. Alors $B = QA + R$, avec $R = R_1$ et $Q = Q_1 + \dfrac{\operatorname{td}(B)}{\operatorname{td}(A)}$ · ■

Il découle d'ailleurs de cette démonstration que le terme dominant du quotient est $\mathrm{td}(Q) = \frac{\mathrm{td}(B)}{\mathrm{td}(A)}$.
On peut même en tirer un algorithme.
Algorithme de la division euclidienne.. On se donne B et A. Pour simplifier les notations, on pose $aX^d := \mathrm{td}(A)$: donc $a \in K^*$ et $d \in \mathbb{N}$. On pose $B_0 := B$, puis l'on définit les B_i par récurrence : si $\deg B_i \geqslant \deg A$, on pose $B_{i+1} := B_i - \dfrac{\mathrm{cd}(B_i)}{a} X^{\deg B_i - d} A$. Par construction, $\deg B_{i+1} < \deg B_i$, donc (récurrence) $\deg B_i \leqslant \deg B - i$. Il existe donc un entier $k \leqslant \deg B - \deg A$ tel que $\deg B_{k+1} < \deg A$. L'algorithme est alors terminé, et l'on a l'égalité :

$$B = \left(\sum_{i=0}^{k} \frac{\mathrm{cd}(B_i)}{a} X^{\deg B_i - d} \right) A + B_{k+1}.$$

Le quotient est donc $Q = \sum\limits_{i=0}^{k} \dfrac{\mathrm{cd}(B_i)}{a} X^{\deg B_i - d}$ et le reste est $R = B_{k+1}$. Le nombre d'étapes de l'algorithme est majoré par $\deg B - \deg A$. Notons que l'on n'a produit que les termes non nuls $\dfrac{\mathrm{cd}(B_i)}{a} X^{\deg B_i - d}$ du quotient.

Exercice 2.

Diviser $B := X^7 + X + 1$ par $A := X^3 + X + 1$.
Solution. On a :

$$\begin{aligned}
X^7 + X + 1 &= X^4(X^3 + X + 1) - X^5 - X^4 + X + 1 \\
-X^5 - X^4 + X + 1 &= -X^2(X^3 + X + 1) - X^4 + X^3 + X^2 + X + 1 \\
-X^4 + X^3 + X^2 + X + 1 &= -X(X^3 + X + 1) + X^3 + 2X^2 + 2X + 1 \\
X^3 + 2X^2 + 2X + 1 &= 1(X^3 + X + 1) + 2X^2 + X
\end{aligned}$$

Donc $Q = X^4 - X^2 - X + 1$ et $R = 2X^2 + X$.

Exercice 3.

Quel est le reste de la division de $P(X)$ par $X - a$?
Solution. C'est un polynôme de degré strictement plus petit que 1, donc constant. En écrivant $P(X) = Q(X)(X - a) + R(X)$ et en évaluant en $X = a$, on voit que $R = P(a)$. On peut le vérifier, et calculer du même coup le quotient, en remarquant que $X - a \mid X^k - a^k$ et que $\dfrac{X^k - a^k}{X - a} = \sum\limits_{i+j=k-1} a^i X^j$ (c'est une identité remarquable, valide dans tout anneau commutatif). On calcule donc le quotient $Q(X) = \dfrac{P(X) - P(a)}{X - a}$:

$$Q(X) = \sum_{k=0}^{d} a_k \sum_{i+j=k-1} a^i X^j = \sum_{j=0}^{d-1} \left(\sum_{i=0}^{d-1-j} a_{i+j+1} a^i \right) X^j.$$

Remarque. Le théorème 7 de la page précédente demeure valable sur un anneau commutatif quelconque, à la seule condition de supposer $\mathrm{cd}(A)$ inversible.

1.2.3 Ordre d'un polynôme

Tout polynôme non nul de $K[X]$ admet une unique écriture :

$$P(X) = a_k X^k + \cdots + a_d X^d\,, \quad \text{avec} \quad a_k \neq 0, \quad a_d \neq 0 \quad \text{et} \quad k \leqslant d.$$

Définition 4. On appelle *ordre de* P ou *valuation de* P l'entier :

$$\omega(P) := \min\{n \in \mathbb{N} \mid a_n \neq 0\}.$$

Par convention, l'ordre du polynôme nul est $+\infty$.
Le symbole $+\infty$ qui apparaît ici est soumis aux axiomes suivants : pour tout $n \in \mathbb{Z}$,

$$n < +\infty \qquad \text{et} \qquad (+\infty) + n = n + (+\infty) = (+\infty) + (+\infty) = +\infty.$$

L'intérêt de cette notion est qu'elle garde un sens pour les *séries formelles*, qui sont des sortes de « polynômes de degré infini » (voir le cours de L2). L'ordre a des propriétés proches de celles du degré. Soient P et Q deux polynômes quelconques. Alors :

$$\omega(PQ) = \omega(P) + \omega(Q) \qquad \text{et} \qquad \omega(P+Q) \geqslant \min\big(\omega(P), \omega(Q)\big),$$

et cette inégalité devient une égalité si $\omega(P) \neq \omega(Q)$ ou si $\omega(P) = \omega(Q) = k < +\infty$ et les coefficients en X^k de P et Q ne sont pas opposés (ainsi que pour $P = Q = 0$). Disons ensuite que les polynômes P_k $(0 \leqslant k \leqslant n)$ (resp. $k \in \mathbb{N}$) forment une famille *à valuations étagées* si, pour tout k, $\omega(P_k) = k$. Alors $(P_0, \ldots, P_n)$ est une base de $K_n[X]$ (resp. $(P_n)_{n\in\mathbb{N}}$ est une base de $K[X]$). Les preuves de ces affirmations sont similaires à celles de la proposition 4 de la page 286 et de la proposition 5 de la page 287. Notons enfin, pour $0 \leqslant k \leqslant n$, C_k le vecteur colonne des coordonnées de P_k dans la base canonique de $K_n[X]$ et soit $A = [C_0; \ldots; C_n] \in M_{n+1}(K)$. Pour que la famille des P_k soit à valuations étagées, il faut, et il suffit, que la matrice A soit triangulaire inférieure à coefficients diagonaux non nuls. Par exemple, le polynôme $X^k(1-X)^{n-k}$ s'écrit $X^k + \cdots + (-1)^{n-k} X^n$: son ordre est k. Les $X^k(1-X)^{n-k}$ $(0 \leqslant k \leqslant n)$ forment donc une base de $K_n[X]$.

1.3 Propriétés arithmétiques des polynômes

Cette section va être l'occasion de découvrir de frappantes analogies entre les anneaux $\mathbb{Z}$ et $K[X]$, dues au fait que ces deux anneaux sont munis d'une division euclidienne.
Remarquons d'abord que tout polynôme non nul est associé à un unique polynôme unitaire : on a vu que deux polynômes P et Q sont associés si, et seulement s'il existe $\lambda \in K^*$ tel que $Q = \lambda P$. L'unique polynôme unitaire associé à $P \neq 0$ est donc $\dfrac{1}{\mathrm{cd}(P)} P$. Pour énumérer les diviseurs d'un polynôme non nul P de degré $d \in \mathbb{N}$, on pourra donc se cantonner aux polynômes unitaires et rechercher ceux-ci de degré inférieur ou égal à d (lemme 6 de la page 288). Ses diviseurs de degré 0 sont les constantes non nulles, qui divisent bien P puisque ce sont des polynômes inversibles ; les quotients correspondants sont ses associés $\lambda P, \lambda \in K^*$. Ses diviseurs Q de degré d sont tels que $\deg \dfrac{P}{Q} = 0$, donc ce sont ses associés $\lambda P, \lambda \in K^*$ et les quotients correspondants sont inversibles.

Définition 5. Les *diviseurs triviaux* d'un polynôme P de degré d (donc non nul) sont les inversibles et ses associés. Un polynôme P non constant est dit *irréductible* si tous ses diviseurs sont triviaux (module II.2, définition 19 de la page 144) et *réductible* dans le cas contraire.

Le polynôme X^2+1 de $\mathbb{R}[X]$ est irréductible. Mais le même polynôme X^2+1, considéré comme élément de $\mathbb{C}[X]$, se factorise en $(X+\mathrm{i})(X-\mathrm{i})$. Cette notion est donc relative au corps de base K.

Exemples.

1. Un polynôme constant n'est ni réductible ni irréductible[1].
2. Le polynôme P de degré supérieur ou égal à 1 est irréductible si l'on a l'implication : $P=QR \Rightarrow \deg Q=0$ ou $\deg R=0$. Il est réductible s'il s'écrit $P=QR$ avec $\deg Q, \deg R \geqslant 1$. En particulier, tout polynôme de degré 1 est irréductible.
3. Soient $a \in K^*$ et $b,c,d \in K$. Les polynômes aX^2+bX+c et aX^3+bX^2+cX+d sont-ils réductibles ? Cherchons des diviseurs non triviaux et unitaires. Si le polynôme aX^2+bX+c a des diviseurs unitaires de degré 1, ceux-ci sont de la forme $Q=X-\alpha$, et le quotient de la forme $R=a'X+b'$. On calcule :

$$P=QR=a'X^2+(b'-\alpha a')X-\alpha b' \iff \begin{cases} a'=a \\ b'-\alpha a'=b \\ -\alpha b'=c \end{cases}$$

Ces équations impliquent $b'=b+\alpha a$, et elles sont compatibles si, et seulement si, $-\alpha(b+\alpha a)=c$. Les diviseurs de la forme $X-\alpha$ de aX^2+bX+c correspondent donc aux $\alpha \in K$ tels que $P(\alpha)=0$.

Les diviseurs non triviaux de aX^3+bX^2+cX+d sont de degré 1 ou 2, les quotients correspondants de degré 2 ou 1. Encore une fois, il suffit de chercher les diviseurs de la forme $Q=X-\alpha$, mais avec quotients $R=a'X^2+b'X+c'$. On calcule encore :

$$P=QR=a'X^3+(b'-\alpha a')X^2+(c'-\alpha b')X-\alpha c' \iff \begin{cases} a'=a \\ b'-\alpha a'=b \\ c'-\alpha b'=c \\ -\alpha c'=d \end{cases}$$

On trouve de nouveau pour condition nécessaire et suffisante de compatibilité l'égalité $P(\alpha)=0$. Dans les deux cas, pour que le polynôme P soit réductible, il faut, et il suffit, qu'il existe $\alpha \in K$ tel que $P(\alpha)=0$.

Théorème 8. Tout polynôme non constant est produit de polynômes irréductibles.

Démonstration. On le prouve par récurrence forte sur $\deg P$. Si $\deg P=1$, P est irréductible. Si $\deg P \geqslant 2$, soit P est irréductible, soit $P=QR$ avec $\deg Q, \deg R \geqslant 1$, donc $\deg Q, \deg R < \deg P$. Par hypothèse de récurrence, Q et R sont produits de polynômes irréductibles, donc $P=QR$ aussi. ∎

Quitte à remplacer chaque facteur irréductible par le polynôme unitaire associé et à sortir les coefficients dominants, on peut donc écrire : $P=\mathrm{cd}(P)P_1\cdots P_r$, où $P_1,\ldots,P_r$ sont irréductibles unitaires.

[1] De la même manière, l'entier naturel 1 n'est ni premier ni réductible.

Corollaire 9. Il y a une infinité de polynômes irréductibles unitaires (donc deux à deux non associés).

Démonstration. Si le corps K est infini, il suffit de considérer les $X - a$, $a \in K$. Dans le cas contraire, on utilise une preuve analogue à celle d'Euclide pour les nombres premiers de $\mathbb{N}$ (module II.1, théorème 35 de la page 85). Supposons qu'il n'y ait qu'un nombre fini de tels polynômes, soient $P_1, \ldots, P_k$; on sait que $k \neq 0$ (en fait, $k \geqslant \operatorname{card} K$ d'après le premier argument). Le polynôme $P = P_1 \cdots P_k + 1$ est de degré supérieur ou égal à k, donc divisible par un polynôme irréductible unitaire Q. Si Q était l'un des P_i, il diviserait $P_1 \cdots P_k$, donc également $P - P_1 \cdots P_k = 1$, ce qui est impossible. ■

Proposition 10. L'anneau $K[X]$ est principal.

Démonstration. Soit I un idéal de $K[X]$. Si $I = \{0\}$, il est principal. Sinon, soit $A \in I \setminus \{0\}$ de degré minimum (les degrés des éléments de $I \setminus \{0\}$ forment une partie non vide de $\mathbb{N}$) ; nous allons prouver que $I = < A >$. L'inclusion $< A > \subset I$ est triviale. Soit donc $B \in I$ quelconque ; on effectue la division euclidienne : $B = QA + R$. Alors $R = B - QA \in I$. Comme $\deg R < \deg A$, la condition de minimalité posée sur A implique $R = 0$, c'est-à-dire $B = QA$. Donc $I \subset < A >$. ■

Les générateurs d'un idéal non nul de $K[X]$ sont tous associés entre eux et l'un d'entre eux (et un seul) est unitaire. Rappelons (module II.2) qu'un pgcd de $F, G \in K[X]$ est un polynôme Δ tel que les diviseurs communs de F et G sont exactement les diviseurs de Δ. Si F et G ne sont pas tous deux nuls et s'ils admettent un pgcd, les associés de celui-ci sont tous leurs pgcd et l'on appellera celui qui est unitaire *le* pgcd de F et G ; on le notera $F \wedge G$. On conviendra que $0 \wedge 0 = 0$.

Théorème 11 (Théorème de Bézout pour les polynômes). Soient F et G deux polynômes non tous deux nuls. Ils admettent alors un pgcd $\Delta = F \wedge G$. C'est le générateur unitaire de l'idéal $< F, G >$ engendré par F et G, et il existe donc $A, B \in K[X]$ tels que $\Delta = AF + BG$.

Démonstration. C'est la proposition 44 de la page 144 du module II.2. ■

Comme pour les nombres entiers, on dira que deux polynômes F et G sont *premiers entre eux* ou *étrangers* s'ils n'ont pas de diviseur commun non inversible. D'après les considérations qui précèdent, cela équivaut à $F \wedge G = 1$.

Corollaire 12. Soit L un corps commutatif contenant le corps K.
Les polynômes $F, G \in K[X]$ sont premiers entre eux en tant qu'éléments de $K[X]$ si, et seulement s'ils le sont en tant qu'éléments de $L[X]$.

Démonstration. Si F et G ont un diviseur commun non constant dans $K[X]$, c'en est aussi un dans $L[X]$. Sinon, la relation $AF + BG = 1$ dans $K[X]$ entraîne qu'ils sont premiers entre eux en tant qu'éléments de $L[X]$. ■

Corollaire 13 (Lemme de Gauß pour les polynômes).
Supposons F et G premiers entre eux et soit $H \in K[X]$ tel que $F \mid GH$. Alors $F \mid H$.

Démonstration. Selon le théorème 11 de la page ci-contre, il existe A et B tels que $AF + BG = 1$. On a donc $H = AFH + BGH$; comme F divise AFH et BGH (car il divise GH par hypothèse), il divise H. ∎

Corollaire 14 (Lemme d'Euclide pour les polynômes).
Tout élément F irréductible est *premier*, c'est-à-dire tel que :

$$F \mid GH \Rightarrow F \mid G \text{ ou } F \mid H.$$

Théorème 15 (Théorème fondamental de l'arithmétique des polynômes).
Tout polynôme non nul P admet une factorisation :

$$P = CP_1 \cdots P_r,$$

où $C \in K^*$ et où $P_1, \ldots, P_r$ sont irréductibles unitaires.
Cette décomposition est unique à l'ordre près des facteurs irréductibles.

Démonstration. L'existence a déjà été établie. Prouvons l'unicité.
On suppose que $CP_1 \cdots P_r = DQ_1 \cdots Q_s$ (avec des notations évidentes). Un produit de polynômes unitaires est unitaire, on a donc $C = \mathrm{cd}(P) = D$, donc $P_1 \cdots P_r = Q_1 \cdots Q_s$. Si $r = 0$, le membre gauche vaut 1, donc $s = 0$. Sinon, P_1 divise l'un des Q_i d'après le lemme d'Euclide. Quitte à renuméroter les Q_i, on peut supposer que P_1 divise Q_1. Comme ils sont unitaires irréductibles, on a $P_1 = Q_1$, donc $P_2 \cdots P_r = Q_2 \cdots Q_s$, et l'on peut réitérer l'argument. ∎

On peut alors, comme au module II.1, équation (11) de la page 87, définir des *valuations*. Pour tout polynôme irréductible unitaire P et pour tout polynôme non nul F, notons $v_P(F)$ le nombre d'occurrences de P dans la décomposition ci-dessus : c'est donc l'unique entier k tel que F est divisible par P^k mais pas par P^{k+1}. On convient que $v_P(0) = +\infty$. Par exemple, l'ordre $\omega(F)$ est égal à $v_X(F)$. On a une décomposition :

$$F = \mathrm{cd}(F) \prod_P P^{v_P(F)}.$$

Le produit est étendu à tous les polynômes irréductibles unitaires P, mais presque tous les $v_P(F)$ sont nuls, donc presque tous les facteurs valent 1.
Rappelons la définition du *plus petit commun multiple* ou ppcm : les multiples communs à F et G (supposés non nuls) sont les éléments de l'idéal $< F > \cap < G >$. Celui-ci est principal, et tout générateur M satisfait la relation : $F \mid H \quad \text{et} \quad G \mid H \Leftrightarrow M \mid H$. On dit que H est un ppcm de F et G. Le générateur unitaire est noté $F \vee G$.

On a les règles suivantes de calcul et de raisonnement : $FG = \prod_P P^{v_P(F)+v_P(G)}$, $F \mid G$ si, et seulement si, $\forall P$, $v_P(F) \leqslant v_P(G)$, $F \wedge G = \prod_P P^{\min\left(v_P(F), v_P(G)\right)}$, $F \vee G = \prod_P P^{\max\left(v_P(F), v_P(G)\right)}$,

etc. Les règles suivantes, qui sont déjà valables pour les entiers, s'étendent sans peine au cas des polynômes :

1. Si F est étranger à $G_1, \ldots, G_k$, alors F est étranger à $G_1 \cdots G_k$.
2. Si $U \neq 0$ et si Δ est un pgcd de F et G, alors $U\Delta$ est un pgcd de UF et UG.

   De même, si $U \mid F$ et $U \mid G$ et si Δ est un pgcd de F et G, alors $\dfrac{\Delta}{U}$ est un pgcd de $\dfrac{F}{U}$ et $\dfrac{G}{U}$.

   Si Δ est un pgcd de F et G, $\dfrac{FG}{\Delta}$ en est un ppcm.

Nous allons maintenant décrire un moyen pratique de calculer le pgcd, l'*algorithme d'Euclide*. Comme dans le cas des entiers, il repose sur les deux idées suivantes. Si $G = 0$, $F \wedge G = \dfrac{1}{\mathrm{cd}(F)} F$ (c'est évident). Si $F = QG + R$, alors $F \wedge G = G \wedge R$. Cela découle de la double inclusion $< F, G > \subset < G, R >$ et $< G, R > \subset < F, G >$, dont la preuve (aisée) est laissée au lecteur ; les idéaux $< G, R >$ et $< F, G >$ étant égaux, leurs générateurs unitaires respectifs $G \wedge R$ et $F \wedge G$ sont égaux.

Algorithme d'Euclide. On se donne F et G tels que $\deg G \leqslant \deg F$ et $F \neq 0$. Le but est de calculer leur pgcd $\Delta = F \wedge G$. Si $G = 0$, le pgcd est $\Delta = \dfrac{1}{\mathrm{cd}(F)} F$. Supposons donc G non nul. On pose $F_0 := F, F_1 := G$, puis l'on définit des polynômes F_i par une récurrence à deux pas. Pour $i \geqslant 1$, si $F_i \neq 0$, on pose $F_{i+1} := F_{i-1} \mod F_i$. Donc, pour $i \geqslant 1$, $\deg F_{i+1} < \deg F_i$. Comme il n'existe pas de suite infinie strictement décroissante d'entiers naturels, il y a un entier k tel que $F_k \neq 0$ et $F_{k+1} = 0$. On a alors $\Delta = \dfrac{1}{\mathrm{cd}(F_k)} F_k$. En effet, le lemme implique que $F_{i-1} \wedge F_i = F_i \wedge F_{i+1}$. Par récurrence, on voit donc que $\Delta = F_0 \wedge F_1 = F_k \wedge F_{k+1}$.

On majore le nombre d'étapes k en remarquant que $\deg F_{i+1} \leqslant \deg F_i - 1$, d'où l'on tire :

$$\deg F_k \leqslant \deg F_0 - k \qquad \text{et} \qquad \deg F_k \leqslant \deg F_1 - (k-1),$$

et finalement $k \leqslant \min(\deg F, \deg G + 1)$.

Exercice 4.

Traiter le cas de : $F_0 = F := X^5 - 1$ et $F_1 = G := X^3 - 1$.

Solution. On a :

$$\begin{aligned} X^5 - 1 = X^2(X^3 - 1) + X^2 - 1 &\Longrightarrow F_2 = X^2 - 1 \\ X^3 - 1 = X(X^2 - 1) + X - 1 &\Longrightarrow F_3 = X - 1 \\ X^2 - 1 = (X+1)(X-1) + 0 &\Longrightarrow F_4 = 0 \end{aligned}$$

Le pgcd est donc $F_3 = X - 1$.

Exercice 5.

On définit des polynômes T_n par la relation de récurrence : $T_0 := 1$, $T_1 := X$ et, pour $n \geqslant 1$, $T_{n+1} := 2XT_n - T_{n-1}$.
Que dire de l'algorithme d'Euclide appliqué à T_n et T_{n-1} ?
Solution. On voit d'abord par récurrence que $\deg T_n = n$.
Par définition, les F_i qui apparaissent dans l'exécution de l'algorithme d'Euclide sont $T_n, T_{n-1}, -T_{n-2}, -T_{n-3}, T_{n-4}, \ldots$. Il y a donc n étapes, ce qui est le « cas le pire » du point de vue de l'estimation ci-dessus. Les T_n sont les *polynômes de Tchebychef*, que nous retrouverons plus loin (page 310).

On montre facilement par récurrence qu'à chaque étape, $F_i \in < F, G >$, donc Δ aussi, ce qui fournit une nouvelle preuve du théorème de Bézout. On peut même exploiter les calculs pour trouver les « coefficients de Bézout » à partir des quotients successifs. Dans l'exemple ci-dessus, on trouve successivement $F_2 = F_0 - Q_1F_1$, $F_3 = F_1 - Q_2F_2 = F_1 - Q_2(F_0 - Q_1F_1)$ et $\Delta = (1 + Q_1Q_2)G - Q_2F = (X^3 + 1)G - XF$. Nous allons systématiser cette idée. On se donne G et F. Le but est de calculer leur pgcd $\Delta := F \wedge G$ et des *coefficients de Bézout* A, B tels que $\Delta = AF + BG$. Si $F = 0$, le pgcd est $\Delta = \dfrac{1}{\mathrm{cd}(G)}G$, et l'on peut prendre $A = 0$ et $B = \mathrm{cd}(G)$.
Supposons donc F non nul. On pose, $F_0 := F$, $F_1 := G$, puis, tant que $F_i \neq 0$, on effectue la division euclidienne : $F_{i-1} := Q_iF_i + F_{i+1}$. D'autre part, on introduit des polynômes A_i, B_i avec l'idée d'avoir à chaque étape : $A_iF + B_iG = F_i$. On pose naturellement $A_0 := 1$, $B_0 := 0$, $A_1 := 0$ et $B_1 := 1$; puis, pour $i \geqslant 1$: $A_{i+1} := A_{i-1} - Q_iA_i$ et $B_{i+1} := B_{i-1} - Q_iB_i$. Supposons (hypothèse de récurrence) que l'on ait $A_{i-1}F + B_{i-1}G = F_{i-1}$ et $A_iF + B_iG = F_i$. Alors, par définition de ces suites, $A_{i+1}F + B_{i+1}G = (A_{i-1} - Q_iA_i)F + (B_{i-1} - Q_iB_i)G = F_{i-1} - Q_iF_i = F_{i+1}$. La relation désirée est établie par récurrence. Pour l'indice k final, cela donne $A_kF + B_kG = F_k$, donc : $\Delta = AF + BG$, avec $A := \dfrac{1}{\mathrm{cd}(F_k)}A_k$ et $B := \dfrac{1}{\mathrm{cd}(F_k)}B_k$. Nous allons voir que les coefficients de Bézout obtenus par cette méthode sont optimaux : ils ont les plus petits degrés possibles. Pour cela, étudions l'équation :

$$PF + QG = H, \tag{5}$$

d'inconnues $P, Q \in K[X]$, où $F, G \in K[X]$ sont fixés tous deux non nuls et $H \in K[X]$. Une condition nécessaire et suffisante est que H soit multiple du pgcd $\Delta = F \wedge G$. En effet, l'existence de P, Q équivaut à la relation $H \in < F, G > = < \Delta >$. Nous supposerons donc cette condition vérifiée : $H = \Delta H'$, avec $H' \in K[X]$. Sous cette hypothèse, il existe au moins une solution de l'équation (5). Soient en effet A_0, B_0 des coefficients de Bézout : $A_0F + B_0G = \Delta$. Alors $P_0 := A_0H'$ et $Q_0 := B_0H'$ sont bien solutions de (5). Pour déterminer toutes les solutions, nous introduisons les notations suivantes : $F' = \dfrac{F}{\Delta}$ et $G' = \dfrac{G}{\Delta}$.
Ce sont deux polynômes premiers entre eux et $A_0F' + B_0G' = 1$.

Proposition 16. Les solutions de l'équation (5) sont les couples :

$$(P, Q) = (P_0 + SG', Q_0 - SF') \text{ avec } S \in K[X].$$

Démonstration.
L'égalité $PF + QG = H$ équivaut à $(P - P_0)F = (Q_0 - Q)G$, donc à $(P - P_0)F' = (Q_0 - Q)G'$.

Comme F' et G' sont premiers entre eux, cette dernière égalité entraîne $F' \mid Q_0 - Q$, donc $Q = Q_0 - SF'$ pour un certain $S \in K[X]$. On a alors $(P - P_0)F' = SF'G'$, d'où $P = P_0 + SG'$. Réciproquement, pour tout $S \in K[X]$:

$$(P_0 + SG')F + (Q_0 - SF')G = P_0F + Q_0G + S(G'F - F'G) = H.$$

En prenant $H := \Delta$ et $H' := 1$, on en déduit qu'un couple (A_0, B_0) de coefficients de Bézout étant donné, les couples de coefficients de Bézout sont les couples $(A, B) := (A_0 + SG', B_0 - SF')$ avec $S \in K[X]$.

Corollaire 17. Il existe un unique couple (A, B) de coefficients de Bézout tels que $\deg A < \deg G'$ et $\deg B < \deg F'$.

Démonstration. Selon nos hypothèses, F' et G' sont non tous deux constants, et F et G non associés. L'unicité vient de ce que, d'après le corollaire précédent, A et B ne peuvent être que $A_0 \mod G'$ et $B_0 \mod F'$. Pour prouver l'existence, on fait la division euclidienne $B_0 = SF' + B$, et l'on pose $A := A_0 + SG'$. Le couple (A, B) est un couple de coefficients de Bézout tel que $\deg B < \deg F'$. Comme $AF' + BG' = 1$ et que AF' et BG' ne sont pas tous deux de degré 0, on a :

$$\deg A + \deg F' = \deg AF' = \deg BG' = \deg B + \deg G' < \deg F' + \deg G',$$

d'où $\deg A < \deg G'$.

Théorème 18.
L'algorithme d'Euclide ainsi *étendu* fournit les coefficients de Bézout de degré minimum.

Démonstration. D'après la preuve du corollaire, il suffit de tester le degré de A, par exemple. Reprenons les notations de l'algorithme. Pour $2 \leqslant i \leqslant k$, on voit par récurrence que $\deg A_i = \deg Q_1 + \cdots + \deg Q_{i-1}$. D'autre part, $\deg Q_i = \deg F_{i-1} - \deg F_i$. On en tire :

$$\deg A_k = \deg F_1 - \deg F_{k-1} < \deg F_1 - \deg F_k = \deg G - \deg \Delta = \deg G'.$$

1.4 Fonctions polynomiales et racines d'un polynôme

Définition 6. Soit $P = \sum_{n \geqslant 0} a_n X^n \in K[X]$ un polynôme sur le corps commutatif K. L'application $x \mapsto P(x) = \sum_{n \geqslant 0} a_n x^n$ de K dans K est appelée *fonction* ou *application polynomiale associée au polynôme* P.

Notons $\widetilde{P}$ cette application. On a donc $\forall a \in K\ ,\ \widetilde{P}(a) = P(a)$, mais il y aura souvent lieu de distinguer le polynôme P (objet « formel ») de la fonction $\widetilde{P}$ (objet « numérique »). L'application polynomiale associée à $\lambda P + \mu Q$ (resp. à PQ) est $\lambda\widetilde{P} + \mu\widetilde{Q}$ (resp. $\widetilde{P}\widetilde{Q}$). Autrement dit, l'application $P \mapsto \widetilde{P}$ de $K[X]$ dans $\mathcal{F}(K, K)$ est à la fois un morphisme

d'anneau et une application K-linéaire. Remarquons également que l'application polynomiale associée à $P \circ Q$ est $\widetilde{P} \circ \widetilde{Q}$. En particulier, si le polynôme P est pair (resp. impair), l'application $\widetilde{P}$ l'est aussi.

Théorème et définition 19. Soit $P \in K[X]$ un polynôme non nul et soit $a \in K$.
(i) On dit que a est *une racine* ou *un zéro de* P si $P(a) = 0$. Pour que a soit un zéro de P, il faut, et il suffit, que $X - a$ divise P.
(ii) On suppose que a est un zéro de P. Il existe alors un unique entier m tel que P est divisible par $(X-a)^m$ mais pas par $(X-a)^{m+1}$. On a $m \in [\![1, \deg P]\!]$. On peut alors écrire $P(X) = (X-a)^m Q(X)$, où $Q(a) \neq 0$. L'entier m est appelé *l'ordre* ou *la multiplicité* de la racine a. La racine a est dite *simple* si $m = 1$, *multiple* sinon.

Démonstration. Selon l'exercice 3 de la page 289, le reste de la division euclidienne de P par $X - a$ est $P(a)$, ce qui prouve la première affirmation. Par ailleurs, si P est divisible par $(X-a)^m$, alors $m = \deg(X-a)^m \leqslant \deg P$ puisque P n'est pas nul. Il y a donc un plus grand tel entier m, et il est compris entre 1 et $\deg P$. Si $Q = \frac{P}{(X-a)^m}$, par définition, Q n'est pas divisible par $X - a$, donc $Q(a) \neq 0$. ∎

Lemme 20. Soient $a_1, \ldots, a_k$ des racines deux à deux distinctes du polynôme non nul P et soient $m_1, \ldots, m_k$ leurs multiplicités.
Alors $P(X) = (X-a_1)^{m_1} \cdots (X-a_k)^{m_k} Q(X)$, où $Q(a_1), \ldots, Q(a_k) \neq 0$.

Démonstration. Puisque les a_i sont deux à deux distinctes, les $X - a_i$ sont premiers entre eux deux à deux, donc également les $(X-a_i)^{m_i}$; comme ils divisent tous P, leur produit divise P (section 1.3). Notant Q le quotient, celui-ci n'est divisible par aucun des $X - a_i$ (sinon l'ordre de la racine a_i de P serait strictement supérieur à m_i), donc les $Q(a_i)$ sont non nuls. ∎

Soient $a_1, \ldots, a_k$ toutes les racines de P et soient $m_1, \ldots, m_k$ leurs multiplicités. L'expression *le nombre de racines de P comptées avec leurs multiplicités* désigne alors la somme $m_1 + \cdots + m_k$ de ces multiplicités.

Théorème 21. Le nombre de racines d'un polynôme non nul P, comptées avec leurs multiplicités, est inférieur ou égal à $\deg P$.

Démonstration. D'après le lemme, $m_1 + \cdots + m_k \leqslant \deg P$. ∎

Corollaire 22. Si un polynôme $P \in K_n[X]$ s'annule en $n+1$ éléments de K deux à deux distincts, il est nul.

Exemple. Le noyau de l'application $P \mapsto \widetilde{P}$ de $K[X]$ dans $\mathcal{F}(K, K)$ est formé des polynômes dont tout $a \in K$ est racine. Si le corps K est infini, un tel polynôme est nul, d'après le corollaire. L'application $P \mapsto \widetilde{P}$ est alors injective et il est donc légitime d'identifier d'identifier un polynôme à la fonction polynomiale associée. Si le corps K est fini, le noyau est l'idéal engendré par le polynôme $\prod_{a \in K} (X - a)$, qui est de degré $\operatorname{card} K$.

Soient $a_1, \ldots, a_n$ des éléments deux à deux distincts de K.
Introduisons, pour tout $i \in [\![1, n]\!]$, les polynômes :

$$L_i := \prod_{j \in [\![1,n]\!] \setminus \{i\}} \frac{X - a_j}{a_i - a_j} .$$

Ces polynômes sont de degré n et $\forall i, j \in [\![1, n]\!]$, $L_i(a_j) = \delta_{i,j}$.

Théorème 23 (Polynôme d'interpolation de Lagrange).
Pour tout $(b_1, \ldots, b_n) \in K^n$, il existe un unique polynôme $P \in K_{n-1}[X]$ tel que $P(a_1) = b_1, \ldots, P(a_n) = b_n$. Il s'agit de

$$P = \sum_{i=1}^{n} b_i L_i.$$

Démonstration. Il est clair que le polynôme indiqué répond à la question. Si un autre polynôme Q y répond aussi, le polynôme $P - Q \in K_{n-1}[X]$ admet les n racines a_i, donc il est nul. ∎

Exemple. Démontrons que l'application $P \mapsto \widetilde{P}$ est surjective si, et seulement si, le corps K est fini. Si K est fini, toute application est polynomiale : en effet, notant $a_1, \ldots, a_n$ les éléments de K (donc $n = \operatorname{card} K$), et reprenant les notations ci-dessus, on voit que toute application $f : K \to K$ est interpolée par le polynôme $\sum_{i=1}^{n} f(a_i) L_i$. Si K est infini, soit $f : K \to K$ l'application telle que $f(0) = 1$ et qui est nulle sur K^*. Si elle était polynomiale, le polynôme correspondant (qui est unique comme on vient de le voir) aurait une infinité de racines (les éléments de K^*), donc serait nul, et f le serait aussi. Donc f n'est pas polynomiale.

Exercice 6.
Calculer $P \mod \prod_{i=1}^{n} (X - a_i)$.
Solution. C'est un polynôme de degré strictement inférieur à n qui prend la valeur $P(a_i)$ en a_i. C'est donc le polynôme $\sum_{i=1}^{n} P(a_i) L_i$.

Définition 7. Un polynôme non nul P est dit *scindé* ou *dissocié* s'il est produit de polynômes du premier degré (ou, de manière équivalente, s'il admet $\deg P$ racines comptées avec leurs multiplicités). Il est de plus dit *séparable* si toutes ses racines sont simples (ou, de manière équivalente, si ses $\deg P$ racines sont distinctes).

Le polynôme $X^2 + 1$ de $\mathbb{R}[X]$ est irréductible, il n'est certainement pas scindé. Mais le même polynôme $X^2 + 1$, considéré comme élément de $\mathbb{C}[X]$, se factorise en $(X + \mathrm{i})(X - \mathrm{i})$, il est donc scindé. Cette notion est donc relative au corps de base K.

Théorème et définition 24. Le corps commutatif K est dit *algébriquement clos* si les conditions équivalentes suivantes sont sont réalisées :
(i) Tout polynôme non constant sur K admet une racine au moins.
(ii) Tout polynôme non constant sur K est scindé.
(iii) Les polynômes irréductibles de $K[X]$ sont de degré 1.

Démonstration. La preuve de l'implication (i) $\Rightarrow$ (ii) se fait par récurrence sur le degré : un polynôme de degré 1 est, par définition, scindé. Un polynôme de degré supérieur ou égal à 2 admet une racine a, donc s'écrit $(X-a)Q$, où Q est scindé par hypothèse de récurrence. La réciproque (ii) $\Rightarrow$ (i) est triviale, puisque tout polynôme non constant scindé admet des facteurs de degré 1, donc des racines. Pour prouver que (ii) $\Rightarrow$ (iii), on remarque que, de façon générale, un polynôme irréductible scindé est nécessairement de degré 1. Supposons enfin que tous les polynômes irréductibles sont de degré 1. L'existence de la décomposition en facteurs irréductibles entraîne alors que tout polynôme non constant est scindé. ∎

Exemple. Le corps $\mathbb{C}$ est algébriquement clos (théorème 38 de la page 306).

On a vu dans l'enseignement secondaire que les racines x_1, x_2 d'un trinôme du second degré aX^2+bX+c sont liées aux coefficients par les égalités $x_1+x_2 = -\dfrac{b}{a}$ et $x_1x_2 = \dfrac{c}{a}$ ·
Nous allons généraliser ces relations. On fixe un entier $n \geqslant 1$.

Définition 8. Soient $x_1, \ldots, x_n$ des éléments du corps K (pas nécessairement distincts). Les *fonctions symétriques élémentaires des* x_i sont les expressions :

$$\forall k \in [\![1,n]\!]\ ,\ \sigma_k(x_1,\ldots,x_n) := \sum_{1\leqslant i_1<\cdots<i_k\leqslant n} x_{i_1}\cdots x_{i_k}.$$

Pratiquement :

$$\begin{aligned}
\sigma_1(x_1,\ldots,x_n) &= x_1+\cdots+x_n\\
\sigma_2(x_1,\ldots,x_n) &= \sum_{1\leqslant i<j\leqslant n} x_ix_j = x_1x_2+\cdots+x_{n-1}x_n\\
\sigma_k(x_1,\ldots,x_n) &= x_1x_2\cdots x_k+\cdots+x_{n-k+1}x_{n-k+2}\cdots x_n\\
\sigma_n(x_1,\ldots,x_n) &= x_1\cdots x_n
\end{aligned}$$

Ainsi, chaque $\sigma_k(x_1,\ldots,x_n)$ est la somme de $\binom{n}{k}$ monômes.

Théorème 25 (Relations entre coefficients et racines).
Soit $P := a_0+\cdots+a_nX^n = a_n(X-x_1)\cdots(X-x_n)$ un polynôme scindé de degré n. On a alors les relations suivantes :

$$\forall k \in [\![1,n]\!]\ ,\ \sigma_k(x_1,\ldots,x_n) = (-1)^k\frac{a_{n-k}}{a_n}\ \cdot$$

Démonstration. En développant le produit $(X-x_1)\cdots(X-x_n)$, on trouve :

$$X^n-(x_1+\cdots+x_n)X^{n-1}+\cdots+(-1)^nx_1\cdots x_n.$$

Le coefficient de degré k est $(-1)^k\sigma_k(x_1,\ldots,x_n)$. On conclut par identification. ∎

Exercice 7.
Avec les mêmes notations, calculer $\big(\sigma_1(x_1,\dots,x_n)\big)^2$, et en déduire une expression de $x_1^2+\cdots+x_n^2$ en fonction des a_i.
Exprimer de même $\dfrac{1}{x_1}+\cdots+\dfrac{1}{x_n}$ en fonction des a_i (en supposant les x_i non nuls).
Solution. De l'égalité (valable dans tout anneau commutatif) :

$$\Big(\sum_{1=1}^{n} x_i\Big)^2=\sum_{i=1}^{n} x_i^2+2\sum_{1\leqslant i<j\leqslant n} x_ix_j,$$

on tire $x_1^2+\cdots+x_n^2=\dfrac{a_{n-1}^2-2a_na_{n-2}}{a_n^2}$.
La condition sur les x_i équivaut à $a_0\neq 0$, et l'on vérifie alors facilement que :

$$\frac{1}{x_1}+\cdots+\frac{1}{x_n}=-\frac{a_1}{a_0}\cdot$$

On va calculer plus généralement les *sommes de Newton* :

$$S_k:=x_1^k+\cdots+x_n^k\ ,\ k\in\mathbb{N}.$$

Tout d'abord, des égalités $P(x_i)=0$, on tire, pour tout $k\in\mathbb{N}$: $a_nx_i^{n+k}+\cdots+a_0x_i^k=0$.
En sommant ces relations pour $i\in[\![1,n]\!]$, on obtient une relation de récurrence à n pas :

$$\forall k\in\mathbb{N}\ ,\ S_{k+n}=-\frac{a_{n-1}}{a_n}S_{k+n-1}-\cdots-\frac{a_0}{a_n}S_k.$$

Il suffit donc de déterminer $S_0,\dots,S_{n-1}$.
Il est évident que $S_0=n$ et que $S_1=\sigma_1(x_1,\dots,x_n)=-\dfrac{a_{n-1}}{a_n}$; S_2 a été calculé plus haut. Plus généralement :

$$\forall k\in[\![0,n-1]\!]\ ,\ (n-k)a_{n-k}=a_nS_k+a_{n-1}S_{k-1}+\cdots+a_{n-k}S_0. \tag{6}$$

Nous démontrerons cette relation à la page 303. On en déduit, par exemple :

$$S_3=\frac{-3a_{n-3}a_n^2+3a_na_{n-1}a_{n-2}-a_{n-1}^3}{a_n^3}\cdot$$

Réciproquement, *si le corps K est de caractéristique nulle*, ces mêmes relations permettent de calculer P à un facteur constant près : on fixe a_n arbitraire non nul, puis l'on calcule $a_{n-1},\dots,a_0$ par récurrence descendante.

1.4.1 Application aux corps finis

Soit K un corps fini commutatif à q éléments.[2]

Théorème 26. Le polynôme $X^q-X\in K[X]$ est scindé et séparable :

$$X^q-X=\prod_{a\in K}(X-a).$$

Démonstration. Le groupe multiplicatif K^* a $(q-1)$ éléments. D'après le théorème de Lagrange,

[2]En fait, tout corps fini est commutatif (théorème de Wedderburn).

tout $x \in K^*$ vérifie donc $x^{q-1} = 1$, donc $x^q = x$. Comme c'est évidemment le cas pour $x = 0$, on voit que tous les éléments de K sont racines de $X^q - X$, de sorte que $\prod_{a \in K} (X - a)$ le divise. Mais ces deux polynômes sont unitaires de même degré q, donc égaux. ■

Théorème 27. Le groupe multiplicatif K^* est cyclique.

Démonstration. On va démontrer, plus généralement, que tout sous-groupe fini du groupe multiplicatif d'un corps commutatif est cyclique. Soit donc G un sous-groupe fini du groupe multiplicatif K^* d'un corps commutatif K. Soit $n := \operatorname{card} G$. Le théorème de Lagrange implique donc que $\forall x \in G, x^n = 1$. Pour tout $x \in G$, l'ordre de x est donc un diviseur de n. Notons, pour tout diviseur d de n, G_d l'ensemble des $x \in G$ d'ordre d et $\psi(d) := \operatorname{card} G_d$. On a donc : $\sum_{d|n} \psi(d) = n$. Par ailleurs, si x est d'ordre d, il engendre un sous-groupe cyclique $< x >$ d'ordre d de G dont tous les éléments vérifient $y^d = 1$. Comme le polynôme $X^d - 1$ admet au plus d racines, ses racines sont exactement les éléments de $< x >$. En particulier, tous les éléments de G_d sont dans $< x >$. Comme le groupe cyclique $< x >$ admet $\varphi(d)$ éléments d'ordre d (cela découle de la proposition 49 de la page 147 du module II.2), on voit que $\psi(d) = \varphi(d)$ dans ce cas. Pour tout diviseur d de n, il y a donc une dichotomie : soit il n'y a aucun élément d'ordre d dans G et $\psi(d) = 0$, soit il y en a et $\psi(d) = \varphi(d)$. Dans tous les cas, $\psi(d) \leqslant \varphi(d)$. Mais (exercice 9 de la page 89 du module II.1), on a : $\sum_{d|n} \varphi(d) = n$. On a donc $\sum_{d|n} \big(\varphi(d) - \psi(d)\big) = 0$ et $\forall d \mid n\ ,\ \varphi(d) - \psi(d) \geqslant 0$, d'où $\forall d \mid n\ ,\ \varphi(d) - \psi(d) = 0$. En particulier, $\psi(n) = \varphi(n)$, qui est non nul, ce qui signifie très précisément que G est cyclique. ■

On a déjà prouvé que les sous-groupes finis de $\mathbb{C}^*$ étaient cycliques. La preuve avait l'avantage d'être explicite en ce sens que l'exponentielle complexe nous permettait de décrire tous les générateurs.

1.5 Polynômes dérivés

Lorsque le corps K est $\mathbb{R}$, la fonction polynomiale $\widetilde{P}$ est dérivable et sa dérivée est elle-même polynomiale.

L'application $\widetilde{P}'$ est associée au polynôme $P' = \sum_{n \geqslant 0} (n+1) a_{n+1} X^n \in \mathbb{R}[X]$. Ce polynôme peut être défini de manière purement algébrique, sans référence à un passage à la limite et, d'ailleurs, sur un corps commutatif quelconque.

Définition 9. Soit $P = \sum_{n \geqslant 0} a_n X^n \in K[X]$ un polynôme sur le corps commutatif K.

Le *polynôme dérivé* est le polynôme $P' := \sum_{n \geqslant 0} (n+1) a_{n+1} X^n \in K[X]$.

L'application $P \mapsto P'$ de $K[X]$ dans lui-même est appelée *dérivation*.

Si P est constant, P' est nul. Plus généralement, $\deg P' \leqslant \deg P - 1$.

Proposition 28. L'application de dérivation est un endomorphisme du K-espace vectoriel $K[X]$ qui satisfait la *règle de Leibniz* :

$$\forall P, Q \in K[X]\ ,\ (PQ)' = P'Q + PQ'.$$

Démonstration. La linéarité de $P \mapsto P'$ est évidente. Les applications $P \mapsto (PQ)'$ et $P \mapsto P'Q + PQ'$ étant linéaires, il suffit de vérifier l'égalité lorsque P est élément de la base canonique ; de même, les applications $Q \mapsto (PQ)'$ et $Q \mapsto P'Q + PQ'$ étant linéaires, il suffit de vérifier l'égalité lorsque Q est élément de la base canonique. Si $P = 1$, l'égalité à démontrer s'écrit $Q' = 0Q + Q'$, qui est évidente ; similairement lorsque $Q = 1$. Supposons donc $P := X^p, Q := X^q$ avec $p, q \geqslant 1$. On calcule alors $(PQ)' = (p+q)X^{p+q-1} = pX^{p-1}X^q + X^p(qX^{q-1}) = P'Q + PQ'$. ∎

Les *dérivées successives* de $P \in K[X]$ sont définies par itération de la dérivation. On les note : $P'' := (P')'$, $P''' := (P'')'$, $P^{(4)} := (P''')'$, et, en général : $P^{(n+1)} := (P^{(n)})'$. Les dérivées successives de $P := X^m$ sont donc :

$$P' = mX^{m-1}, P'' = m(m-1)X^{m-2}, \ldots, P^{(k)} = \frac{m!}{(m-k)!}X^{m-k}, \ldots, P^{(m)} = m!.$$

Selon les conventions usuelles sur l'itération, on pose $P^{(0)} := P$.

Corollaire 29 (Formule de Leibniz). Soient $P, Q \in K[X]$ et soit $n \in \mathbb{N}$. Alors :

$$(PQ)^{(n)} = \sum_{i=0}^{n} \binom{n}{i} P^{(i)} Q^{(n-i)}.$$

Démonstration. Comme pour la proposition 28, on peut prendre $P = X^p$ et $Q = X^q$. Il s'agit alors de démontrer la formule : $\dfrac{(p+q)!}{(p+q-n)!} = \displaystyle\sum_{i+j=n, i\leqslant p, j\leqslant q} \frac{p!}{(p-i)!}\frac{q!}{(q-j)!}$. Mais celle-ci est une conséquence facile de l'exercice II.1.0 de la page 105. ∎

Corollaire 30. Soient $P_1, \ldots, P_n \in K[X]$. Alors :

$$(P_1 \cdots P_n)' = P_1' \cdots P_n + \cdots + P_1 \cdots P_n' = \sum_{i=1}^{n} \left(\prod_{j \in [\![0,n]\!] \setminus \{i\}} P_j \right) P_i'.$$

Démonstration. Cette formule se prouve par récurrence sur n, à l'aide de la règle de Leibniz : nous laisserons ce soin au lecteur. ∎

Corollaire 31. Soient $P, Q \in K[X]$ et soit $n \geqslant 1$. Alors :

$$(P^n)' = nP^{n-1}P' \qquad \text{et} \qquad (P \circ Q)' = (P' \circ Q)Q'.$$

Démonstration. La première formule se prouve en posant $P_1 = \cdots = P_n = P$ dans le corollaire précédent. Pour la deuxième, on remarque que les deux membres de l'égalité

sont fonctions linéaires de P (mais pas de Q !), et que l'on peut donc prendre $P = X^n$. Si $n = 0$, l'égalité est évidente, si $n \geqslant 1$, c'est celle qui vient d'être établie. ■

Exemple. Soit $P(X) := \prod_{i=1}^{n}(X - a_i)$. Alors $P'(X) = \sum_{i=1}^{n} \prod_{j\in[\![1,n]\!]\setminus\{i\}} (X - a_j)$. Si l'on évalue P' en a_k, on constate que tous les termes de cette somme s'annulent, sauf celui d'indice k, qui vaut : $P'(a_k) = \prod_{j\in[\![1,n]\!]\setminus\{k\}} (a_k - a_j)$.

Si l'on pose $\omega = \prod_{i\in[\![1,n]\!]} (X - a_i)$, on en déduit, en supposant les a_i deux à deux distincts, la formule : $L_i = \dfrac{\omega}{\omega'(a_i)(X - a_i)}$.

Application aux sommes de Newton. Reprenons les notations $S_k := x_1^k + \cdots + x_n^k$ et $P := a_0 + \cdots + a_n X^n = a_n(X - x_1)\cdots(X - x_n)$, où $a_n \neq 0$. Pour tout $k \in \mathbb{N}$, le quotient de la division euclidienne de $X^{k+1}P'$ par P est $S_0X^k + \cdots + S_k$. En effet, les calculs précédents montrent que $P' = \dfrac{P}{X - x_1} + \cdots + \dfrac{P}{X - x_n}$, d'où :

$$X^{k+1}P' = \left(\sum_{i=1}^{n} \frac{X^{k+1} - x_i^{k+1}}{X - x_i}\right) P + \sum_{i=1}^{n} x_i^{k+1} \frac{P}{X - x_i} \cdot$$

Comme $\displaystyle\sum_{i=1}^{n} \frac{X^{k+1} - x_i^{k+1}}{X - x_i} = S_0X^k + \cdots + S_k \in K[X]$ et comme le degré de $\displaystyle\sum_{i=1}^{n} x_i^{k+1} \frac{P}{X - x_i}$ est strictement inférieur à n, c'est bien une division euclidienne. Les polynômes $X^{k+1}P'$ et $(S_0X^k + \cdots + S_k)P$ ont donc même terme de degré n. Or, le coefficient du premier est $(n - k)a_{n-k}$ et celui du second est $a_nS_k + a_{n-1}S_{k-1} + \cdots + a_{n-k}S_0$, ce qui établit la relation de récurrence (6), page 300, dont nous avions laissé la preuve en suspens.

1.5.1 Cas d'un corps de caractéristique nulle

Nous supposons maintenant que le corps K est de caractéristique nulle. Le cas où la caractéristique est non nulle sera abordé à la page 305.

Proposition 32. Le noyau de l'application de dérivation est formé des constantes. Pour un polynôme P non constant, $\deg P' = \deg P - 1$.

Démonstration. Si $P \in K$, on sait déjà que $P' = 0$. Sinon, si l'on note $\mathrm{td}(P) = ax^n$, avec $a \in K^*$ et $n \geqslant 1$, on a $P = aX^n + Q$ où $\deg Q \leqslant n - 1$.

On en tire $P' = naX^{n-1} + Q'$ où $\deg Q' \leqslant n - 2$. Comme $na \in K^*$ *parce que la caractéristique de K est nulle*, on a $\mathrm{td}(P') = naX^{n-1}$ et $\deg P' = n - 1$. ■

Corollaire 33. Soient $P \in K[X]$ et $n \in \mathbb{N}$. Alors :

$$P^{(n)} = 0 \Longleftrightarrow \deg P < n.$$

Lemme 34. Soient U et V deux éléments quelconques de $K[X]$ et soit $P \in K_n[X]$. Alors :

$$P(U+V) = \sum_{k=0}^{n} \frac{P^{(k)}(U)}{k!} V^k = \sum_{k \geqslant 0} \frac{P^{(k)}(U)}{k!} V^k.$$

Démonstration. Les deux membres de la première égalité sont des fonctions linéaires de $P \in K_n[X]$. On peut donc prendre pour P un élément de la base canonique de $K_n[X]$; supposons $P := X^m$, avec $m \leqslant n$. Alors :

$$P^{(k)} = \frac{m!}{(m-k)!} X^{m-k} \quad \text{si} \quad k \leqslant m \qquad \text{et} \qquad P^{(k)} = 0 \quad \text{si} \quad m < k \leqslant n.$$

On a donc, d'après la formule du binôme :

$$\sum_{k=0}^{n} \frac{P^{(k)}(U)}{k!} V^k = \sum_{k=0}^{m} \frac{1}{k!} \frac{m!}{(m-k)!} U^{m-k} V^k = (U+V)^m = P(U+V).$$

La deuxième égalité vient de ce que, si $P \in K_n[X]$ et si $k > n$, alors $P^{(k)} = 0$. ■

Théorème 35 (Formules de Taylor). Soient $P \in K[X]$ et $a \in K$. Alors :

$$P(X+a) = P(a) + XP'(a) + \cdots + \frac{X^n}{n!} P^{(n)}(a) = \sum_{k \geqslant 0} \frac{X^k}{k!} P^{(k)}(a),$$

$$P(X+a) = P(X) + aP'(X) + \cdots + \frac{a^n}{n!} P^{(n)}(X) = \sum_{k \geqslant 0} \frac{a^k}{k!} P^{(k)}(X),$$

$$P(X) = P(a) + P'(a)(X-a) + \cdots + \frac{P^{(n)}(a)}{n!} (X-a)^n = \sum_{k \geqslant 0} \frac{P^{(k)}(a)}{k!} (X-a)^k,$$

$$P(X) = P(0) + P'(0)X + \cdots + \frac{P^{(n)}(0)}{n!} X^n = \sum_{k \geqslant 0} \frac{P^{(k)}(0)}{k!} X^k,$$

où n est n'importe quel entier supérieur ou égal à $\deg P$.

Démonstration. Il suffit d'appliquer le lemme, d'abord avec $U := a$ et $V := X$, ensuite avec $U := X$ et $V := a$, enfin avec $U := a$ et $V := X - a$; puis de faire $a := 0$. ■

Théorème 36. (i) Soit P un polynôme non nul sur le corps K. Pour que $a \in K$ soit racine multiple de P, il faut, et il suffit, que a soit racine de P et de P'.
(ii) Soit m un entier non nul. Pour que $a \in K$ soit racine de P d'ordre supérieur ou égal à m, il faut, et il suffit, que a soit racine de $P, P', \ldots, P^{(m-1)}$, c'est-à-dire que l'on ait $P(a) = P'(a) = \cdots = P^{(m-1)}(a) = 0$.

Démonstration. (i) Si $a \in K$ est racine multiple de P, alors :

$$P = (X-a)^2 Q, \qquad P' = (X-a)\big(2Q + (X-a)Q'\big) \quad \text{et} \quad P(a) = P'(a) = 0.$$

Si $P(a) = P'(a) = 0$, on a $P = (X-a)Q$ et $P' = (X-a)Q' + Q$, d'où $Q(a) = 0$, et $(X-a)^2 \mid P$.
(ii) D'après le théorème 35 de la page ci-contre, le reste de la division de P par $(X-a)^m$ est $P(a)+(X-a)P'(a)+\cdots+\dfrac{(X-a)^{m-1}}{(m-1)!}P^{(m-1)}(a)$. Le polynôme P est donc divisible par $(X-a)^m$ si, et seulement si, l'on a $P(a) = P'(a) = \cdots = P^{(m-1)}(a) = 0$. ■

Exemple. On suppose P scindé : $P = C(X-a_1)^{\mu_1}\cdots(X-a_k)^{\mu_k}$, où $C \neq 0$, les racines a_i sont deux à deux distinctes et les multiplicités $\mu_i \in \mathbb{N}^*$. Le pgcd de P et P' doit diviser P, aucune racine simple de P ne doit en être racine et chaque racine multiple a_i doit en être racine avec la multiplicité $\mu_i - 1$. Donc : $P \wedge P' = (X-a_1)^{\mu_1-1}\cdots(X-a_k)^{\mu_k-1}$. Le polynôme $\dfrac{P}{P \wedge P'} = C(X-a_1)\cdots(X-a_k)$ a donc les mêmes racines que P, mais elles sont simples, il est donc séparable. On l'appelle *radical de* P.

1.5.2 Cas d'un corps de caractéristique non nulle

Indiquons rapidement comment certains des résultats précédents doivent être modifiés lorsque le corps K est de caractéristique $p > 0$. Tout d'abord, *le noyau de l'application de dérivation est formé des polynômes de la forme* $P(X^p)$. Soit en effet $P = \sum\limits_{n\geqslant 0} a_n X^n$. L'égalité $P' = 0$ équivaut à $\forall n, na_n = 0$. Cela revient à dire que les seuls coefficients a_n non nuls sont ceux pour lesquels $p \mid n$ et l'on a alors : $P = \sum\limits_{m\geqslant 0} a_{pm}X^{pm} = Q(X^p)$, où $Q(X) = \sum\limits_{m\geqslant 0} a_{pm}X^m$.

La formule de Taylor n'a évidemment aucun sens (sauf pour des degrés strictement inférieurs à p), puisqu'elle comporte des divisions par $k!$. On peut cependant généraliser le lemme de la manière suivante. On introduit, pour tout $k \in \mathbb{N}$, l'unique endomorphisme Δ_k de $K[X]$ tel que $\Delta_k(X^n) = 0$ si $n < k$ et $\Delta_k(X^n) = \binom{n}{k}X^{n-k}$ si $n \geqslant k$. Les Δ_k sont parfois appelés *dérivations supérieures*. Le même calcul que précédemment donne la formule :

$$\forall P \in K[X]\ ,\ \forall U, V \in K[X]\ ,\ P(U+V) = \sum_{k\geqslant 0} \Delta_k P(U)V^k.$$

Naturellement, Δ_1 est l'endomorphisme de dérivation. Sans hypothèse sur la caractéristique, on prouve facilement : $\Delta_k(PQ) = \sum\limits_{i+j=k} \Delta_i(P)\Delta_j(Q)$ et $\Delta_k \circ \Delta_\ell = \binom{k+\ell}{k}\Delta_{k+\ell}$. En caractéristique nulle, on déduirait de cette dernière relation que $\Delta_k(P) = \dfrac{P^{(k)}}{k!}$, puis, de la première, la formule de Leibniz (et du lemme, les formules de Taylor). Ici, les relations entre les Δ_k sont beaucoup plus compliquées.
La caractérisation des racines multiples par $P(a) = P'(a) = 0$ est valable en toute caractéristique, mais pas la caractérisation de l'ordre d'une racine, comme le montre l'exemple de X^p dont toutes les dérivées sont nulles.

2 Polynômes sur les corps $\mathbb{R}$ et $\mathbb{C}$

Les corps $\mathbb{R}$ et $\mathbb{C}$ étant infinis, nous identifierons tout polynôme P de $\mathbb{R}[X]$ ou de $\mathbb{C}[X]$ avec la fonction polynomiale associée. Si $P \in \mathbb{R}[X]$, celle-ci est continue et dérivable et sa dérivée (en tant que fonction de variable réelle) est l'application $x \mapsto P'(x)$ associée au polynôme dérivé. Les deux notions de dérivée coïncident donc parfaitement. Les faits suivants sont démontrés dans les modules IV.6 et IV.7, ou s'en déduisent immédiatement.

Soient $t \mapsto u(t)$ et $t \mapsto v(t)$ deux fonctions dérivables d'un intervalle ouvert I de $\mathbb{R}$ dans $\mathbb{C}$ (ce sont donc des fonctions vectorielles). Soient $\lambda, \mu \in \mathbb{C}$.
Alors les fonctions $f = \lambda u + \mu v$ et $g = uv$ sont dérivables de dérivées respectives $f' = \lambda u' + \mu v'$ et $g' = u'v + uv'$.
Soit $t \mapsto x(t)$ une fonction dérivable d'un intervalle ouvert I de $\mathbb{R}$ dans $\mathbb{C}$. Pour tout $P \in \mathbb{C}[X]$, la fonction $t \mapsto P\big(x(t)\big)$ est dérivable et sa dérivée est la fonction $t \mapsto P'\big(x(t)\big)x'(t)$.

Théorème 37 (Inégalité des accroissements finis). Soit $P \in \mathbb{C}[X]$ et soient $a, b \in \mathbb{C}$. On suppose que $|P'| \leqslant M$ sur le segment $[a, b]$, où M désigne un réel positif. Alors $|P(b) - P(a)| \leqslant M\,|b - a|$.

2.1 Applications du théorème de d'Alembert-Gauß

Nous avons démontré au module II.5 le théorème de d'Alembert-Gauß (théorème 29 de la page 275), selon lequel *tout polynôme non constant de* $\mathbb{C}[X]$ *admet au moins une racine*. En vertu du théorème 24 de la page 299, on en déduit immédiatement :

Théorème 38. Le corps $\mathbb{C}$ des nombres complexes est algébriquement clos : tout polynôme non nul de $\mathbb{C}[X]$ est scindé. En particulier, un polynôme de $\mathbb{R}[X]$ est scindé si, et seulement si, toutes ses racines dans $\mathbb{C}$ sont réelles.

Corollaire 39. Les polynômes irréductibles unitaires de $\mathbb{R}[X]$ sont les polynômes du premier degré et les trinômes du second degré $X^2 + pX + q \in \mathbb{R}[X]$ dont le discriminant $p^2 - 4q$ est strictement négatif.

Démonstration. Les polynômes $X^2 + pX + q \in \mathbb{R}[X]$ dont le discriminant $p^2 - 4q$ est strictement négatif n'ont pas de racines et sont donc irréductibles (puisqu'ils sont de degré 2). Soit $P \in \mathbb{R}[X]$ un polynôme irréductible unitaire. En tant que polynôme sur $\mathbb{C}$, il admet une racine $z \in \mathbb{C} \setminus \mathbb{R}$. Comme P est à coefficients réels, il découle des règles de calcul sur les conjugués (section 2.3 du module II.5) que $P(\overline{z}) = \overline{P(z)} = 0$. Puisque P admet les deux racines distinctes z et $\overline{z}$, il est divisible par $(X-z)(X-\overline{z}) = X^2+pX+q$, où $p = -2\,\mathrm{Re}(z)$ et $q = |z|^2$. On a donc $P = X^2 + pX + q$. De plus, les racines de P n'étant pas réelles, $p^2 - 4q < 0$. En fait, $p^2 - 4q = -4(\mathrm{Im}\,z)^2$. ∎

Corollaire 40. Soit $P \in \mathbb{R}[X]$ unitaire tel que $\forall x \in \mathbb{R},\ P(x) > 0$.
Il existe alors $A, B \in \mathbb{R}[X]$ tels que $P = A^2 + B^2$.

Démonstration. Il y a deux preuves possibles. On peut décomposer P sur $\mathbb{R}$. Ses facteurs sont des trinômes du second degré à discriminant négatif, donc de la forme $(X + p/2)^2 + \left(\sqrt{q - p^2/4}\right)^2$.
Or, dans tout anneau commutatif, les produits de sommes de deux carrés sont des sommes de deux carrés en vertu de l'identité : $(a^2 + b^2)(c^2 + d^2) = (ad - bc)^2 + (ac + bd)^2$.
On peut également décomposer P sur $\mathbb{C}$. Comme il est à coefficients réels, pour toute racine a de multiplicité μ, $\overline{a}$ est aussi une racine de multiplicité μ. Comme il n'a aucune racine réelle, on peut choisir une racine dans chaque paire de conjuguées et former deux polynômes $F = \prod(X - a_i)$ et $G = \prod(X - \overline{a_i})$ tels que $P = FG$. Alors $F \in \mathbb{C}[X]$ s'écrit $A + \mathrm{i}B$ avec $A, B \in \mathbb{R}[X]$, et $G = A - \mathrm{i}B$, d'où $P = A^2 + B^2$. ∎

Corollaire 41.
Soit $P \in \mathbb{R}[X]$ tel que $\forall x \in \mathbb{R},\ P(x) \geqslant 0$. Il existe alors $A, B \in \mathbb{R}[X]$ tels que $P = A^2 + B^2$.

Démonstration. On peut évidemment supposer $P \neq 0$. Son coefficient dominant C est strictement positif, puisque $P(x) \sim Cx^d$ lorsque $x \to +\infty$ (avec $d = \deg P$). Quitte à diviser P par $C = \left(\sqrt{C}\right)^2$, on peut supposer P unitaire. Soit a une racine réelle de P, de multiplicité $\mu \in \mathbb{N}^*$: donc $P = (X - a)^\mu Q$, où $Q(a) \neq 0$. Si μ était impair, $P(x)$ serait du signe de $Q(a)(x - a)$ au voisinage de a, donc il changerait de signe. Donc les multiplicités de toutes les racines réelles de P sont paires. Ainsi, $P = Q^2 P_1$, où $Q \in \mathbb{R}[X]$ est scindé et où P_1 est strictement positif sur $\mathbb{R}$. On applique alors à P_1 le corollaire précédent. ∎

Prenons par exemple $P := X^6 + X^4 + X^2 + 1$, qui est unitaire, réel et ne prend que des valeurs strictement positives sur $\mathbb{R}$. On remarque que $(X^2 - 1)P = X^8 - 1$. Les racines de P sont donc les éléments de $\mu_8 \setminus \{\pm 1\}$ (module II.5, exemple de la page 259), donc $\pm\mathrm{i}$ et $\pm\dfrac{\sqrt{2}}{2} \pm \mathrm{i}\dfrac{\sqrt{2}}{2}$.
Prenons $F := (X - \mathrm{i})\left(X - \dfrac{\sqrt{2}}{2} - \mathrm{i}\dfrac{\sqrt{2}}{2}\right)\left(X + \dfrac{\sqrt{2}}{2} + \mathrm{i}\dfrac{\sqrt{2}}{2}\right)$. Alors :

$$F = (X - \mathrm{i})(X^2 - \mathrm{i}) = X^3 - \mathrm{i}X^2 - \mathrm{i}X - 1 \qquad \text{et} \qquad P = (X^3 - 1)^2 + (X^2 + X)^2.$$

2.2 Cyclotomie

Rappelons (module II.5) que l'on note $\mu_n := \{\mathrm{e}^{2k\mathrm{i}\pi/n} \mid k \in [\![0, n-1]\!]\}$ l'ensemble des racines $n^{\text{èmes}}$ de l'unité dans $\mathbb{C}$.

Théorème 42. Soit $n \geqslant 1$. Le polynôme $X^n - 1 \in \mathbb{C}[X]$ est scindé et séparable. Il admet la factorisation suivante :

$$X^n - 1 = \prod_{j \in \mu_n} (X - j) = \prod_{k=0}^{n-1} (X - \mathrm{e}^{2k\mathrm{i}\pi/n}).$$

Démonstration.
Les n éléments de μ_n sont racines de $X^n - 1$, qui est de degré n et unitaire. ∎

Exemple. Notons $P(X) := X^n - 1$. Alors $P'(1) = n$, mais la factorisation donne aussi : $P'(1) = \prod_{j\in\mu_n\setminus\{1\}} (1-j) = \prod_{k=1}^{n-1}(1-\mathrm{e}^{2k\mathrm{i}\pi/n})$ qui se réécrit :

$$\prod_{k=1}^{n-1} \mathrm{e}^{k\mathrm{i}\pi/n} \prod_{k=1}^{n-1}(\mathrm{e}^{-k\mathrm{i}\pi/n} - \mathrm{e}^{k\mathrm{i}\pi/n}) = \mathrm{e}^{(n-1)\mathrm{i}\pi/2} \prod_{k=1}^{n-1}\left(-2\mathrm{i}\sin\frac{k\pi}{n}\right) \cdot$$

On a donc en fin de compte $\prod_{k=1}^{n} \sin\frac{k\pi}{n} = \frac{n}{2^{n-1}}$ ·

Corollaire 43. La décomposition en facteurs irréductibles de X^n-1 sur $\mathbb{R}$ est donnée par les formules suivantes :

si $n = 2p$, $\quad X^n - 1 = (X-1)(X+1)\prod_{k=1}^{p-1}\left(X^2 - 2X\cos\frac{2k\pi}{n} + 1\right)$,

si $n = 2p+1$, $\quad X^n - 1 = (X-1)\prod_{k=1}^{p}\left(X^2 - 2X\cos\frac{2k\pi}{n} + 1\right)$.

Démonstration. Le conjugué de $j_k = \mathrm{e}^{2k\mathrm{i}\pi/n}$ est $\overline{j_k} = j_{n-k}$.
D'autre part, $(X - \mathrm{e}^{\mathrm{i}\theta})(X - \mathrm{e}^{-\mathrm{i}\theta}) = X^2 - 2X\cos\theta + 1$. ■

Exercice 8.

Factoriser $X^5 - 1$ et $X^6 - 1$ sur $\mathbb{R}$.

Solution. D'après le calcul de μ_5 au module II.5 (exercice 10 de la page 259), on a :

$$X^5 - 1 = (X-1)\Big(X^2 + \frac{1-\sqrt{5}}{2}X + 1\Big)\Big(X^2 + \frac{1+\sqrt{5}}{2}X + 1\Big).$$

D'après le calcul de μ_6 : $X^6 - 1 = (X-1)(X+1)\left(X^2 - X + 1\right)\left(X^2 + X + 1\right)$.

Exercice 9.

En développant et en factorisant $(X+1)^n - (X-1)^n$, trouver une formule simple pour $\sum_{k=1}^{n-1} \operatorname{cotan}^2 k\pi/n$.

Solution. On a $(X+1)^n - (X-1)^n = 2\left(\binom{n}{1}X^{n-1} + \binom{n}{3}X^{n-3} + \cdots\right)$ (formule du binôme). Le terme dominant est donc $2nX^{n-1}$. Les racines x sont caractérisées par $\frac{x+1}{x-1} = j \in \mu_n$ (et 1 n'est pas racine), ce sont donc les $(n-1)$ complexes $x = \frac{j+1}{j-1}$ où j parcourt $\mu_n\setminus\{1\}$ (il est clair que $\frac{x+1}{x-1} = 1$ est impossible). Comme $\frac{\mathrm{e}^{\mathrm{i}\theta}+1}{\mathrm{e}^{\mathrm{i}\theta}-1} = -\mathrm{i}\operatorname{cotan}\theta/2$, on a finalement :

$$2\left(\binom{n}{1}X^{n-1} + \binom{n}{3}X^{n-3} + \cdots\right) = 2n\prod_{k=1}^{n-1}\left(X + \mathrm{i}\operatorname{cotan}\frac{k\pi}{n}\right).$$

Les calculs de sommes de Newton donnent alors : $\sum_{k=1}^{n-1} \operatorname{cotan}^2\frac{k\pi}{n} = \frac{(n-1)(n-2)}{3}$ ·

2.2.1 Polynômes cyclotomiques

Après la décomposition en facteurs irréductibles de $X^n - 1$ dans $\mathbb{C}[X]$ et dans $\mathbb{R}[X]$, nous allons étudier la décomposition dans $\mathbb{Q}[X]$. Notons $E_n = \{k \in [\![0, n-1]\!] \mid k \wedge n = 1\}$.
Alors $\mu_n^* = \{\mathrm{e}^{2k\mathrm{i}\pi/n} \mid k \in E_n\}$ est l'ensemble des racines primitives $n^{\text{èmes}}$ de l'unité.

Définition 10. Le $n^{\text{ème}}$ *polynôme cyclotomique* est le polynôme unitaire Φ_n dont les racines sont les racines primitives $n^{\text{èmes}}$ de l'unité :

$$\Phi_n := \prod_{j \in \mu_n^*} (X - j) = \prod_{k \in E_n} (X - \mathrm{e}^{2k\mathrm{i}\pi/n}).$$

Le polynôme Φ_n est donc de degré $\operatorname{card} \mu_n^* = \operatorname{card} E_n = \varphi(n)$ (indicatrice d'Euler). On a évidemment $\Phi_1 = X - 1$. Soit p premier. Tous les entiers de $[\![1, p-1]\!]$ sont premiers avec p, et l'on a donc :

$$\Phi_p = \frac{X^p - 1}{X - 1} = X^{p-1} + \cdots + X + 1.$$

Exercice 10.
Calculer Φ_4, Φ_6 et Φ_8.
Solution. On a :

- $\mu_4^* = \{\mathrm{i}, -\mathrm{i}\}$, d'où $\Phi_4 = X^2 + 1$;
- $\mu_6^* = \left\{\frac{1}{2} + \mathrm{i}\frac{\sqrt{3}}{2}, \frac{1}{2} - \mathrm{i}\frac{\sqrt{3}}{2}\right\}$, d'où $\Phi_6 = X^2 - X + 1$;
- $\mu_8^* = \left\{\frac{\sqrt{2}}{2} + \mathrm{i}\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} - \mathrm{i}\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} + \mathrm{i}\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} - \mathrm{i}\frac{\sqrt{2}}{2}\right\}$, d'où :

$$\Phi_8 = \left(\left(X - \frac{\sqrt{2}}{2}\right)^2 + \frac{1}{2}\right)\left(\left(X + \frac{\sqrt{2}}{2}\right)^2 + \frac{1}{2}\right) = X^4 + 1.$$

Proposition 44. Soit $n \geqslant 1$. Alors : $X^n - 1 = \prod_{d \mid n} \Phi_d$.

Démonstration. Soit j une racine $n^{\text{ème}}$ de l'unité et soit d son ordre en tant qu'élément du groupe μ_n : c'est l'unique diviseur de n tel que j est racine primitive $d^{\text{ème}}$. Le groupe μ_n est l'union disjointe des μ_d^* pour $d \mid n$, et l'on a :

$$X^n - 1 = \prod_{j \in \mu_n} (X - j) = \prod_{d \mid n} \Big(\prod_{j \in \mu_n^*} (X - j) \Big) = \prod_{d \mid n} \Phi_d. \qquad \blacksquare$$

Soient par exemple p premier et $k \geqslant 1$. Alors $\Phi_{p^k} = \dfrac{X^{p^k} - 1}{X^{p^{k-1}} - 1} = \sum_{i=0}^{p-1} X^{ip^{k-1}}$. Plus généralement, on peut déduire de la proposition une expression explicite de Φ_n, grâce à la formule d'inversion de Möbius (module II.2, exercice II.2.29 de la page 158). Anticipant un peu sur la fin de

ce module, nous considérons le groupe multiplicatif $\mathbb{C}(X)^*$ du corps des fractions rationnelles. De la formule de Möbius on déduit :

$$\Phi_n = \prod_{d|n} (X^d - 1)^{\mu(n/d)}.$$

Cette formule permet un nouveau calcul de Φ_4, Φ_6 et Φ_8.
On a $\mu(1) = \mu(6) = 1$, $\mu(2) = \mu(3) = -1$ et $\mu(4) = \mu(8) = 0$, d'où :

$$\Phi_4 = \frac{X^4 - 1}{X^2 - 1} = X^2 + 1,$$
$$\Phi_6 = \frac{(X^6 - 1)(X - 1)}{(X^2 - 1)(X^3 - 1)} = X^2 - X + 1,$$
$$\Phi_8 = \frac{X^8 - 1}{X^4 - 1} = X^4 + 1.$$

La formule ne met en jeu que des polynômes à coefficients rationnels, et l'on en déduit donc *a priori* que $\Phi_n \in \mathbb{Q}(X)$. On va démontrer qu'en fait, $\Phi_n \in \mathbb{Z}[X]$.

Lemme 45. Soient $A, B \in \mathbb{Q}[X]$ des polynômes à coefficients rationnels et $C := AB$ leur produit. On suppose que A est unitaire et que A et C sont à coefficients entiers : $A, C \in \mathbb{Z}[X]$. Alors B est à coefficients entiers.

Démonstration. C'est un cas particulier du fait suivant : si l'on effectue la division euclidienne de $C \in \mathbb{Z}[X]$ par un polynôme unitaire $A \in \mathbb{Z}[X]$, le quotient est dans $\mathbb{Z}[X]$. En effet, dans l'algorithme de la division euclidienne, les seules divisions sont par le coefficient dominant de A. ■

Théorème 46. Les polynômes cyclotomiques sont à coefficients entiers : $\Phi_n \in \mathbb{Z}[X]$.

Démonstration. On le démontre par récurrence forte sur n. Pour $n = 1$, c'est acquis. Supposons $n > 1$. On applique alors le lemme à $B := \Phi_n$ et $C := X^n - 1$: le polynôme A est le produit des Φ_d pour tous les d diviseurs stricts de n ; il est évidemment unitaire, et il est à coefficients entiers par hypothèse de récurrence. ■

2.3 Polynômes de Tchebychef

On a remarqué au module II.5 (exemple de la page 267) que les $\cos n\theta$ et $\dfrac{\sin n\theta}{\sin\theta}$ s'expriment comme des polynômes en $\cos\theta$. L'étude de ces polynômes est l'objet de cette section. Nous considérerons d'ailleurs l'expression $\dfrac{\sin n\theta}{\sin\theta}$ comme définie pour tout θ ; pour $\theta = k\pi$, le prolongement par continuité permet de lui attribuer la valeur :

$$\lim_{\alpha\to 0} \frac{\sin n(k\pi + \alpha)}{\sin(k\pi + \alpha)} = \lim_{\alpha\to 0} \frac{(-1)^{nk} \sin n\alpha}{(-1)^k \sin\alpha} = (-1)^{(n-1)k} n.$$

Définition 11. Les *polynômes de Tchebychef de première espèce* sont les polynômes T_n définis par la relation de récurrence à deux pas :

$$T_0 := 1\ ,\ T_1 := X \qquad \text{et} \qquad \forall n \geqslant 1, T_{n+1} := 2XT_n - T_{n-1}.$$

Les *polynômes de Tchebychef de deuxième espèce* sont les polynômes U_n définis par la relation de récurrence à deux pas :

$$U_0 := 1\ ,\ U_1 := 2X \qquad \text{et} \qquad \forall n \geqslant 1, U_{n+1} := 2XU_n - U_{n-1}.$$

Proposition 47. Soit $n \in \mathbb{N}$. Le polynôme T_n est l'unique polynôme tel que :

$$\forall \theta \in \mathbb{R}\ ,\ T_n(\cos\theta) = \cos n\theta.$$

Le polynôme U_n est l'unique polynôme tel que :

$$\forall \theta \in \mathbb{R} \setminus \pi\mathbb{Z}\ ,\ U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta}\ .$$

Démonstration. Ces égalités sont évidemment vérifiées pour $n = 0$ et 1. On prouve qu'elles le sont pour tout $n \geqslant 2$ par une récurrence à deux pas, à l'aide des identités :

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta$$

$$\frac{\sin(n+1)\theta}{\sin\theta} + \frac{\sin(n-1)\theta}{\sin\theta} = 2\cos\theta\frac{\sin n\theta}{\sin\theta}\ .$$

L'unicité vient de ce que les valeurs de T_n et U_n sont ainsi imposées en tout $\cos\theta$, donc sur $[-1,1]$ (ou $]-1,1[$ pour U_n), c'est-à-dire en une infinité de points. ∎

Corollaire 48. (i) Les polynômes T_n et U_n ont même parité que n.
(ii) Pour $n \geqslant 1$, $T_n' = nU_{n-1}$.
(iii) Pour tout n, $XU_n + (X^2-1)U_n' = (n+1)T_{n+1}$.

Démonstration. Toutes les relations à démontrer sont des égalités entre polynômes ; on peut donc se restreindre à $]-1,1[$, qui est infini. Il suffit d'évaluer ces égalités en $\cos\theta$. L'assertion (i) s'obtient alors en remplaçant θ par $\theta+\pi$, l'assertion (ii) (resp. l'assertion (iii)), en dérivant par rapport à θ l'égalité $\cos n\theta = T_n(\cos\theta)$ (resp. l'égalité $\sin(n+1)\theta = \sin\theta\, U_n(\cos\theta)$). ∎

Théorème 49. Les polynômes de Tchebychef sont scindés et séparables. Ils admettent les décompositions :

$$\forall n \geqslant 1\ ,\ T_n = 2^{n-1}\prod_{k=1}^{n}\left(X - \cos\frac{(2k-1)\pi}{2n}\right),$$

$$\forall n \geqslant 1\ ,\ U_n = 2^{n}\prod_{k=1}^{n}\left(X - \cos\frac{k\pi}{n+1}\right).$$

Démonstration. On vérifie d'abord par récurrence que le terme dominant de T_n est,

pour $n \geqslant 1$, $2^{n-1}X^n$; et celui de U_n est, pour $n \geqslant 0$, $2^n X^n$. Par ailleurs, on a :

$$T_n\left(\cos\frac{(2k-1)\pi}{2n}\right) = \cos\frac{(2k-1)\pi}{2} = 0 \quad \text{et} \quad U_n\left(\cos\frac{k\pi}{n+1}\right) = \frac{\sin k\pi}{\sin\frac{k\pi}{n+1}} = 0.$$

La fonction $\cos$ étant strictement décroissante sur $[0,\pi]$, les $\cos\frac{2k-1}{2n}\pi$ (resp. les $\cos\frac{k}{n+1}\pi$) sont deux à deux distincts pour $k \in [\![1,n]\!]$, et l'on a donc n racines distinctes de T_n (resp. de U_n), qui est de degré n : ce sont donc toutes les racines, et elles sont simples. ■

2.4 Nombres algébriques

Proposition et définition 50. Un complexe x est dit *algébrique sur* $\mathbb{Q}$ s'il existe un polynôme non nul $P \in \mathbb{Q}[X]$ tel que $P(x) = 0$. Les nombres complexes algébriques sur $\mathbb{Q}$ sont appelés *nombres algébriques*, les complexes non algébriques sont appelés *nombres transcendants*. L'ensemble des nombres algébrique est noté $\overline{\mathbb{Q}}$. Soit x un nombre algébrique. Le noyau de l'application $P \mapsto P(x)$ de $\mathbb{Q}[X]$ dans $\mathbb{C}$ est un idéal non trivial de $\mathbb{Q}[X]$. Son unique générateur unitaire est appelé *polynôme minimal de* x. C'est un polynôme irréductible dans $\mathbb{Q}[X]$ et scindé séparable dans $\mathbb{C}[X]$. Son degré est appelé *degré de* x.

Démonstration. Le noyau d'un morphisme d'anneaux est un idéal. Il est ici non trivial par définition d'un nombre algébrique. Soit P son générateur unitaire. Si $P = QR$ et une décomposition dans $\mathbb{Q}[X]$, alors $Q(x)R(x) = 0$, donc Q ou R est dans l'idéal, donc multiple de P. Celui-ci est donc bien irréductible. Il est évidemment scindé dans $\mathbb{C}[X]$. Le radical de P est dans $\mathbb{Q}[X]$, il divise P et x en est racine : il est donc associé à P, qui est donc séparable. ■

Exemples.

1. Les nombres algébriques de degré 1 sont les rationnels : tout rationnel x est racine de $X - x \in \mathbb{Q}[X]$ et l'unique racine de $aX + b \in \mathbb{Q}[X]$ est $-\dfrac{b}{a} \in \mathbb{Q}$.
2. Les nombres algébriques de degré 2 sont les nombres de la forme $a+b\sqrt{D}$, où $a, b, D \in \mathbb{Q}$, D n'est pas un carré de $\mathbb{Q}$ et $b \neq 0$ (si $D < 0$, la notation abusive $\sqrt{D}$ représente ici $\mathrm{i}\sqrt{-D}$). Soit en effet $x = a+b\sqrt{D}$. Alors x est racine de $X^2-2aX+a^2-Db^2 \in \mathbb{Q}[X]$. Comme D n'est pas un carré, x n'est pas rationnel : il est donc de degré 2. Réciproquement, si x est de degré 2, il est racine de X^2+pX+q, donc de la forme $a+b\sqrt{D}$, où $D = p^2-4q$ n'est pas un carré (sinon x serait rationnel).
3. Une racine primitive $n^{\text{ème}}$ de l'unité est algébrique de degré $\varphi(n)$: cela découle de l'irréductibilité de $\Phi_n \in \mathbb{Q}[X]$ (que nous n'avons pas démontrée).

Lemme 51. Soient $x_1, \ldots, x_n \in \mathbb{C}$. On suppose que toutes les sommes de Newton $S_k(x_1, \ldots, x_n)$ sont rationnelles : $\forall k \in \mathbb{N}$, $x_1^k + \cdots + x_n^k \in \mathbb{Q}$.
Alors $(X - x_1)\cdots(X - x_n) \in \mathbb{Q}[X]$ et les x_i sont algébriques.

Démonstration. En partant de $a_n = 1$, les relations de récurrence descendante (6) de la page 300 montrent que tous les coefficients de $(X - x_1)\cdots(X - x_n)$ sont rationnels. ■

Lemme 52. Soient $x_1, \ldots, x_m, y_1, \ldots, y_n$ des complexes. On note $p = mn$ et l'on réindexe la famille des $x_i + y_j$ en $u_1, \ldots, u_p$ et la famille des $x_i y_j$ en $v_1, \ldots, v_p$.
On a alors, pour tout $k \in \mathbb{N}$:

$$S_k(v_1, \ldots, v_p) = S_k(x_1, \ldots, x_m) S_k(y_1, \ldots, y_n),$$

$$S_k(u_1, \ldots, u_p) = \sum_{\ell=0}^{k} \binom{k}{\ell} S_\ell(x_1, \ldots, x_m) S_{k-\ell}(y_1, \ldots, y_n).$$

Démonstration. D'abord :

$$S_k(v_1, \ldots, v_p) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i^k y_j^k = \left(\sum_{i=1}^{m} x_i^k \right) \left(\sum_{j=1}^{n} y_j^k \right) = S_k(x_1, \ldots, x_m) S_k(y_1, \ldots, y_n).$$

Ensuite, $S_k(u_1, \ldots, u_p) = \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} (x_i + y_j)^k$ se développe en :

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{\ell=0}^{k} \binom{k}{\ell} x_i^\ell y_j^{k-\ell} = \sum_{\ell=0}^{k} \binom{k}{\ell} \sum_{i=1}^{m} \sum_{j=1}^{n} x_i^\ell y_j^{k-\ell} = \sum_{\ell=0}^{k} \binom{k}{\ell} S_\ell(x_1, \ldots, x_m) S_{k-\ell}(y_1, \ldots, y_n).$$

■

Théorème 53. L'ensemble $\overline{\mathbb{Q}}$ des nombres algébriques est un sous-corps de $\mathbb{C}$.

Démonstration. Il contient 1 et c'est un sous-anneau d'après les lemmes 51 et 52. Si x est algébrique non nul, il est racine de $a_0 + \cdots + a_n X^n$ avec $a_0 a_n \neq 0$, et $\dfrac{1}{x}$ est alors racine de $a_n + \cdots + a_0 X^n$, donc algébrique. ■

C'est un cas particulier d'un théorème valable en toute généralité, mais cette démonstration repose sur le lemme 51 de la page ci-contre qui n'est valable qu'en caractéristique 0. On peut démontrer (et nous ne le ferons pas ici) que le corps $\overline{\mathbb{Q}}$ est algébriquement clos.

Corollaire 54. Le complexe $z \in \mathbb{C}$ est algébrique si, et seulement si, $\operatorname{Re} z$ et $\operatorname{Im} z$ le sont.

Démonstration.
Il est en effet évident que z est algébrique si, et seulement si, son conjugué $\overline{z}$ l'est. ■

Théorème 55. Le corps $\overline{\mathbb{Q}}$ est dénombrable. En conséquence, presque tous les nombres complexes et presque tous les nombres réels sont transcendants.

Démonstration. Tout nombre algébrique est racine d'un polynôme non nul à coefficients entiers, obtenu, par exemple, en réduisant au même dénominateur les coefficients du polynôme minimal. Pour tout entier n, notons E_n l'ensemble des racines des polynômes non nuls de $\mathbb{Z}[X]$ de degré inférieur ou égal à n et dont la somme des valeurs absolues des coefficients est inférieur ou égal à n. Ces polynômes sont en nombre fini, donc E_n est fini. De plus, $\overline{\mathbb{Q}} = \bigcup\limits_{n \geqslant 0} E_n$, qui est dénombrable.

Comme $\mathbb{R}$ et $\mathbb{C}$ ne sont pas dénombrables, l'ensemble des nombres réels transcendants et l'ensemble des nombres complexes transcendants sont non dénombrables. ■

Cet argument ne permet cependant pas de trouver des nombres transcendants, ni de prouver qu'un nombre donné l'est. Les nombres π et e sont transcendants, mais la preuve en est difficile. Nous exhiberons un nombre transcendant moins « naturel » , mais avec une preuve élémentaire.

Lemme 56. Soit x un nombre algébrique réel de degré $d \geqslant 2$.
Il existe alors une constante $C > 0$ telle que :

$$\forall \frac{p}{q} \in \mathbb{Q}\ ,\ \left|x - \frac{p}{q}\right| \geqslant \frac{C}{q^d}\ .$$

Démonstration. Soit P le polynôme obtenu en chassant les dénominateurs des coefficients du polynôme minimal de x : il est donc à coefficients entiers, de degré $d \geqslant 2$ et irréductible dans $\mathbb{Q}[X]$. Par conséquent, $P\left(\frac{p}{q}\right)$ est de la forme $\frac{a}{q^d}$, où a est un entier non nul (sinon P aurait une racine rationnelle, donc un facteur de degré 1, donc il serait réductible ou de degré 1). Ainsi, $\left|P(x) - P\left(\frac{p}{q}\right)\right| \geqslant \frac{1}{q^d}$ · Soit $A > 0$ un majorant de $|P'|$ sur $[x-1, x+1]$.

D'après l'inégalité des accroissements finis (pour les fonctions dérivables sur $\mathbb{R}$), si $\left|x - \frac{p}{q}\right| \leqslant 1$,

$$\left|P(x) - P\left(\frac{p}{q}\right)\right| \leqslant A\left|x - \frac{p}{q}\right| \quad \text{et} \quad \left|x - \frac{p}{q}\right| \geqslant \frac{1}{Aq^d}\ .$$

On peut donc prendre, dans tous les cas, $C = \min\left(1, \frac{1}{A}\right)$. ■

Ce lemme justifie la définition suivante.

Définition 12. Un *nombre de Liouville* est un réel x tel que :

$$\forall d \geqslant 2\ ,\ \forall C > 0\ ,\ \exists (p,q) \in \mathbb{Z} \times \mathbb{N}^*\ :\ 0 < \left|x - \frac{p}{q}\right| < \frac{C}{q^d}.$$

Il résulte donc du lemme que *tout nombre de Liouville est transcendant*.

Exemple. Le nombre $x := \sum\limits_{n \geqslant 0} 10^{-n!}$ est un nombre de Liouville, donc transcendant.

Soit en effet $\frac{p_n}{q_n}$ la forme irréductible de $\sum\limits_{k=0}^{n} 10^{-k!}$: donc $p_n = \sum\limits_{k=0}^{n} 10^{n!-k!}$ et $q_n = 10^{n!}$ (on a bien $p_n \wedge q_n = 1$ car $p_n \equiv 1 \pmod{10}$).
De plus, $x - \frac{p_n}{q_n} = \sum\limits_{k \geqslant n+1} 10^{-k!}$, qui est non nul et vérifie :

$$\left|x - \frac{p_n}{q_n}\right| < \sum_{k \geqslant n+1} 10^{-k!} \leqslant \sum_{k \geqslant (n+1)!} 10^{-k} \leqslant 10^{-(n+1)!} \sum_{k \geqslant 0} 10^{-k} < \frac{2}{10^{(n+1)!}}.$$

On a donc $0 < \left|x - \frac{p_n}{q_n}\right| < \frac{2}{q_n^{n+1}}$, d'où on tire facilement la conclusion.

3 Fractions et fonctions rationnelles

3.1 Le corps des fractions rationnelles

Définition 13. Le corps des fractions (voir module II.2, théorème 56 de la page 153) de l'anneau intègre $K[X]$ est noté $K(X)$ et appelé *corps des fractions rationnelles sur le corps* K. Les éléments de $K(X)$ sont appelés *fractions rationnelles sur* K.

Un élément de $K(X)$ est donc une fraction $\dfrac{A}{B}$, où $A, B \in K[X]$ et où $B \neq 0$. Les fractions rationnelles $\dfrac{A_1}{B_1}$ et $\dfrac{A_2}{B_2}$ sont égales si, et seulement si, $A_1B_2 = A_2B_1$. Chacune des écritures particulières $\dfrac{A}{B}$ d'une fraction rationnelle $R \in K(X)$ peut donc être considérée comme un *représentant* particulier de R. Si B divise A, on identifie la fraction rationnelle $\dfrac{A}{B}$ au quotient de A par B dans l'anneau $K[X]$. Les opérations dans le corps $K(X)$ sont définies par les formules suivantes :

$$\frac{A_1}{B_1} + \frac{A_2}{B_2} := \frac{A_1B_2 + A_2B_1}{B_1B_2},$$
$$\frac{A_1}{B_1}\frac{A_2}{B_2} := \frac{A_1A_2}{B_1B_2},$$
$$\text{si } A \neq 0\ ,\ \left(\frac{A}{B}\right)^{-1} := \frac{B}{A}\ \cdot$$

Pour tout polynôme non nul P, on appelle *dérivée logarithmique de* P la fraction rationnelle $\dfrac{P'}{P}\ \cdot$ De la règle de Leibniz, on déduit que *la dérivée logarithmique de* PQ *est égale à la somme des dérivées logarithmiques de* P *et de* Q. Par exemple, pour un polynôme scindé, on trouve :

$$P = C\prod_{i=1}^{n}(X - a_i)^{\mu_i} \Longrightarrow \frac{P'}{P} = \sum_{i=1}^{n}\frac{\mu_i}{X - a_i}\ \cdot$$

Exercice 11.

Montrer que les $\dfrac{1}{X-a}$ $(a \in K)$ et que les $\dfrac{1}{X^n}$ $(n \in \mathbb{N})$ forment deux familles libres du K-espace vectoriel $K(X)$. Qu'en déduire au sujet de la dimension de celui-ci ?

Solution. Soient $a_1, \ldots, a_n \in K$ deux à deux distincts et $\lambda_1, \ldots, \lambda_n \in K$ tels que $\sum\limits_{i=1}^{n}\lambda_i\dfrac{1}{X-a_i} = 0$. En multipliant cette égalité par $\prod\limits_{i=1}^{n}(X - a_i)$ on obtient l'égalité $\sum\limits_{i=1}^{n}\lambda_iL_i = 0$, où les L_i sont, à un facteur constant près, les polynômes de Lagrange. En évaluant cette égalité en un a_i particulier, on trouve $\lambda_i = 0$: la famille est bien libre.

La dimension de $K(X)$ est donc au moins égale au cardinal de K. Si la famille des $\dfrac{1}{X^n}$ est liée, il existe un $n \in \mathbb{N}$ tel que les $\dfrac{1}{X^k}$ $(0 \leqslant k \leqslant n)$ sont liés. Soient $\lambda_0, \ldots, \lambda_n \in K$ tels que $\sum\limits_{i=0}^{n} \lambda_i \dfrac{1}{X^i} = 0$. En multipliant par X^n, on obtient l'égalité $\sum\limits_{i=0}^{n} \lambda_{n-i} X^i = 0$, d'où $\lambda_0, \ldots, \lambda_n = 0$: la famille est bien libre. La dimension de $K(X)$ est donc infinie.

Proposition et définition 57. Soit $F \in K(X)$. La valeur de $\deg A - \deg B$ est la même pour toute écriture $F = \dfrac{A}{B}$.
On l'appelle *degré* ou *degré total* de F et on le note $\deg F$.

Démonstration. Soient $\dfrac{A_1}{B_1}, \dfrac{A_2}{B_2}$ deux écritures de F. On a $A_1 B_2 = A_2 B_1$, d'où $\deg A_1 + \deg B_2 = \deg A_2 + \deg B_1$, d'où $\deg A_1 - \deg B_1 = \deg A_2 - \deg B_2$ car on peut soustraire $\deg B_1 + \deg B_2 \neq -\infty$. ∎

Proposition 58. Soient F et G deux fractions rationnelles. On a alors :

$$\deg(F+G) \leqslant \max(\deg F, \deg G), \qquad \text{et} \qquad \deg(FG) = \deg(F) + \deg G.$$

De plus, si $\deg F \neq \deg G$, alors $\deg(F+G) = \max(\deg F, \deg G)$.

Démonstration. Écrivons $F = \dfrac{A}{B}$ et $G = \dfrac{C}{D}$.
L'inégalité à démontrer résulte de la relation $\deg(AD+BC) \leqslant \max\big(\deg(AD), \deg(BC)\big)$, dont on peut soustraire $\deg(BD) = \deg B + \deg D \in \mathbb{N}$; le cas d'égalité s'obtient alors en remarquant que, si $\deg F \neq \deg G$, alors $\deg(AD) \neq \deg(BC)$. On obtient $\deg(FG)$ par soustraction des égalités $\deg(AC) = \deg A + \deg C$ et $\deg(BD) = \deg B + \deg D \in \mathbb{N}$. ∎

Proposition et définition 59. Toute fraction rationnelle $F \in K[X]$ admet un unique représentant $\dfrac{A_0}{B_0}$ tel que A_0 et B_0 sont premiers entre eux et B_0 est unitaire. Tous les représentants $\dfrac{A}{B}$ sont alors de la forme $\dfrac{PA_0}{PB_0}$, où $P \in K[X]$ est non nul. Les fractions $\dfrac{A}{B} := \dfrac{PA_0}{PB_0}$ telles que A et B sont premiers entre eux sont celles telles que $P \in K^*$: on les appelle *formes irréductibles de* F.
Le représentant $\dfrac{A_0}{B_0}$ est parfois appelé *la* forme irréductible de F.

Démonstration. Soit $\dfrac{A_1}{B_1}$ un représentant arbitraire de F. Parmi les pgcd de A et B, no-

tons Δ celui qui a même coefficient dominant que B. Alors $A_0 = \dfrac{A_1}{\Delta}$ et $B_0 = \dfrac{B_1}{\Delta}$ sont premiers entre eux, B_0 est unitaire et $\dfrac{A_0}{B_0} = F$. Soit maintenant $\dfrac{A}{B}$ un représentant quelconque de F. De l'égalité $AB_0 = A_0B$, on déduit que B_0 divise A_0B, donc il divise B d'après le lemme de Gauß : il existe donc un unique $P \in K[X]$ non nul tel que $B = PB_0$, ce qui entraîne $A = PA_0$. Les affirmations restantes (y compris l'unicité de *la* forme irréductible de F) se déduisent alors immédiatement. ∎

Corollaire 60. Soient $P \in K[X]$ un polynôme non constant et $F \in K(X)$.
Si $P(F) = 0$, alors F est constante.

Démonstration. Notons $P := a_0 + \cdots + a_nX^n$. Soit $\dfrac{A}{B}$ la forme irréductible de F. L'égalité $P(F) = 0$ entraîne $a_0B^n + a_1B^{n-1}A + \cdots + a_nA^n = 0$, donc $B \mid A^n$, donc $B \in K^*$, puisque A et B sont premiers entre eux (lemme de Gauß), donc $B = 1$, et $a_0 + a_1A + \cdots + a_nA^n = 0$, d'où $\deg A = 0$. ∎

Soit L un corps commutatif contenant K et soit $\dfrac{A_0}{B_0}$ une forme irréductible de $F \in K(X)$. C'est d'ailleurs aussi une forme irréductible de $F \in L(X)$ en vertu du corollaire 12 page 292. Pour tout élément $\alpha \in L$, on dit que F *est définie en* α, ou que α *est substituable dans* F, si $B_0(\alpha) \neq 0$. On pose alors $F(\alpha) := \dfrac{A_0(\alpha)}{B_0(\alpha)} \in L$, que l'on appelle *valeur de* F *en* α.

Ces définitions ne dépendent pas du choix d'une forme irréductible : si $\dfrac{A_1}{B_1}$ est une autre forme irréductible de F, les conditions $B_0(\alpha) \neq 0$ et $B_1(\alpha) \neq 0$ sont équivalentes et $\dfrac{A_0(\alpha)}{B_0(\alpha)} = \dfrac{A_1(\alpha)}{B_1(\alpha)}$.
De plus, pour toute écriture $\dfrac{A}{B}$ de F telle que $B(\alpha) \neq 0$, on a encore $F(\alpha) = \dfrac{A(\alpha)}{B(\alpha)}$.
Réciproquement, si F n'est pas définie en α, alors $\dfrac{A(\alpha)}{B(\alpha)}$ n'est définie pour *aucune* écriture $\dfrac{A}{B}$ de F.

Exemples.

1. La fraction rationnelle $\dfrac{X}{X^2+X} \in K(X)$ est définie en 0, mais cette écriture particulière ne convient pas pour en calculer la valeur, alors que l'écriture $\dfrac{X-1}{X^2-1}$ de la même fraction rationnelle convient.
2. La fraction rationnelle $\dfrac{1}{X^2-2} \in \mathbb{Q}[X]$ est définie en tout point de $\mathbb{Q}$ mais pas en tout point de $\mathbb{R}$. La fraction rationnelle $\dfrac{1}{X^2+1} \in \mathbb{R}[X]$ est définie en tout point de $\mathbb{R}$ mais pas en tout point de $\mathbb{C}$.
3. Si $L = K(X)$, tout G non constant est substituable dans tout $F \in K(X)$, d'après le corollaire 60. On note $F(G)$ ou $F \circ G$ la fraction rationnelle ainsi obtenue.

Proposition et définition 61. Soit F une fraction rationnelle non nulle sur K.
(i) On appelle *pôle de F dans L* un élément $\alpha \in L$ en lequel F n'est pas définie. Pour toute forme irréductible $\dfrac{A_0}{B_0}$ de F, la multiplicité de α en tant que racine de B_0 est la même : on l'appelle *l'ordre du pôle α*.
(ii) On appelle *zéro de F dans L* un élément $\alpha \in L$ en lequel $F(\alpha) = 0$. Pour toute écriture $\dfrac{A}{B}$ de F, α est alors une racine de A. Pour toute forme irréductible $\dfrac{A_0}{B_0}$ de F, la multiplicité de α en tant que racine de B_0 est la même : on l'appelle *l'ordre du zéro α*.
(iii) Il existe un unique $m \in \mathbb{Z}$ tel que $F = (X - \alpha)^m G$, où G est définie en α et $G(\alpha) \neq 0$: si $m > 0$, α est un zéro d'ordre m de F ; si $m < 0$, c'en est un pôle d'ordre $-m$; si $m = 0$, α n'est ni un zéro ni un pôle. L'entier m est appelé *l'ordre de F en α*.

Démonstration. C'est immédiat d'après ce qui précède. ∎

3.2 Propriétés arithmétiques de $K(X)$

Ces propriétés reposent sur le fait que $K(X)$ est le corps des fractions de $K[X]$, qui possède une division euclidienne et qui est principal. Elles sont donc très proches de celles de $\mathbb{Q}$.

Théorème 62. Toute fraction rationnelle non nulle F peut s'écrire de manière unique :

$$F = C P_1^{\mu_1} \cdots P_r^{\mu_r},$$

où $C \in K^*$, où $P_1, \ldots, P_r$ sont irréductibles unitaires et où $\mu_1, \ldots, \mu_r$ sont des entiers relatifs non nuls.

Démonstration. Soit $\dfrac{A_0}{B_0}$ la forme irréductible de F. Alors la factorisation ci-dessus s'obtient à partir des factorisations $A_0 = C \prod_{j=1}^{p} Q_j^{\nu_j}$ et $B_0 = \prod_{k=1}^{q} R_k^{\xi_k}$ dans $K[X]$ en posant $r := p + q$, et :

$$\begin{cases} (P_i, \mu_i) := (Q_i, \nu_i) \text{ pour } i = 1, \ldots, p, \\ (P_i, \mu_i) := (R_{i-p}, -\xi_{i-p}) \text{ pour } i = p+1, \ldots, p+q. \end{cases}$$ ∎

Corollaire 63. Si m est l'ordre de F en $\alpha \in K$, le polynôme irréductible unitaire $X - \alpha$ apparaît dans cette factorisation avec l'exposant m.

Exemple. Pour que F soit un carré dans $K(X)$, il faut, et il suffit, que C soit un carré dans K^* et que tous les μ_i soient pairs. Par exemple, X n'est pas un carré dans $K(X)$.

Proposition et définition 64. Toute fraction rationnelle $F \in K(X)$ admet une unique décomposition $F = Q + G$, où $Q \in K[X]$ et $\deg G < 0$. Le polynôme Q est appelé *partie entière de* F.

Démonstration. Soit $\dfrac{A_0}{B_0}$ la forme irréductible de F. Alors $R := A_0 - QB_0$ est un polynôme vérifiant $\deg R < \deg B_0$, puisque $G = \dfrac{R}{B_0}$ est de degré strictement négatif.

Ainsi, $Q = A_0 \text{ div } B_0$ et $G = \dfrac{R}{B_0}$, où $R = A_0 \mod B_0$. ∎

Pour tout $n \in \mathbb{Z}$, notons $K_n(X)$ le sous–espace vectoriel de $K(X)$ formé des fractions rationnelles de degré inférieur ou égal à n. On voit donc que $K(X) = K[X] \oplus K_{-1}(X)$. Nous allons maintenant décrire les éléments de $K_{-1}(X)$.

Lemme 65. Soit $F := \dfrac{A}{B}$ une fraction rationnelle non nulle de degré strictement négatif. On suppose que $B = \prod_{i=1}^{k} B_i$, où les B_i sont premiers entre eux deux à deux. Il existe alors une unique décomposition $F = \sum_{i=1}^{k} \dfrac{A_i}{B_i}$ avec $\deg \dfrac{A_i}{B_i} < 0$.
Si $\dfrac{A}{B}$ est irréductible, les $\dfrac{A_i}{B_i}$ le sont.

Démonstration. Si $k = 1$, c'est évident.
Si $k = 2$, l'écriture voulue équivaut à $A = A_1B_2 + A_2B_1$, qui découle du théorème de Bézout. La condition sur les deux degrés équivaut alors à l'une quelconque des deux conditions, par exemple $\deg A_1 < \deg B_1$. Il suffit, pour réaliser celle-ci, de remplacer A_1 par $A_1 - SB_1 = A_1 \mod B_1$ (donc $S = A_1 \text{ div } B_1$) et A_2 par $A_2 + SB_2$. On passe au cas général par récurrence, en remarquant que B_k et $\prod_{i=1}^{k-1} B_i$ sont premiers entre eux. Supposons $\dfrac{A}{B}$ irréductible. Tout diviseur commun de A_i et B_i, divise B et chacun des k facteurs qui composent A, donc il est constant. ∎

Exemple. Si $\deg A < k$, la fraction rationnelle $F := \dfrac{A}{(X - x_1)\cdots(X - x_k)}$ (où les x_i sont deux à deux distincts) admet donc une unique décomposition

$$F = \frac{\lambda_1}{X - x_1} + \cdots + \frac{\lambda_k}{X - x_k}\ .$$

Notant $\omega = (X - x_1)\cdots(X - x_k)$, on voit que cela équivaut à $A = \sum \lambda_i \omega'(x_i) L_i$, où les L_i sont les polynômes de Lagrange $\dfrac{\omega}{\omega'(x_i)(X - x_i)}\cdot$ On a donc :

$$\frac{A}{(X - x_1)\cdots(X - x_k)} = \frac{A(x_1)}{\omega'(x_1)(X - x_1)} + \cdots + \frac{A(x_k)}{\omega'(x_k)(X - x_k)}\ .$$

On peut retrouver ce résultat de manière plus « intuitive » : pour chaque i, on multiplie l'égalité $F = \dfrac{\lambda_1}{X - x_1} + \cdots + \dfrac{\lambda_k}{X - x_k}$ par $X - x_i$ puis l'on fait « tendre X vers x_i ». Le membre droit « tend vers λ_i », le membre gauche vers $\dfrac{A(x_i)}{\omega'(x_i)}$.

Lemme 66. Soit $F := \dfrac{A}{B}$ une fraction rationnelle non nulle de degré strictement négatif. On suppose que $B = Q^m$, où $m \geqslant 1$.
Il existe alors une unique décomposition $F = \displaystyle\sum_{i=1}^{m} \frac{A_i}{Q^i}$ avec $\deg A_i < \deg Q$.

Démonstration. Pour $m = 1$, c'est immédiat.
Si $m \geqslant 2$, l'égalité voulue équivaut à : A_1 est la partie entière de QF et $\displaystyle\sum_{i=1}^{m-1} \frac{A_{i+1}}{Q^i}$ est la décomposition de $QF - A_1 = \dfrac{A - A_1 Q^{m-1}}{Q^{m-1}}$. On conclut par récurrence sur m. ∎

Théorème et définition 67. Soit $\dfrac{A}{B}$ la forme irréductible d'une fraction rationnelle F et soit $B = \displaystyle\prod_{i=1}^{k} F_i^{\mu_i}$ la décomposition de B en facteurs irréductibles unitaires. On a alors une unique décomposition :

$$F = E + \sum_{i=1}^{k} \sum_{j=1}^{\mu_i} \frac{A_{i,j}}{F_i^j},$$

où E et les $A_{i,j}$ sont des polynômes et où $\deg A_{i,j} < \deg F_i$ pour $i \in [\![1, k]\!]$, $j \in [\![1, \mu_i]\!]$.
Cette écriture est la *décomposition en éléments simples de* F.

Démonstration. Elle découle de la proposition et des deux lemmes précédents. On obtient E comme partie entière de F. Puis l'on décompose $F - E$ en une somme de $\dfrac{A_i}{B_i} = \dfrac{A_i}{F_i^{\mu_i}}$; puis chaque $\dfrac{A_i}{F_i^{\mu_i}}$ en une somme $\displaystyle\sum_{j=1}^{\mu_i} \frac{A_{i,j}}{F_i^j}$. ∎

Définition 14. Soit α un pôle de F dans K, de sorte que la décomposition en éléments simples de F comporte un terme de la forme $\dfrac{a}{X - \alpha}$, avec $a \in K$.
La constante a est appelée *résidu de* F *en le pôle* α. Par convention, le résidu de F en un point de K qui n'est pas un pôle est nul.

Si l'on note E la partie entière de $(X - \alpha)F$, on vérifie facilement que résidu de F en α est, dans tous les cas, égal à $E(\alpha)$.

Dans le cas où le dénominateur de F est scindé, on dispose d'une méthode efficace pour déterminer sa décomposition en éléments simples.

Théorème 68 (Division par puissances croissantes).
Soient $P, Q \in K[X]$ tels que $Q(0) = \mathrm{tc}(Q) \neq 0$. Pour tout entier m, il existe un unique polynôme $S \in K_{m-1}[X]$ tel que X^m divise $P - SQ$.

Démonstration. Prouvons d'abord l'unicité. Si deux polynômes S_1, S_2 conviennent, X^m divise $(S_1 - S_2)Q$. Mais l'hypothèse sur Q implique qu'il est premier avec X^m, qui divise donc $S_1 - S_2$, qui est donc nul puisque de degré strictement inférieur à m. L'existence s'obtient par un algorithme analogue à celui de la division euclidienne. On pose $P_0 := P$. Alors $P_1 := P_0 - \dfrac{\mathrm{tc}\, P_0}{\mathrm{tc}\, Q} Q$ a pour terme constant 0, il est donc divisible par X. Supposons déterminé P_i non nul, de terme de plus bas degré $s_i X^{m_i}$. On pose alors $P_{i+1} := P_i - \dfrac{s_i}{\mathrm{tc}\, Q} X^{m_i} Q$, qui est divisible par X^{m_i+1}. Ainsi, $P_{i+1} = 0$ ou $m_{i+1} > m_i$. On arrête au premier indice k tel que $P_{k+1} = 0$ ou $m_{k+1} > m$, et l'on a alors $S = \sum_{i=0}^{k} s_i X^{m_i}$. ■

Appliquons cet algorithme à la décomposition en éléments simples de la fraction rationnelle $F := \dfrac{A}{(X-\alpha)^m B}$, supposée de degré strictement négatif, où $B(\alpha) \neq 0$. On a une unique écriture : $F = \dfrac{C}{B} + \dfrac{a_1}{X-\alpha} + \cdots + \dfrac{a_m}{(X-\alpha)^m}$, où $\deg C < \deg B$ et où $a_1, \ldots, a_m \in K$. En effet, cette égalité équivaut à :

$$A(X+\alpha) = (a_1 X^{m-1} + \cdots + a_m) B(X+\alpha) + X^m C(X+\alpha)$$

(les notations $A(X+\alpha)$, $B(X+\alpha)$ et $C(X+\alpha)$ représentent ici bien entendu des compositions des polynômes A, B, C par le polynôme $X+\alpha$). On détermine les coefficients a_i (donc C) en effectuant la division par puissances croissantes de $P = A(X+\alpha)$ par $Q = B(X+\alpha)$. Celle-ci est possible parce que, par hypothèse, $Q(0) = B(\alpha) \neq 0$. On réitère alors le procédé avec $G = \dfrac{C}{B}$.

Plus généralement, si $F := \dfrac{A}{\prod_{i=1}^{k} (X-\alpha_i)^{\mu_i}}$, supposé de degré strictement négatif, on se débarrasse successivement des facteurs $(X-\alpha_i)^{\mu_i}$ du dénominateur.

Exercice 12.

Décomposer en éléments simples $F := \dfrac{4X^3}{(X^2-1)^2}$.

Solution. On veut avoir : $F = \dfrac{a}{X-1} + \dfrac{b}{(X-1)^2} + \dfrac{c}{X+1} + \dfrac{d}{(X+1)^2}$.

Une division selon les puissances croissantes donne : $4(X+1)^3 = (1+2X)(X+2)^2 +$ un multiple de X^2 ; de même, $4(X-1)^3 = (-1+2X)(X-2)^2 +$ un multiple de X^2.
Finalement : $a = 2, b = 1, c = 2, d = -1$.

3.3 Fonctions rationnelles

Dans cette section, le corps K est supposé infini. Pour toute fraction rationnelle $F \in K(X)$ écrite sous forme irréductible $\dfrac{A}{B}$, l'ensemble des pôles de F dans K, qui est l'ensemble des racines de B dans K, est fini. Son complémentaire dans K est l'ensemble $\mathcal{D}_F$ des $\alpha \in K$ tels que $F(\alpha)$ est défini. On l'appelle *ensemble de définition de* F.

Proposition 69. Soient $F, G \in K(X)$.
On suppose que, pour tout $\alpha \in \mathcal{D}_F \cap \mathcal{D}_G$, $F(\alpha) = G(\alpha)$. Alors $F = G$.

Démonstration. Si $F = \dfrac{A}{B}$ et $G = \dfrac{C}{D}$, on a $(AD - BC)(\alpha) = 0$ pour tout α dans l'ensemble infini $\mathcal{D}_F \cap \mathcal{D}_G$, donc $AD - BC = 0$. ∎

On pourra donc sans inconvénient identifier la fraction rationnelle F avec la *fonction rationnelle* $x \mapsto F(x)$ de $\mathcal{D}_F$ dans K. Cette identification est évidemment compatible avec les combinaisons linéaires, la multiplication et l'inversion des éléments non nuls. La seule petite difficulté est celle-ci : il peut arriver que $F + G$ ou FG soit défini en un point où F ou G n'est pas défini. Par exemple $\dfrac{1}{X} - \dfrac{1}{X}$ et $\dfrac{X-1}{X}\dfrac{X}{X-1}$ sont définies partout.

Proposition et définition 70. Soit $F := \dfrac{A}{B} \in K(X)$.
La fraction rationnelle $\dfrac{A'B - AB'}{B^2}$ ne dépend que de F et non de l'écriture particulière choisie. On l'appelle *dérivée de* F et on la note F'. L'application $F \mapsto F'$ est un endomorphisme du K-espace vectoriel $K(X)$, qui vérifie de plus la règle de Leibniz :

$$\forall F, G \in K(X)\ ,\ (FG)' = F'G + FG'.$$

Démonstration. Soit $\dfrac{A_0}{B_0}$ la forme irréductible de F. Il existe un polynôme P tel que $A = PA_0$ et $B = PB_0$, d'où $A'B - AB' = P^2(A_0'B_0 - A_0B_0')$, et enfin :

$$\frac{A'B - AB'}{B^2} = \frac{A_0'B_0 - A_0B_0'}{B_0^2}\ .$$

Les autres affirmations découlent d'un calcul direct, qui est laissé au lecteur. ∎

Naturellement, dans le cas du corps $\mathbb{R}$, cette définition algébrique coïncide avec celle du cours d'analyse.

Exemple. La dérivée $n^{ème}$ de $\dfrac{1}{(X-\alpha)^m}$ est $\dfrac{(-1)^n m(m+1)\cdots(m+n-1)}{(X-\alpha)^{m+n}}$.
Pour calculer les dérivées successives de F lorsque le corps est algébriquement clos, il suffit donc d'utiliser la décomposition en éléments simples de F.

Proposition 71. On suppose le corps K de caractéristique 0. Si $F \in K(X)$ est non nulle et de degré non nul, on a $\deg F' = (\deg F) - 1$.

Démonstration.
Si $F = \dfrac{A}{B}$, on a $\deg A \neq \deg B$, et les coefficients dominants $(\deg A)(\mathrm{cd}\, A)(\mathrm{cd}\, B)$ et $(\deg B)(\mathrm{cd}\, A)(\mathrm{cd}\, B)$ de $A'B$ et de $B'A$ sont différents. Le numérateur $AB' - BA'$ de F' est donc de degré $\deg A + \deg B - 1$ et :

$$\deg F' = \deg A + \deg B - 1 - 2\deg B = \deg F - 1.$$

Corollaire 72. Si K est de caractéristique 0, l'endomorphisme $F \mapsto F'$ de $K(X)$ a pour noyau K.

Démonstration. On écrit $F = E + G$, avec $E \in K[X]$ et $\deg G < 0$. Alors $F' = E' + G'$, avec $E' \in K[X]$ et $\deg G' < 0$. Si $F' = 0$, on a $E' = G' = 0$, donc $E \in K$ et $G = 0$.

On sait que $K(X) = K[X] \oplus K_{-1}(X)$ et que les X^n forment une base de $K[X]$. Si le corps K est algébriquement clos, il résulte du théorème 67 de la page 320 que $K_{-1}(X)$ admet pour base la famille des $(X-\alpha)^{-m}$, où $\alpha \in K$ et $m \in \mathbb{N}^*$. Plus généralement, pour chaque $n \geqslant 1$, $K_{-n}(X)$ admet pour base la famille des $(X-\alpha)^{-m}$, où $\alpha \in K$ et $m \geqslant n$. L'ensemble des fractions rationnelles dont tous les résidus sont nuls est donc $K[X] \oplus K_{-2}(X)$.

Théorème 73. On suppose K de caractéristique 0 et algébriquement clos. Alors l'endomorphisme $F \mapsto F'$ de $K(X)$ a pour image l'ensemble des fractions rationnelles dont tous les résidus sont nuls.

Démonstration. Notons D cet endomorphisme. On sait que la restriction de D à $K[X]$ est surjective et que l'image de $K_{-1}(X)$ par D est incluse dans $K_{-2}(X)$. Par ailleurs, si $m \geqslant 2$, la fraction rationnelle $\dfrac{1}{(X-\alpha)^m}$ est la dérivée de la fraction rationnelle $-\dfrac{1}{m-1}\dfrac{1}{(X-\alpha)^{m-1}}$.

3.4 Développements limités

Soit $F \in \mathbb{C}(X)$ une fraction rationnelle non nulle, d'ordre m en 0. Alors $F = X^m G$, où G est définie en 0 et $a = G(0) \neq 0$. Ainsi, lorsque le complexe z tend vers 0, selon les notations du module IV.8, $F(z) \sim az^m$. De même, la fonction rationnelle F est définie en 0 si, et seulement si, $m \geqslant 0$. Bien que défini algébriquement, l'ordre a donc dans ce cas une signification analytique.

Dans cette section, on notera $K(X)_0$ le sous-anneau de $K(X)$ formé des fractions rationnelles définies en 0, qui est stable par dérivation. Pour tout $m \geqslant 0$, l'ensemble des fractions rationnelles dont l'ordre en 0 est supérieur ou égal à m est l'idéal de $K(X)_0$ engendré par X^m. On notera $F \equiv G \pmod{X^m}$ la congruence modulo cet idéal, c'est à dire la relation d'équivalence « $F - G$ est d'ordre supérieur ou égal à m en 0 ».

Proposition et définition 74. Soit $F \in K(X)_0$. Pour tout $m \geqslant 0$, il existe un unique $P \in K_m[X]$ tel que $F \equiv P \pmod{X^{m+1}}$; on l'appelle *développement limité de F à l'ordre m* et on le note DL_m de F ou $DL_m(F)$.

Démonstration. Si $\dfrac{A}{B}$ est une forme irréductible de F, le fait que $F \in K(X)_0$ signifie que X ne divise pas B. La proposition est alors une conséquence immédiate de la division de A par B selon les puissances croissantes. ■

Les règles connues en analyse concernant les développements limités s'appliquent alors, avec l'avantage supplémentaire que $F \in K(X)_0$ admet des développements limités à tous les ordres.

1. Le DL_m de F est la troncature à l'ordre m de $DL_{m+1}(F)$.
2. Le DL_m de $\lambda F + \mu G$ est $\lambda DL_m(F) + \mu DL_m(G)$.
3. Le DL_m de FG est la troncature à l'ordre m de $DL_m(F)DL_m(G)$.
4. La dérivée de $DL_m(F)$ est $DL_{m-1}(F')$. Cette règle, qui est fausse en analyse, découle du calcul suivant : si $F = P + X^{m+1}G$, où $G \in K(X)_0$, alors $F' = P' + X^m H$, où $H = (m+1)G + XG' \in K(X)_0$.

Exemple. De l'égalité $\dfrac{1 - X^{m+n}}{1 - X} = 1 + X + \cdots + X^{m+n-1}$, on déduit que :

$$DL_{m+n-1}\left(\frac{1}{1-X}\right) = 1 + X + \cdots + X^{m+n-1},$$

puis, en dérivant $(n-1)$ fois :

$$\frac{1}{(1-X)^n} \equiv \sum_{k=0}^{m} \binom{k+n-1}{k} X^k \pmod{X^{m+1}}.$$

Exercice type corrigé

II.6.0 Nous allons déterminer les coefficients des polynômes de Tchebychef T_n. Une méthode analogue permettrait de calculer ceux des polynômes U_n. Nous fixons $n \in \mathbb{N}^*$ et nous écrivons $T_n = a_0 + a_1 X + \cdots + a_n X^n$.

1) Calculer $a_0 = T_n(0)$ et $a_1 = T_n'(0)$ *(on rappelle que $T_n' = nU_{n-1}$)*.

2) Démontrer l'équation différentielle : $(1 - X^2)T_n'' - XT_n' + n^2 T_n = 0$.

3) En déduire la relation de récurrence : $a_{k+2} = \dfrac{k^2 - n^2}{(k+1)(k+2)} a_k$.

4) Calculer explicitement les a_k (distinguer deux cas selon la parité de n).

Solution.

1) On a $a_0 = T_n(\cos \pi/2) = \cos n\pi/2$, qui est nul si n est impair et vaut $(-1)^p$ si $n = 2p$.

De même, on a $a_1 = T_n'(0) = nU_{n-1}(\cos \pi/2)$, donc $a_1 = n\dfrac{\sin n\pi/2}{\sin \pi/2}$, qui est nul si n est pair et vaut $n(-1)^p$ si $n = 2p+1$.

2) Selon le corollaire 48 de la page 311, on a directement $T'_n = nU_{n-1}$ et (après un décalage d'indice) $XU_{n-1} + (X^2-1)U'_{n-1} = nT_n$. Il suffit donc de remplacer U_{n-1} par $\dfrac{1}{n}T'_n$ dans cette dernière relation.
3) Les coefficients en X^k dans les polynômes $(1-X^2)T''_n$, XT'_n et n^2T_n valent respectivement $(k+1)(k+2)a_{k+2} - k(k-1)a_k$, ka_k et n^2a_k, d'où, d'après la question précédente, l'égalité $(k+1)(k+2)a_{k+2} - k(k-1)a_k - ka_k + n^2a_k = 0$. La relation voulue s'en déduit immédiatement.
4) Si $n = 2p$, les $a_{2\ell+1}$ sont nuls (car $a_1 = 0$), et l'on a :

$$a_{2\ell+2} = -4\frac{p^2-\ell^2}{(2\ell+1)(2\ell+2)}a_{2\ell}.$$

Comme $a_0 = (-1)^p$, on trouve :

$$a_{2\ell} = \frac{(-4)^\ell(-1)^p}{2\ell!}\prod_{j=0}^{\ell-1}(p-j)(p+j) = \frac{(-1)^{\ell+p}2^{2\ell}p}{2\ell!}\frac{(p+\ell-1)!}{(p-\ell)!}\ .$$

Si $n = 2p+1$, les $a_{2\ell}$ sont nuls (car $a_0 = 0$), et l'on a :

$$a_{2\ell+1} = -4\frac{p(p+1)-\ell(\ell-1)}{2\ell(2\ell+1)}a_{2\ell-1}.$$

Comme $a_1 = n(-1)^p$, un calcul similaire donne :

$$a_{2\ell+1} = \frac{(-1)^{\ell+p}2^{2\ell}(2p+1)}{(2\ell+1)!}\frac{(p+\ell)!}{(p-\ell)!}\ .$$

On déduit aisément de ces formules l'égalité $a_n = 2^{n-1}$ (théorème 49 de la page 311). On retrouve de même le fait que T_n a même parité que n (corollaire 48 de la page 311).

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

II.6.1 Soit $a \in K$, et soit φ_a l'automorphisme $P \mapsto P(X+a)$ du K-espace vectoriel $K_n[X]$. Vérifier que $a \mapsto \varphi_a$ est un morphisme du groupe additif K dans le groupe linéaire de $K_n[X]$. Soit M_a la matrice dont les colonnes sont les vecteurs colonnes des coordonnées des images $(X+a)^k$ des éléments de la base canonique de $K_n[X]$ dans cette même base canonique. Écrire la matrice M_a.
Vérifier que $a \mapsto M_a$ est un morphisme du groupe additif K dans le groupe linéaire $GL_{n+1}(K)$.

II.6.2 Vérifier que l'application $\Delta_a : P \mapsto \varphi_a(P) - P$ (voir l'exercice II.6.1 de la page précédente) est une application linéaire de $K_n[X]$ dans $K_{n-1}[X]$. On suppose K de caractéristique nulle. Quelle est l'image par Δ_1 de la base des polynômes de Newton $\binom{X}{n} := \dfrac{X(X-1)\cdots(X-n+1)}{n!}$?
En déduire que cette application est surjective et de noyau K.
Application : Soit $k \in \mathbb{N}$. Montrer l'existence d'un unique polynôme $P_k \in \mathbb{Q}(X)$, de terme dominant $\dfrac{X^{k+1}}{k+1}$, tel que, pour tout $n \in \mathbb{N}^* : 1^k + \cdots + n^k = P_k(n)$.

II.6.3 Que donne la division euclidienne de $X^n - 1$ par $X^m - 1$?
Que donne l'algorithme d'Euclide appliqué à $X^n - 1$ et $X^m - 1$?

II.6.4 $*$ Soient $A, B \in K[X]$ des polynômes non constants de degrés respectifs n, p. Montrer que, pour que A et B ne soient pas étrangers, il faut, et il suffit, qu'il existe des polynômes $P \in K_{p-1}[X]$ et $Q \in K_{n-1}[X]$ non tous deux nuls tels que $AP + BQ = 0$. Vérifier que cette condition est, à son tour, équivalente à la suivante : les éléments $A, XA, \ldots, X^{p-1}A, B, XB, \ldots, X^{q-1}B$ de $K_{n+p-1}[X]$ sont liés. On écrit $A = a_0 + \cdots + a_n X^n$ et $B = b_0 + \cdots + b_p X^p$, et l'on forme la matrice[3] $S(A, B) \in M_{n+p}(K)$ ainsi définie :

$$\begin{pmatrix} a_0 & \ldots & a_n & 0 & \ldots & 0 & 0 & 0 \\ 0 & a_0 & \ldots & a_n & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \ldots & 0 & a_0 & \ldots & a_n \\ b_0 & \ldots & \ldots & b_p & 0 & \ldots & \ldots & 0 \\ 0 & b_0 & \ldots & \ldots & b_p & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & 0 & b_0 & \ldots & \ldots & b_p \end{pmatrix} \quad (p \text{ lignes avec les } a_i,\ n \text{ lignes avec les } b_j).$$

Vérifier que $A \wedge B = 1$ si, et seulement si, $S(A, B)$ est inversible.

II.6.5 Soit $P := X^n + a_{n-1}X^{n-1} + \cdots + a_n \in \mathbb{Z}[X]$ un polynôme unitaire à coefficients entiers et soit $x \in \mathbb{Q}$ une racine rationnelle de P. Démontrer que $x \in \mathbb{Z}$: on dit que l'anneau $\mathbb{Z}$ est *intégralement clos*. (Écrire x sous forme irréductible $\dfrac{p}{q}$, montrer que $q \mid p^n$ et appliquer le lemme de Gauß dans $\mathbb{Z}$.) En déduire que $\sqrt{2}$, $\sqrt{3}$ et $\sqrt{2} + \sqrt{3}$ ne sont pas rationnels. (Pour ce dernier, penser au polynôme $(X^2 - 5)^2 - 24$.)

II.6.6 (i) Soit $P \in \mathbb{C}[X]$. Démontrer que $P(\mathbb{Z}) \subset \mathbb{Q}$ si, et seulement si, $P \in \mathbb{Q}[X]$. (Si $\deg P = n$, interpoler P en $0, \ldots, n$.)
(ii) On prend $P \in \mathbb{Q}[X]$ que l'on décompose dans la base des polynômes de Newton : $P = \displaystyle\sum_{n \geqslant 0} a_n \binom{X}{n}$.
Démontrer que $P(\mathbb{Z}) \subset \mathbb{Z}$ si, et seulement si, les $a_n \in \mathbb{Z}$.
Indication : appliquer P successivement à $0, 1, \ldots$.

[3] C'est la *matrice de Sylvester*, dont le déterminant sert à définir le *résultant* : voir le chapitre sur les polynômes à plusieurs indéterminées du cours de L2.

II.6.7 Soient $x_1, \ldots, x_n \in K$ deux à deux distincts et $L_1, \ldots, L_n \in K_{n-1}[X]$ les polynômes de Lagrange correspondants. On définit la matrice de Vandermonde : $A = (a_{i,j})_{1 \leqslant i,j \leqslant n} \in M_n(K)$ par $a_{i,j} = x_i^{j-1}$ Soit par ailleurs B la matrice dont la $j^{ème}$ colonne est formée des coordonnées de L_j dans la base canonique de $K_{n-1}[X]$. Démontrer que $AB = I_n$.

II.6.8 $*$ Avec les notations de l'exercice précédent, on pose $\omega := \prod_{i=1}^{n}(X - x_i)$, de sorte que $L_i = \dfrac{\omega}{\omega'(x_i)(X - x_i)}$.

Démontrer que, pour tout polynôme $P \in K_{n-1}[X]$, $P = \sum_{i=1}^{n} P(x_i)L_i$.

En déduire, pour tout polynôme $P \in K_{n-2}[X]$, l'égalité $\displaystyle\sum_{i=1}^{n} \frac{P(x_i)}{\omega'(x_i)} = 0$. Que vaut $\displaystyle\sum_{i=1}^{n} \frac{x_i^{n-1}}{\omega'(x_i)}$?

II.6.9 $*$ Soit $\theta \in \mathbb{R}$.
Démontrer que $X^n \sin\theta - X \sin n\theta + \sin(n-1)\theta$ est divisible par $X^2 - 2X\cos\theta + 1$.

II.6.10 Montrer que si l'on suppose K de caractéristique $p > 0$ dans l'exercice II.6.2 de la page ci-contre, et si l'on prend $n = p$, alors le polynôme $X^p - X$ est dans le noyau et le polynôme X^{p-1} n'est pas dans l'image de Δ_1.

II.6.11 $*$ On suppose que K est un anneau commutatif et que $A, B \in K[X]$ et $A \neq 0$. Démontrer qu'il existe $Q, R \in K[X]$ tels que :

$$\big(\mathrm{cd}(A)\big)^k B = QA + R, \text{ où } k := \max(0, \deg B - \deg A + 1).$$

et $\deg R < \deg A$. Dans le cas où $\mathrm{cd}(A)$ est inversible, montrer que l'on retrouve l'existence et l'unicité du quotient et du reste.

II.6.12 Montrer que le polynôme $P \circ (X + P)$ est divisible par le polynôme P. À quelle condition $P \circ (X + P)$ est-il irréductible ?

II.6.13 Sur le corps $\mathbf{H}$ des quaternions, le polynôme $X^2 + 1$ a une infinité de racines. (Les chercher sous la forme $b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$.) Le groupe multiplicatif $\mathbf{H}^*$ a des sous-groupes finis non cycliques. (Considérer le groupe $\{\pm 1, \pm\mathbf{i}, \pm\mathbf{j}, \pm\mathbf{k}\}$.)

II.6.14 $*$ Soit K un corps fini à q éléments.
(i) De la factorisation $X^{q-1} - 1 = \prod_{a \in K^*}(X - a)$, déduire une expression exacte des polynômes de Lagrange L_i correspondant aux $(q-1)$ paires de K^*.
(ii) En prenant $K = \mathbb{Z}/p\mathbb{Z}$ (p premier), déduire le théorème de Wilson : $(p-1)! \equiv -1 \pmod{p}$.

II.6.15 $*$ Dans cet exercice, on fait appel à la terminologie du module V.2.
(i) On tire au hasard et sans remise des entiers dans $[\![1, n]\!]$. Après k tirages, le nombre d'entiers *distincts* qui ont été tirés est D_k. Ainsi $D_0 = 0$ et $D_1 = 1$ avec probabilité 1, et, pour $k \geqslant 2$, D_k prend une valeur dans $[\![1, \min(n, k)]\!]$. Démontrer, pour tout $m \in [\![0, n]\!]$ et pour tout $k \in \mathbb{N}$, la formule :

$$\mathbb{P}(D_{k+1} = m) = \frac{m}{n}\mathbb{P}(D_k = m) + \frac{n - m + 1}{n}\mathbb{P}(D_k = m - 1).$$

(ii) Soit, pour tout $k \in \mathbb{N}$, $P_k := \sum_{m=0}^{n} \mathbb{P}(D_k = m)X^m \in K_n[X]$. Vérifier que $P_k(1) = 1$ et que $P_k'(1)$ est l'espérance de D_k.

(iii) Vérifier que $P_0 = 1$, $P_1 = X$ et, pour $k \in \mathbb{N}$: $P_{k+1} = XP_k + \dfrac{X - X^2}{n} P'_k$.

(iv) Soit $n \in \mathbb{N}^*$ fixé. Pour tout polynôme $P \in K[X]$, on pose $\Phi(P) = XP + \dfrac{X - X^2}{n} P'$. Démontrer que Φ induit un endomorphisme de $K_n[X]$. Étudier l'effet de Φ sur la base des $X^k(1-X)^{n-k}$. Exprimer 1 dans cette base. (Développer $(X+1-X)^n$ par la formule du binôme.) On pose $P_0 := 1$ et $P_{k+1} := \Phi(P_k)$. Calculer P_k puis $P'_k(1)$.

Exercices sur la section 2

II.6.16
Calculer les polynômes d'interpolation de Lagrange L_i correspondant aux n points de μ_n.

II.6.17 Calculer $\prod_{k=0}^{n-1} \cos(\theta + k\pi/n)$ et $\prod_{k=0}^{n-1} \sin(\theta + k\pi/n)$.

II.6.18 En évaluant en $X = \mathrm{e}^{\mathrm{i}\theta}$ la factorisation de $X^{2p} - 1$ dans $\mathbb{R}[X]$, retrouver la factorisation de $\dfrac{\sin p\theta}{\sin \theta}$ donnée en 2.3.

II.6.19 $*$ (i) Vérifier les inégalités $\sin x \leqslant x \leqslant \tan x$ sur $[0, \pi/2[$ et les égalités :

$$\operatorname{cotan}^2 x + 1 = \frac{1}{\sin^2 x} \quad \text{et} \quad \operatorname{cotan}(\pi - x) = -\operatorname{cotan} x.$$

(ii) À l'aide de la formule $\sum_{k=1}^{n-1} \operatorname{cotan}^2 \dfrac{k\pi}{n} = \dfrac{(n-1)(n-2)}{3}$, en déduire un encadrement de $\sum_{k=1}^{p} \dfrac{1}{k^2}$ (où $n = 2p$), puis : $\sum_{k=1}^{+\infty} \dfrac{1}{k^2} = \dfrac{\pi^2}{6}$.

II.6.20 Soit $x \in \mathbb{R}$. Étudier les suites définies par :

$$t_0 := 1, t_1 := x, t_{n+1} := 2xt_n - t_{n-1} \qquad \text{et} \qquad u_0 := 1, u_1 := 2x, u_{n+1} := 2xu_n - u_{n-1}.$$

Il y a une discussion selon que $|x|$ est inférieur, égal ou supérieur à 1.

II.6.21 Démontrer les propriétés suivantes des polynômes de Tchebychef :

$$T_n^2 + (1 - X^2)U_{n-1}^2 = 1 \qquad \text{et} \qquad \begin{pmatrix} T_{n+1} \\ U_n \end{pmatrix} = \begin{pmatrix} X & X^2 - 1 \\ 1 & X \end{pmatrix}^n \begin{pmatrix} X \\ 1 \end{pmatrix}.$$

Exercices sur la section 3

II.6.22 Soit $P \in K_n[X]$. Montrer que $Q := (1 + X)^n P\left(\dfrac{X}{1 + X}\right) \in K_n[X]$.

Si $Q = \sum_{k=0}^{n} a_k X^k$, montrer que $P = \sum_{k=0}^{n} a_k X^k (1 - X)^{n-k}$. En déduire, pour tout $i \in [\![0, n]\!]$, l'écriture de X^i dans la base de $K_n[X]$ formée des $X^k(1 - X)^{n-k}$.

II.6.23 On définit une suite $(F_n)_{n\in\mathbb{N}}$ de fractions rationnelles par la relation de récurrence : $F_0 := 0$ et, pour tout $n \in \mathbb{N}$, $F_{n+1} := \dfrac{F_n + X}{1 - XF_n}$. Montrer que F_n peut s'écrire $\dfrac{A_n}{B_n}$, où $A_0 = 0$, $B_0 = 1$, et, pour tout $n \in \mathbb{N}$, $A_{n+1} = A_n + XB_n$ et $B_{n+1} = B_n - XA_n$. Démontrer que $B_n + \mathrm{i}A_n = (1 + \mathrm{i}X)^n$ et en déduire des expressions explicites de A_n et B_n. En substituant $\tan\theta$ à X, déterminer les zéros de A_n et ceux de B_n et les zéros et les pôles de F_n, et démontrer l'égalité $F_n(\tan\theta) = \tan n\theta$.

II.6.24 Démontrer que la fraction rationnelle $\dfrac{aX+b}{cX+d}$ est non constante si, et seulement si, $ad-bc \neq 0$. On dira alors que c'est une *homographie* et on la notera H_A, où A est la matrice $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2(K)$. Démontrer que $H_A \circ H_B = H_{AB}$, que $H_{I_2} = X$. En déduire que l'ensemble des homographies, muni de la loi $\circ$, est un groupe et que $A \mapsto H_A$ est un morphisme de $GL_2(K)$ sur ce groupe. Montrer que le noyau de ce morphisme est le sous-groupe K^*I_2 de $GL_2(K)$ formé des matrices scalaires inversibles.

II.6.25 Décomposer en éléments simples les fractions rationnelles suivantes :
$\dfrac{1}{T_n}$ sur $\mathbb{R}$; $\dfrac{1}{X^n-1}$ sur $\mathbb{C}$ et sur $\mathbb{R}$; $\dfrac{1}{X^q-X} \in K(X)$ sur le corps fini K à q éléments.

II.6.26 Soit α une racine simple de B, donc un pôle simple de $F := \dfrac{A}{B}$ ·
Démontrer que le résidu de F en α est $\dfrac{A(\alpha)}{B'(\alpha)}$ ·

II.6.27 Décomposer en éléments simples la fraction rationnelle $\dfrac{1}{X+1}\cdots\dfrac{1}{X+p}$ $(p \geqslant 2)$. En déduire une expression de la somme de la série $\sum\limits_{n\geqslant 0} \dfrac{1}{(n+1)\cdots(n+p)}$.

Indication : si l'on note $H_n = \sum\limits_{k=1}^{n} \dfrac{1}{k}$, *on a, lorsque* $n \to +\infty$, $H_n = \log n + \gamma + \mathrm{o}(1)$, *où* $\gamma \in \mathbb{R}$ *est la* constante d'Euler.

II.6.28 Calculer la dérivée $n^{ème}$ de la fraction rationnelle $\dfrac{1}{X^2-2X\cos\theta+1}$ ·
Décomposer celle-ci en éléments simples.

II.6.29 Décomposer en éléments simples $\dfrac{1}{(1-X^2)(1-X^3)}$ sur $\mathbb{C}$ et en déduire une expression du nombres de solutions entières de $2x+3y=n$.
Considérer les développements limités de cette fonction rationnelle.

II.6.30 $*$ Soit $(P_n)_{n\in\mathbb{N}}$ une suite de polynômes non constants.
On définit une suite $(F_n)_{n\in\mathbb{N}}$ de fractions rationnelles par la relation de récurrence : $F_0 := P_0$ et, pour tout $n \in \mathbb{N}$, $F_{n+1} := P_{n+1} + \dfrac{1}{F_n}$ · Montrer que, pour tout n, $\deg F_n = \deg P_n$ et que F_n admet la forme irréductible $\dfrac{A_n}{B_n}$, où $A_0 = P_0, B_0 = 1$ et $A_{n+1} = A_nP_{n+1} + B_n, B_{n+1} = A_n$.

II.6.31 Montrer que l'ordre de $F(1/X)$ en 0 est $-\deg F$.

II.6.32 $*$ Soit F une fraction rationnelle. On définit une suite $(F_n)_{n\in\mathbb{N}}$ de fractions rationnelles par la relation de récurrence : $F_0 := F$ et, pour $n \geqslant 0$, si F_n n'est pas un polynôme, $F_n := Q_n + \dfrac{1}{F_{n+1}}$, où Q_n est la partie entière de F_n. Démontrer que cette suite est finie.

Les F_n *admettent des formes irréductibles* $\dfrac{A_n}{A_{n+1}}$ *telles que* $A_{n+1} = A_{n-1} \mod A_n$.

II.6.33 Pour tout $n \in \mathbb{N}$, démontrer que $K_n(X) = K_n[X] \oplus K_{-1}(X)$.

Espaces vectoriels de dimension finie

II.7

Ce module est la continuation à la fois du module II.4 sur le calcul matriciel et du module II.3 sur l'algèbre linéaire générale, ces deux modules se rejoignant dans l'étude des espaces vectoriels *de dimension finie*. Nous utiliserons donc constamment les notions et résultats des modules précités.
Pour bien apprendre et assimiler l'algèbre linéaire, il faut être « bilingue » :

- il faut comprendre et savoir parler le langage « géométrique », dans lequel les mots importants sont « espaces vectoriels, applications linéaires, noyau, image, etc ».
- il faut comprendre et savoir parler le langage « matriciel ».

Le langage géométrique est souple et puissant. Le langage matriciel, seul compris par un ordinateur, est conceptuellement plus simple et il conduit à des algorithmes. En schématisant, le langage géométrique est celui du raisonnement, le langage matriciel celui du calcul. Traduire d'un langage à l'autre est l'une des difficultés de l'algèbre linéaire, lors de l'apprentissage.
La notion de « dimension » joue un rôle essentiel en mathématiques et dans leurs applications, c'est l'analogue du « nombre de degrés de liberté » en physique. Cette notion a un aspect intuitif : nous savons qu'une droite, un cercle, plus généralement une courbe, sont des objets à une dimension, alors que le plan, une surface (par exemple celle de la terre), l'intérieur d'une ellipse, sont des objets à deux dimensions ; enfin l'espace dans lequel nous vivons a trois dimensions, un cylindre plein aussi. Dans ces exemples, le « nombre de dimensions » est le nombre de coordonnées nécessaires pour repérer un point de l'objet considéré ; dans le cas d'un espace vectoriel E, ce nombre sera simplement le nombre d'éléments d'une base de E.

1 Espaces vectoriels de dimension finie

La notion de dimension d'un espace vectoriel a déjà été rencontrée dans les modules II.3, II.4 et II.6. Notre objectif est ici d'en donner une définition précise, puis d'en étudier les propriétés simples.
Dans tout ce module, on désigne par K un corps commutatif, et les espaces vectoriels envisagés sont des K-espaces vectoriels.

1.1 Définition de la dimension

Définition 1. Un espace vectoriel E est dit *de dimension finie* s'il possède une partie génératrice *finie*, il est dit *de dimension infinie* dans le cas contraire.

Observer qu'à ce stade, nous n'avons pas encore défini la *dimension* d'un espace vectoriel, mais seulement le fait que cet espace soit « de dimension finie » (voir le théorème et définition 5 de la page suivante).

Proposition 1. Soit E un espace vectoriel de dimension finie. Toute partie génératrice de E contient une partie génératrice *finie*.

Démonstration. Soient G une partie génératrice finie de E et S une partie génératrice quelconque. Soit $x \in G$. Puisque S est génératrice, x est combinaison linéaire *d'un nombre fini* d'éléments de S : il existe une partie *finie* T_x de S telle que x appartienne à $\mathrm{Vect}(T_x)$, sous-espace vectoriel de E engendré par T_x. La réunion T des T_x, $x \in G$, est une partie finie de S, car G est finie et chaque T_x est finie. Pour tout $x \in G$, $\mathrm{Vect}(T)$ contient $\mathrm{Vect}(T_x)$, car $T \supset T_x$, donc $x \in \mathrm{Vect}(T)$. C'est vrai pour tout $x \in G$, donc $G \subset \mathrm{Vect}(T)$. Comme G engendre E, $\mathrm{Vect}(T) = E$, *i.e.* T engendre E. ■

Proposition 2. Soient E un espace vectoriel et $x_1, \ldots, x_n$ des vecteurs de E ($n \geqslant 1$). Considérons des vecteurs $y_1, \ldots, y_m \in E$ qui sont combinaisons linéaires de $x_1, \ldots, x_n$. Si $m > n$, la famille $(y_1, \ldots, y_m)$ est *liée*.

Démonstration. Par récurrence sur n. Supposons que $n = 1$. Si $y_1 = 0$ ou $y_2 = 0$, $(y_1, \ldots, y_m)$ est liée. Sinon, il existe $a_1, a_2 \in K^*$ tels que $y_1 = a_1 x_1,\ y_2 = a_2 x_1$. Alors $a_2 y_1 - a_1 y_2 = 0$, donc la famille (y_1, y_2) est liée, *a fortiori* $(y_1, \ldots, y_m)$ est liée.

Supposons $n \geqslant 2$, le résultat étant vrai à l'ordre $n-1$.

Soient $E, x_1, \ldots, x_n$ et $m, y_1, \ldots, y_m$ comme dans l'énoncé. Il existe des scalaires $a_{i,j}$ tels que :

$$\begin{cases} y_1 = a_{1,1}x_1 + \cdots + a_{1,n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ y_m = a_{m,1}x_1 + \cdots + a_{m,n}x_n. \end{cases}$$

Si $a_{j,n} = 0$ pour $j = 1, \ldots, m$, les vecteurs $y_1, \ldots, y_m$ sont combinaisons linéaires de $x_1, \ldots, x_{n-1}$. D'après l'hypothèse de récurrence, $(y_1, \ldots, y_m)$ est liée. Si les $a_{j,n}$ ne sont pas tous nuls on peut supposer (en réordonnant les y_j) que $a_{m,n} \neq 0$.

Si $j \in [\![1, m-1]\!]$, posons $\lambda_j := a_{j,n}/a_{m,n}$; alors $y_j - \lambda_j y_m$ est combinaison linéaire de $x_1, \ldots, x_{n-1}$. Comme $m - 1 > n - 1$, l'hypothèse de récurrence montre qu'il existe $b_1, \ldots, b_{m-1} \in K$ non tous nuls tels que :

$$b_1(y_1 - \lambda_1 y_m) + \cdots + b_{m-1}(y_{m-1} - \lambda_{m-1} y_m) \ = \ 0, \quad \text{soit :}$$

$$b_1 y_1 + \cdots + b_{m-1} y_{m-1} - (b_1\lambda_1 + \cdots + b_{m-1}\lambda_{m-1}) y_m \ = \ 0\,.$$

Comme $b_1, \ldots, b_{m-1}$ ne sont pas tous nuls, la famille $(y_1, \ldots, y_m)$ est liée. ■

À ce stade, nous pouvons donner une preuve complète du théorème 30 de la page 196, pour un espace vectoriel de dimension finie.

Théorème 3. Soient E un espace vectoriel de dimension finie, L une partie libre de E et G une partie génératrice de E. Il existe alors une *base finie* $\mathcal{B}$ de E telle que $L \subset \mathcal{B} \subset L \cup G$, en particulier L est finie.

Démonstration. D'après la proposition 1 de la page précédente, G contient une partie génératrice finie G'. Nous pouvons donc supposer G finie (sinon, la remplacer par G').
L'ensemble des parties de G est fini. Parmi les parties S de G telles que $L \cup S$ soit libre (par exemple la partie vide), il en existe donc une, soit S, dont le cardinal soit *maximum*. Alors $\mathcal{B} := L \cup S$ est une partie libre *maximale* parmi les parties de $L \cup G$, donc $\mathcal{B}$ est une base de E (proposition 28 de la page 195).
Soit $n := \operatorname{card}(G)$. Tout élément de E est combinaison linéaire des n éléments de G ; si $\mathcal{B}$ possédait $n+1$ éléments distincts, ces éléments seraient linéairement dépendants, d'après la proposition 2 de la page précédente, ce qui est absurde puisque $\mathcal{B}$ est libre. Ainsi $\mathcal{B}$ est finie (et a au plus n éléments). ∎

Théorème 4. Tout espace vectoriel E de dimension finie possède une base finie.

Démonstration. Il suffit d'appliquer le théorème 3 en prenant $L := \varnothing$ (qui est bien libre) et $G := E$ (qui est bien génératrice). ∎

Théorème et définition 5. Soit E un espace vectoriel de dimension finie. Toutes les bases de E sont finies et ont le même nombre d'éléments. Ce nombre est appelé *dimension* de E et noté $\dim(E)$.

Démonstration. Soient $\mathcal{B}, \mathcal{B}'$ deux bases de E. En prenant $L := G := \mathcal{B}$ dans le théorème 3, on voit que $\mathcal{B}$ est finie, de même $\mathcal{B}'$ est finie.
Soient $n := \operatorname{card}(\mathcal{B})$, $n' := \operatorname{card}(\mathcal{B}')$. Les n éléments de $\mathcal{B}$ engendrent E, et les n' éléments de $\mathcal{B}'$ sont linéairement indépendants, donc $n' \leqslant n$ (proposition 2 de la page ci-contre). De même $n \leqslant n'$, d'où $n = n'$. ∎

Désormais, l'expression « E est de dimension finie n » signifiera que E est de dimension finie, et que sa dimension est l'entier n.
Noter que la notion de dimension est relative au corps de base K considéré (voir l'exercice II.7.10). Ainsi $\mathbb{R}$ est de dimension 1 lorsqu'on le considère comme espace vectoriel sur lui-même. Par contre $\mathbb{R}$ est un $\mathbb{Q}$-espace vectoriel de dimension infinie (*cf.* les exercices II.3.6 et II.3.24 du module II.3).

Exemples.

1. Un espace vectoriel est de dimension 0 si, et seulement s'il est trivial $\big($égal à $\{0\}\big)$, *i.e.* si la partie vide en est une base.
2. Soit $n \geqslant 1$ un entier. L'espace vectoriel K^n est de dimension finie n, puisqu'il possède une base à n éléments, à savoir sa *base canonique*.

3. Soient $n \in \mathbb{N}$ et $K_n[X]$ l'ensemble des polynômes de degré au plus n à coefficients dans K. C'est un sous-espace vectoriel de $K[X]$, et $(1, X, X^2, \ldots, X^n)$ est une base de $K_n[X]$, dite *canonique*, *cf.* le module II.6. Ainsi $K_n[X]$ est de dimension finie $n+1$:
$$\dim\big(K_n[X]\big) \;=\; n+1\,. \tag{1}$$
Par contre, l'espace vectoriel $K[X]$ est de dimension infinie, car il possède une base infinie, à savoir $(X^k)_{k\geqslant 0}$.
4. Soient m, n deux entiers strictement positifs et $M_{m,n}(K)$ l'espace vectoriel des matrices de type (m,n) à coefficients dans K. Les matrices *élémentaires* $E_{i,j}$, où $(i,j) \in [\![1,m]\!] \times [\![1,n]\!]$, forment une base de $M_{m,n}(K)$ (*cf.* le module II.4), donc $M_{m,n}(K)$ est de dimension finie mn :
$$\dim\big(M_{m,n}(K)\big) \;=\; mn\,. \tag{2}$$

La dimension d'un espace vectoriel *caractérise* cet espace vectoriel à un isomorphisme près, comme le prouve le théorème suivant.

Théorème 6.

1. Deux espaces vectoriels de dimension finie sont isomorphes si, et seulement s'ils ont même dimension.
2. Soit n un entier strictement positif. Pour qu'un espace vectoriel soit de dimension n, il faut, et il suffit, qu'il soit isomorphe à K^n.

Démonstration. Soient V, W deux espaces vectoriels isomorphes de dimension finie et u un isomorphisme de V sur W. Étant donnée une base $\mathcal{B}$ de V, la remarque de la page 193 montre que $u(\mathcal{B})$ est une base de W. Puisque u est une bijection, $\mathcal{B}$ et $u(\mathcal{B})$ ont même cardinal. D'où $\dim(V) = \dim(W)$.

Soient ensuite E un espace vectoriel de dimension finie $n \geqslant 1$ et $(e_1, \ldots, e_n)$ une base de E. L'application $(a_1, \ldots, a_n) \mapsto a_1e_1 + \cdots + a_ne_n$ de K^n dans E est un isomorphisme de K^n sur E (théorème 24 de la page 194). Réciproquement, si E est isomorphe à K^n, le début de cette preuve montre que E a même dimension que K^n, c'est-à-dire n.

Enfin, si V, W sont deux espaces vectoriels de même dimension finie n, chacun d'eux est isomorphe à K^n, vu ce qui précède, donc V et W sont isomorphes (si $n = 0$, V et W sont triviaux, donc isomorphes). ∎

Définition 2. On appelle *droite* (vectorielle) tout espace vectoriel de dimension 1 et *plan* (vectoriel) tout espace vectoriel de dimension 2. Soit E un espace vectoriel de dimension finie $n \geqslant 1$. On appelle *hyperplan* (vectoriel) de E tout sous-espace vectoriel de E de dimension $n-1$.

Nous retrouvons les définitions 5 et 12 du module II.3 (*cf.* l'exemple de la page 339). Si E est une droite, le seul hyperplan de E est $\{0\}$; si E est un plan, les hyperplans de E sont les droites de E. Enfin, dans l'espace usuel $\mathbb{R}^3$, les hyperplans sont les plans vectoriels (*i.e.* passant par l'origine).

Donnons quelques conséquences du théorème 3 de la page 333.

Proposition 7. Soit E un espace vectoriel de dimension finie. De tout système générateur de E, on peut extraire une base.

Démonstration. Si G est une partie génératrice de E, le théorème 3 de la page 333 montre que G contient une base de E (prendre pour L la partie vide).

Soit maintenant $(x_i)_{i\in I}$ une famille génératrice de E. L'ensemble formé des x_i est une partie génératrice de E, qui contient une base finie $\mathcal{B}$ de E (théorème 3 de la page 333). On a $\mathcal{B} := \{x_{i_1}, \ldots, x_{i_n}\}$, où $i_1, \ldots, i_n \in I$ sont distincts. Ainsi $(x_{i_1}, \ldots, x_{i_n})$ est une base de E, extraite de la famille donnée. ∎

Théorème 8 (Théorème de la base incomplète). Dans un espace vectoriel E de dimension finie n, toute famille libre $(x_1, \ldots, x_p)$ peut être complétée en une base $(x_1, \ldots, x_p, x_{p+1}, \ldots, x_n)$ de E, en particulier $p \leqslant n$. On peut de plus imposer à $x_{p+1}, \ldots, x_n$ d'appartenir à une partie génératrice donnée G de E.

Démonstration. Soit $L := \{x_1, \ldots, x_p\}$.

Il existe une base $\mathcal{B}$ de E telle que $L \subset \mathcal{B} \subset (L \cup G)$ (théorème 3 de la page 333). Alors $p \leqslant n = \operatorname{card}(\mathcal{B})$. Si $p = n$, $L = \mathcal{B}$, et $(x_1, \ldots, x_p)$ est une base de E. Si $p < n$, les $n-p$ éléments $x_{p+1}, \ldots, x_n$ de $\mathcal{B}\backslash L$ appartiennent à G. Ainsi $\mathcal{B} = \{x_1, \ldots, x_p, x_{p+1}, \ldots, x_n\}$, donc $(x_1, \ldots, x_p, x_{p+1}, \ldots, x_n)$ est une base de E. ∎

Théorème 9 (Caractérisation des bases). Soient E un espace vectoriel de dimension finie n et $x_1, \ldots, x_p$ des vecteurs de E ($p \geqslant 1$).

1. Si $x_1, \ldots, x_p$ forment un système libre, on a $p \leqslant n$.
   Si $p = n$, ce système est une base de E.
2. Si $x_1, \ldots, x_p$ forment un système générateur de E, on a $p \geqslant n$.
   Si $p = n$, ce système est une base de E.

Démonstration. Pour une famille, l'assertion 1 résulte du théorème 8. Pour une partie, on applique ce qui précède à la famille libre $(x_1, \ldots, x_p)$.

Pour l'assertion 2, si le système est une famille, la proposition 7 montre qu'on peut en *extraire* une base $\mathcal{B}$ de E. Les n vecteurs formant $\mathcal{B}$ sont *certains* des x_i, d'où $n \leqslant p$. Si $n = p$, $\mathcal{B} = (x_1, \ldots, x_p)$, d'où la conclusion. Pour une partie, on applique ce qui précède à la famille correspondante. ∎

On retiendra en particulier la propriété suivante :

PROPRIÉTÉ. *Dans un espace vectoriel E de dimension finie n, considérons une famille de vecteurs $(x_1, \ldots, x_n)$ (le même n que la dimension de l'espace). Pour prouver que cette famille est une base de E, on peut* soit *montrer qu'elle est libre,* soit *qu'elle est génératrice. Vu le théorème précédent, il n'est pas nécessaire de vérifier les* deux *propriétés.*

Exemples.

1. Dans $\mathbb{R}^3$, $x_1 := (0,1,-1)$ et $x_2 := (1,-2,2)$ ne sont pas proportionnels. Puisque $\dim(\mathbb{R}^3) = 3$, il existe $x_3 \in \{e_1, e_2, e_3\}$ tel que (x_1, x_2, x_3) soit une base de $\mathbb{R}^3$. Comme $e_1 = 2x_1 + x_2$, e_1 ne convient pas, mais e_2 convient. En effet, les égalités $x_1 = e_2 - e_3$ et $x_2 = e_1 - 2e_2 + 2e_3$ donnent $e_3 = e_2 - x_1$ et $e_1 = 2x_1 + x_2$; ainsi $\mathrm{Vect}(x_1, x_2, e_2)$ contient e_1, e_2, e_3, donc est égal à $\mathbb{R}^3$. La famille (x_1, x_2, e_2) engendre donc $\mathbb{R}^3$, *i.e.* est une base de $\mathbb{R}^3$. Nous avons noté (e_1, e_2, e_3) la base canonique de $\mathbb{R}^3$.
2. Soient $n \in \mathbb{N}$ et $P_0, \ldots, P_n \in K[X]$ des polynômes tels que $\deg(P_k) = k$ pour tout $k = 0, \ldots, n$. À l'aide de la propriété précédente, le lecteur montrera que $(P_0, \ldots, P_n)$ est une base de $K_n[X]$.

Il n'est pas évident *a priori* qu'un sous-espace vectoriel d'un espace vectoriel de dimension finie soit de dimension finie. C'est cependant vrai :

Théorème 10. Soient E un espace vectoriel de dimension finie n et V un sous-espace vectoriel de E. Alors V est de dimension finie, et l'on a $\dim(V) \leqslant \dim(E)$. De plus, $\dim(V) = \dim(E)$ si, et seulement si, $V = E$.

Démonstration. D'après le théorème 9 de la page précédente, toute partie libre de V a au plus n éléments, car c'est une partie libre de E. Parmi les parties libres de V il en existe donc une, notons-la L, dont le cardinal p soit *maximum*, et en outre $p \leqslant n$. Vu la maximalité de p, L est une partie libre *maximale* de V, donc est une base de V (proposition 28 de la page 195). Ainsi V possède une base à p éléments, donc est de dimension finie p. Enfin, si $p = n$, L est une partie libre de E à n éléments, c'est donc une base de E, d'après le théorème 9 de la page précédente. En particulier L engendre E. Or L engendre aussi V, d'où $V = E$. ∎

Vu le théorème 6 de la page 334, il n'y a donc pas d'application linéaire *injective* f d'un espace E de dimension infinie dans un espace F de dimension finie : f définirait un isomorphisme de E sur un sous-espace vectoriel de F.

Donnons une caractérisation des nombres algébriques, définis dans le module II.6 (section 2.4), en termes de dimension. Rappelons que $x \in \mathbb{C}$ est dit *algébrique* s'il existe $P \in \mathbb{Q}[X]$ tel que $P(x) = 0$ et $P \neq 0$.

Notons $\mathbb{Q}[x]$ le sous-$\mathbb{Q}$-espace vectoriel de $\mathbb{C}$ engendré par les x^n, $n \in \mathbb{N}$.

Exercice 1.

Montrer que $x \in \mathbb{C}$ est algébrique si, et seulement si, $\mathbb{Q}[x]$ est de dimension finie en tant que $\mathbb{Q}$-espace vectoriel.

Solution. La condition est suffisante : supposons que $\mathbb{Q}[x]$ soit de dimension finie n. D'après le théorème 9 de la page précédente, la famille $(1, x, \ldots, x^n)$, formée d'éléments de $\mathbb{Q}[x]$, est liée sur $\mathbb{Q}$, ce qui signifie qu'il existe un polynôme non nul $P \in \mathbb{Q}[X]$, de degré au plus n, tel que $P(x) = 0$. Ainsi x est algébrique.

Supposons inversement que x soit algébrique, et considérons son polynôme minimal :

$$P := X^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0 \in \mathbb{Q}[X],$$

défini dans le module II.6. Puisque $P(x) = 0$, on a $x^n = -\big(a_{n-1}x^{n-1} + \cdots + a_1 x + a_0\big)$. Soit V le sous-$\mathbb{Q}$-espace vectoriel de $\mathbb{C}$ engendré par $1, x, \ldots, x^{n-1}$, il est de dimension finie n, par définition de P. L'égalité ci-dessus montre que $x^n \in V$. On en déduit que $x^p \in V$ pour tout $p \in \mathbb{N}$ (raisonner par récurrence sur p), ce qui montre que $V = \mathbb{Q}[x]$. Ainsi $\mathbb{Q}[x]$ est de dimension finie.

Exercice 2.

Montrer alors que l'ensemble $\overline{\mathbb{Q}}$ des nombres algébriques est un sous-corps de $\mathbb{C}$, résultat prouvé dans le module II.6 par une autre méthode.

Solution. Soient $x, y \in \overline{\mathbb{Q}}$ et n (resp. p) le degré du polynôme minimal de x (resp. y). Nous avons vu que $(1, x, \ldots, x^{n-1})$ est une base de $\mathbb{Q}[x]$, de même $(1, y, \ldots, y^{p-1})$ est une base de $\mathbb{Q}[y]$. Soit A le sous-$\mathbb{Q}$-espace vectoriel (de dimension finie) de $\mathbb{C}$ engendré par les $x^j y^k$, (j, k) décrivant $[\![0, n-1]\!] \times [\![0, p-1]\!]$.

De l'égalité ci-dessus donnant x^n et de son analogue pour y, on déduit que, pour *tous* $j, k \in \mathbb{N}$, le produit $x^j y^k$ appartient à A. Il en résulte aussitôt que A est un sous-anneau de $\mathbb{C}$. Ainsi A contient $x + y, x - y, xy$, mais aussi toutes les puissances de ces éléments. Par suite $\mathbb{Q}[x + y] \subset A$ est de dimension finie, en vertu du théorème précédent, et donc $x + y$ est algébrique (voir ci-dessus) ; de même $x - y$ et xy sont algébriques. Si $x \neq 0$, l'algébricité de $1/x$ se déduit de l'exercice 4 de la page 341.

Donnons ici, en dimension finie, une preuve complète du théorème 35 de la page 198 :

Théorème 11. Soient E un espace vectoriel de dimension finie et V un sous-espace vectoriel de E. Alors V possède (au moins) un *supplémentaire* dans E. Plus précisément, soit $(e_1, \ldots, e_p)$ une base de V, complétons-la en une base $(e_1, \ldots, e_p, e_{p+1}, \ldots, e_n)$ de E. Alors le sous-espace vectoriel W de E engendré par $e_{p+1}, \ldots, e_n$ est un supplémentaire de V dans E.

Démonstration. Posons $n := \dim(E)$ et $p := \dim(V)$, de sorte que $p \leqslant n$, vu le théorème précédent. Soit $(e_1, \ldots, e_p)$ une base de V. On peut la compléter en une base $(e_1, \ldots, e_n)$ de E (théorème 8 de la page 335). Soit donc $W := \mathrm{Vect}(e_{p+1}, \ldots, e_n)$. Montrons que $E = V \oplus W$.

Soit $x \in E$. Écrivons $x = a_1 e_1 + \cdots + a_n e_n$, où les a_i sont des scalaires. Alors $x = v + w$, où $v := a_1 e_1 + \cdots + a_p e_p \in V$ et $w := a_{p+1} e_{p+1} + \cdots + a_n e_n \in W$. Ainsi $V + W = E$ (*cf.* la proposition 7 de la page 171).

Il suffit maintenant de montrer que $V \cap W = \{0\}$ (*cf.* la proposition 32 de la page 196). Soit donc $x \in V \cap W$. Il existe des scalaires $a_1, \ldots, a_n$ tels que :

$$a_1 e_1 + \cdots + a_p e_p \;=\; x \;=\; a_{p+1} e_{p+1} + \cdots + a_n e_n\,, \;\text{ donc :}$$

$$a_1 e_1 + \cdots + a_p e_p + (-a_{p+1}) e_{p+1} + \cdots + (-a_n) e_n \;=\; 0\,.$$

Puisque $(e_1, \ldots, e_n)$ est une famille libre, tous les a_i sont nuls, donc $x = 0$. ∎

Théorème 12. Soient V, W deux sous-espaces vectoriels *supplémentaires* d'un espace vectoriel E de dimension finie. Alors $\dim(E) = \dim(V) + \dim(W)$.

Démonstration. Soient S une base de V et T une base de W.

La proposition 34 de la page 197 montre que S, T sont *disjointes* et que $S \cup T$ est une base de E. Ainsi $\dim(E) = \mathrm{card}\,(S \cup T) = \mathrm{card}\,(S) + \mathrm{card}\,(T) = \dim(V) + \dim(W)$. ∎

Proposition 13. Soient E et F deux espaces vectoriels de dimension finie. Le produit $E \times F$ est de dimension finie, et l'on a : $\dim(E \times F) = \dim(E) + \dim(F)$.

La dimension du *produit* $E \times F$ est la *somme* des dimensions de E et F, ce n'est pas le produit de ces dimensions.

Démonstration. Soient $(x_1, \ldots, x_n)$ une base de E et $(y_1, \ldots, y_p)$ une base de F $\big(n := \dim(E),\ p := \dim(F)\big)$. La définition de l'espace vectoriel $E \times F$ (définition 13 de la page 182) montre que, si $a_1, \ldots, a_n, b_1, \ldots, b_p \in K$, on a :

$$a_1(x_1, 0) + \cdots + a_n(x_n, 0) + b_1(0, y_1) + \cdots + b_p(0, y_p) \ = \ (x, y) \quad \text{où :}$$
$$x := a_1x_1 + \cdots + a_nx_n \qquad \text{et} \qquad y := b_1y_1 + \cdots + b_py_p\,.$$

Il en résulte aussitôt que la famille $\big((x_1, 0), \ldots, (x_n, 0), (0, y_1), \ldots, (0, y_p)\big)$ est une base de $E \times F$, et cette base a $n + p$ éléments, d'où la conclusion. ∎

En utilisant la notion de base duale (*cf.* la proposition-définition 25 de la page 194), on prouve le résultat suivant.

Proposition 14. Soit E un espace vectoriel de dimension finie. Alors l'espace dual E^* est de dimension finie et $\dim E^* = \dim E$.

1.2 Applications linéaires en dimension finie

Définition 3. Soient E, F deux espaces vectoriels et $f : E \to F$ une application linéaire. On dit que f est *de rang fini* lorsque l'image de f est de dimension finie. La dimension de cette image est alors appelée *rang* de f. Ainsi $\operatorname{rang} f \ := \ \dim(\operatorname{Im} f)$.

Si F est de dimension finie, toute application $f \in \mathcal{L}(E, F)$ est de rang fini, vu le théorème 10 de la page 336.

Exemple. Soient E un espace vectoriel de dimension finie, V, W deux sous-espaces vectoriels supplémentaires de E et $p \in L(E)$ le projecteur de E sur V parallèlement à W. L'image de p est V, donc $\operatorname{rang}(p) = \dim(V)$.

Le théorème suivant est essentiel :

Théorème 15 (Théorème du rang). Soient E un espace vectoriel de dimension finie, F un espace vectoriel quelconque et $f : E \to F$ une application linéaire.
Alors f est de rang fini, et son rang est donné par la formule :

$$\operatorname{rang} f \ = \ \dim(\operatorname{Im} f) \ = \ \dim(E) - \dim(\operatorname{Ker} f)\,. \tag{3}$$

Démonstration. Par hypothèse, E possède une partie génératrice finie S. Puisque l'application f de E dans $\operatorname{Im} f$ est linéaire surjective, la proposition 20 de la page 187 montre que $f(S)$ est une partie génératrice de $\operatorname{Im} f$. Comme $f(S)$ est finie, $\operatorname{Im} f$ est de dimension finie, *i.e.* f est de rang fini.

Soit maintenant H un supplémentaire de $\operatorname{Ker} f$ dans E ; un tel H existe, d'après le théorème 11 de la page 337. D'après le théorème 36 de la page 198, la restriction de f à H induit un isomorphisme de H sur $\operatorname{Im} f$. Comme H est de dimension finie (car E l'est), le théorème 6 de la page 334 montre que $\operatorname{Im} f$ est de dimension finie, égale à celle de H. Cela étant, puisque $E = \operatorname{Ker} f \oplus H$, le théorème 12 de la page 337 donne : $\dim(E) = \dim(\operatorname{Ker} f) + \dim(H) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f)$. ■

Exemple. Soit E un espace vectoriel de dimension finie $n \geqslant 1$. Si α est une forme linéaire non nulle sur E, $\alpha : E \to K$ est surjective. Le théorème précédent montre donc que le noyau de α est de dimension $n-1$, *i.e.* est un hyperplan de E (définition 2 de la page 334). Soit inversement H un hyperplan de E. Soit $(e_1, \ldots, e_{n-1})$ une base de H, complétée en une base $(e_1, \ldots, e_n)$ de E. Soit α la seule forme linéaire sur E telle que $\alpha(e_i) = 0$ pour $i = 1, \ldots, n-1$ et $\alpha(e_n) = 1$. Alors $\operatorname{Ker}(\alpha) = H$, donc les deux notions d'hyperplan (définition 2 de la page 334 et définition 12 de la page 181) coïncident.

Voici quelques conséquences du théorème du rang.

Théorème 16 (formule de Grassmann).
Soient E un espace vectoriel et V, W deux sous-espaces vectoriels de dimension finie de E. La somme $V + W$ est de dimension finie, à savoir :

$$\dim(V + W) \;=\; \dim(V) + \dim(W) - \dim(V \cap W)\,. \tag{4}$$

Démonstration. Soit f l'application $(v, w) \mapsto v + w$ du produit $V \times W$ dans E. Elle est linéaire, et son image est la somme $V + W$ (voir la définition 6 de la page 171). Il en résulte d'abord que $V + W$ est de dimension finie. Le noyau de f est formé des couples $(v, w) \in V \times W$ tels que $v + w = 0$, c'est-à-dire des couples $(x, -x)$, où $x \in V \cap W$. Ainsi $x \mapsto (x, -x)$ est un isomorphisme de $V \cap W$ sur $\operatorname{Ker} f$. La proposition 13 de la page précédente et le théorème du rang permettent de conclure. En effet, $\dim(V + W)$ vaut :

$$\dim(\operatorname{Im} f) = \dim(V \times W) - \dim(\operatorname{Ker} f) = \dim(V) + \dim(W) - \dim(V \cap W)\,. \quad ■$$

Proposition 17. Soient E un espace vectoriel de dimension finie n et V, W deux sous-espaces vectoriels de E. Pour que E soit somme directe de V et de W, il faut, et il suffit, qu'on ait $\dim(V) + \dim(W) = n$ et que, d'autre part, l'une au moins des conditions $V + W = E$ et $V \cap W = \{0\}$ soit vérifiée.

Démonstration. Si $E = V \oplus W$, on a $V \cap W = \{0\}$ *et* $V + W = E$, par définition ; de plus $\dim(V) + \dim(W) = n$, d'après le théorème 12 de la page 337.

Supposons que $\dim(V) + \dim(W) = n$, soit $\dim(V + W) + \dim(V \cap W) = n$, vu le théorème précédent. Le théorème 10 de la page 336 donne les équivalences suivantes :

$$V + W = E \iff \dim(V + W) = n \iff \dim(V \cap W) = 0 \iff V \cap W = \{0\}\,.$$

Ainsi chacune des conditions $V + W = E$ et $V \cap W = \{0\}$ implique l'autre et, si ces deux conditions sont vraies, on a bien $E = V \oplus W$. ■

Proposition 18. Soient E et F deux espaces vectoriels de dimension finie et $f : E \to F$ une application linéaire. Posons $n := \dim(E)$ et $p := \dim(F)$. Alors :

1. On a $\operatorname{rang}(f) \leqslant n$, et $\operatorname{rang}(f) = n$ si, et seulement si, f est *injective*.
2. On a $\operatorname{rang}(f) \leqslant p$, et $\operatorname{rang}(f) = p$ si, et seulement si, f est *surjective*.
3. Dans tous les cas, $\operatorname{rang}(f) \leqslant \min(n, p)$.

Démonstration. Posons $r := \operatorname{rang}(f) = \dim(\operatorname{Im} f)$ et $s := \dim(\operatorname{Ker} f)$, de sorte que $r = n - s$ (théorème du rang). Ainsi $r \leqslant n$; de plus $r = n$ si, et seulement si, $s = 0$, *i.e.* si $\operatorname{Ker} f = \{0\}$, autrement dit si f est injective.
Comme $\operatorname{Im} f$ est un sous-espace vectoriel de F, on a $r \leqslant p$, et $r = p$ si, et seulement si, $\operatorname{Im} f = F$ (théorème 10 de la page 336), autrement dit si f est surjective. ■

Ce résultat permet de caractériser les isomorphismes en dimension finie :

Théorème 19 (Caractérisation des isomorphismes).
Soient E, F deux espaces vectoriels de *même* dimension finie n et $f : E \to F$ une application linéaire. Pour que f soit un isomorphisme (*i.e.* soit bijective), il faut, et il suffit, qu'elle soit injective, ou qu'elle soit surjective, ou encore que le rang de f soit égal à n.
Ce théorème s'applique en particulier lorsque E est un espace vectoriel de dimension finie n et f est un *endomorphisme* de E.

Démonstration. La proposition précédente prouve en effet l'équivalence des propriétés « f est injective », « $\operatorname{rang}(f) = n$ » et « f est surjective ». ■

Exemple. Dans le module II.6, nous avons déjà rencontré l'*interpolation de Lagrange* pour les polynômes (théorème 23 de la page 298). Rappelons ce dont il s'agit.
On se donne n éléments $a_1, \ldots, a_n$ de K distincts ($n \geqslant 1$), ainsi que n éléments $b_1, \ldots, b_n$ de K arbitraires. Il existe alors un unique polynôme P appartenant à $K_{n-1}[X]$ et prenant en chaque a_i la valeur b_i. Nous allons retrouver ce résultat.
L'application $f : Q \longmapsto \big(Q(a_1), \ldots, Q(a_n)\big)$ de $K_{n-1}[X]$ dans K^n est linéaire.
Soit $Q \in \operatorname{Ker}(f)$: pour tout $i \in [\![1, n]\!]$, $Q(a_i) = 0$, *i.e.* a_i est une racine de Q, autrement dit $X - a_i$ divise Q. Les a_i étant distincts, Q est multiple de $S := \prod_{i=1}^{n} (X - a_i)$, qui est le ppcm des $X - a_i$. Or $\deg(S) = n$ et $\deg(Q) < n$, donc Q ne peut être multiple de S que s'il est nul. Cela prouve que $\operatorname{Ker}(f)$ est trivial, c'est-à-dire que f est injective (elle est linéaire). Comme $\dim\big(K_{n-1}[X]\big) = n = \dim(K^n)$, le théorème précédent montre que f est bijective. Ainsi $(b_1, \ldots, b_n) \in K^n$ possède un unique antécédent P par f, *i.e.* il existe un unique $P \in K_{n-1}[X]$ tel que $P(a_i) = b_i$ pour $i = 1, \ldots, n$.
Noter que l'existence de P a été démontrée ici à l'aide du théorème 19, ce qui illustre la puissance de ce théorème. Par contre, nous n'avons pas retrouvé la formule *explicite* obtenue dans le module II.6.
Partant inversement des polynômes d'interpolation de Lagrange $(L_i)_{i \in [\![1,n]\!]}$, on peut montrer qu'ils forment une famille libre, donc une base. En effet, soient $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$ tels

que $P := \lambda_1 L_1 + \cdots + \lambda_n L_n = 0$. On a, pour tout $i \in [\![1,n]\!]$: $0 = P(a_i) = \lambda_i$.
Pour tout $i \in [\![1,n]\!]$, on définit $\delta_i : E \mapsto \mathbb{K}$ par $\delta_i : Q \in E \mapsto Q(a_i)$; c'est un élément de E^*. On vérifie que $(\delta_i)_{i\in[\![1,n]\!]}$ est la base duale de $(L_i)_{i\in[\![1,n]\!]}$.

Exercice 3.

1. Soient E un $\mathbb{K}$-espace vectoriel de dimension finie $n \in \mathbb{N}^*$ et $\mathcal{C}$ une base de E^*. Montrer qu'il existe une unique base $\mathcal{B}$ de E dont $\mathcal{C}$ soit la base duale.
2. Soit $V \subset E^*$ un sous-espace vectoriel. On suppose que :
$$\{x \in E \mid \varphi(x) = 0,\ \forall \varphi \in V\} = \{0\}.$$
Montrer que $V = E^*$.

Solution.

1. Notons $\mathcal{C} = (\psi_i)_{i\in[\![1,n]\!]}$.
On définit une application $f : E \to \mathbb{K}^n$ par $f : x \in E \mapsto (\psi_1(x), \ldots, \psi_n(x))$. Cette application est évidemment linéaire. Montrons qu'elle est injective.
Si $\psi_1(x) = \cdots = \psi_n(x) = 0$, alors pour tout $\varphi \in E^*$, $\varphi(x) = 0$ (puisque la famille $\mathcal{C}$ est génératrice). En particulier si $\pi_1, \ldots, \pi_n \in E^*$ sont les formes linéaires coordonnées associées à une base de E, on a $\pi_1(x) = \cdots = \pi_n(x) = 0$, donc $x = 0$.
On a $\dim E = \dim \mathbb{K}^n$ donc f est un isomorphisme (théorème 19). Par suite $f^{-1} : \mathbb{K}^n \to E$ est aussi un isomorphisme et l'image $\mathcal{B} := (x_i)_{i\in[\![1,n]\!]}$ par f^{-1} de la base canonique de $\mathbb{K}^n$ est une base de E. Les (ψ_i) sont les formes linéaires coordonnées associées à la base $\mathcal{B}$, donc $\mathcal{C}$ est la base duale de $\mathcal{B}$. Si $\mathcal{B}'$ est une base de E dont $\mathcal{C}$ est la base duale, alors $f(\mathcal{B}')$ est la base canonique de $\mathbb{K}^n$, donc $\mathcal{B}' = \mathcal{B}$.
2. On raisonne par l'absurde en supposant $V \neq E^*$. On choisit une base $\mathcal{S} := (\psi_i)_{i\in[\![1,p]\!]}$ de V. On a $p < n$. On complète $\mathcal{S}$ en une base $\mathcal{C} := (\psi_i)_{i\in[\![1,n]\!]}$ de E^*. D'après la première question, il existe une base $\mathcal{B} := (x_i)_{i\in[\![1,n]\!]}$ de E dont $\mathcal{C}$ est la base duale. On a $\psi_1(x_{p+1}) = \cdots = \psi_p(x_{p+1}) = 0$, donc $x_{p+1} = 0$, ce qui est impossible.

Exercice 4.

Soient K un corps et $\mathcal{A}$ une K-algèbre (*cf.* le module II.4), de dimension finie $n \geqslant 1$ sur K et sans diviseurs de zéro : étant donnés $a, b \in \mathcal{A}$, $ab = 0$ si, et seulement si, $a = 0$ ou $b = 0$. Montrer que $\mathcal{A}$ est un corps.

Solution. Soit $a \in \mathcal{A}$, $a \neq 0$. Montrons que a possède un inverse multiplicatif. L'application $f : x \mapsto ax$ de $\mathcal{A}$ dans $\mathcal{A}$ est K-linéaire (par définition d'une algèbre), elle est injective ($x \neq 0$ implique $ax \neq 0$ car $a \neq 0$), donc surjective, vu le théorème précédent. Ainsi l'élément neutre 1 de $\mathcal{A}$ appartient à $\operatorname{Im} f$: il existe $b \in \mathcal{A}$ tel que $ab = 1$, et b convient.

Définition 4. Soit $(x_i)_{i\in I}$ une famille de vecteurs d'un espace vectoriel E. On dit que cette famille est *de rang fini* si le sous-espace vectoriel V de E engendré par les x_i est de dimension finie.
Dans ce cas, la dimension de V est appelée *rang* de la famille $(x_i)_{i\in I}$.

Exemple. Soient $x_1 := (1,4,3)$, $x_2 := (1,1,0)$, $x_3 := (1,-1,-2) \in \mathbb{R}^3$.
La famille (x_1, x_2) est libre (x_1 et x_2 ne sont pas colinéaires).
En revanche, $2x_1 - 5x_2 + 3x_3 = 0$, donc (x_1, x_2) est une base du sous-espace vectoriel de $\mathbb{R}^3$ engendré par x_1, x_2, x_3. Le rang de la famille (x_1, x_2, x_3) est donc 2.

Proposition 20. Soient E un espace vectoriel de dimension finie n et $(x_1, \ldots, x_p)$ une famille finie de vecteurs de E.

1. La famille $(x_1, \ldots, x_p)$ est de rang fini, notons-le r, et $r \leqslant \min(n,p)$.
2. On a $r \leqslant p$, et $r = p$ si, et seulement si, cette famille est *libre*.
3. On a $r \leqslant n$, et $r = n$ si, et seulement si, cette famille est *génératrice*.

Démonstration. L'application linéaire $f : (a_1, \ldots, a_p) \mapsto a_1x_1 + \cdots + a_px_p$ de K^p dans E a pour image $V = \mathrm{Vect}(x_1, \ldots, x_p)$. En particulier V est de dimension finie, car E l'est, ce qui prouve l'assertion 1. Or $r = \mathrm{rang}(f)$, donc les assertions 2 et 3 résultent du théorème 24 de la page 194 et de la proposition 18 de la page 340. ■

2 Applications linéaires et matrices

Dans cette section, nous allons généraliser les résultats du module II.4 reliant les applications linéaires de K^p dans K^n et les matrices.

2.1 Écriture matricielle d'une application linéaire

Soient V un espace vectoriel de dimension finie $n \geqslant 1$ et $\mathcal{B} := (e_1, \ldots, e_n)$ une base de V fixée. Étant donné un vecteur $x \in V$, notons $x_1, \ldots, x_n \in K$ ses composantes dans la base $\mathcal{B}$, caractérisées par l'égalité suivante :

$$x := x_1e_1 + x_2e_2 + \cdots + x_ne_n = \sum_{i=1}^{n} x_ie_i\,. \tag{5}$$

L'application $x \longmapsto (x)_{\mathcal{B}} := (x_1, \ldots, x_n)$ est un isomorphisme de V sur K^n, l'isomorphisme réciproque est $(x_1, \ldots, x_n) \mapsto x_1e_1 + \cdots + x_ne_n$.
Dans la suite, il sera commode d'identifier K^n à $M_{n,1}(K)$. Ainsi, à tout $x \in V$ comme ci-dessus, on associe le vecteur colonne $X := {}^t(x)_{\mathcal{B}}$, soit :

$$X := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in M_{n,1}(K)\,. \tag{6}$$

L'application $x \mapsto X$ est alors un isomorphisme de V sur $M_{n,1}(K) = K^n$.

Définition 5.
Soient V un espace vectoriel de dimension finie $n \geqslant 1$, $\mathcal{B} := (e_1, \ldots, e_n)$ une base de V et $u_1, \ldots, u_p$ des vecteurs de V ($p \geqslant 1$).
On appelle *matrice de la famille* $(u_1, \ldots, u_p)$ *dans la base* $\mathcal{B}$ la matrice de type (n,p) définie ainsi : pour $j = 1, \ldots, p$, la $j^{\text{ème}}$ colonne de cette matrice est formée des composantes de u_j dans la base $\mathcal{B}$. Cette matrice sera notée $\mathrm{Mat}_{\mathcal{B}}(u_1, \ldots, u_p)$.

Pour $j = 1, \ldots, p$, soit U_j le vecteur colonne des composantes de u_j dans la base $\mathcal{B}$. Ainsi $\mathrm{Mat}_{\mathcal{B}}(u_1, \ldots, u_p) = [U_1; \ldots; U_p]$ ce qui signifie, rappelons-le, que $\mathrm{Mat}_{\mathcal{B}}(u_1, \ldots, u_p) \in M_{n,p}(K)$ a pour colonnes $U_1, \ldots, U_p$.

Si l'on note $a_{i,j}$ les coefficients de cette matrice, on a donc :

$$u_j \;=\; a_{1,j}e_1 + \cdots + a_{n,j}e_n \;=\; \sum_{i=1}^{n} a_{i,j}e_i \quad (j = 1, \ldots, p)\,. \tag{7}$$

Les trois notions de rang : rang d'une famille de vecteurs, rang d'une matrice, rang d'une application linéaire, sont liées, nous allons le voir. Le *rang* d'une matrice $M \in M_{n,p}(K)$ a été défini, dans le module II.4, comme la *dimension* du sous-espace vectoriel de K^n engendré par les colonnes de M.

Théorème 21. Conservons les hypothèses et les notations de la définition 5 de la page précédente. Soit A la matrice de la famille $(u_1, \ldots, u_p)$ dans la base $\mathcal{B}$. Alors le rang de la famille de vecteurs $(u_1, \ldots, u_p)$ est égal au rang de la matrice A.

Démonstration. Soit f l'isomorphisme de $M_{n,1}(K) = K^n$ sur V appliquant tout vecteur colonne ${}^t(x_1, \ldots, x_n)$ sur $x_1e_1 + \cdots + x_ne_n$. Pour $j = 1, \ldots, p$, soit U_j le $j^{ème}$ vecteur colonne de A, formé des composantes de u_j dans la base $\mathcal{B}$, de sorte que $f(U_j) = u_j$. Par restriction, f induit un isomorphisme du sous-espace vectoriel de K^n engendré par $U_1, \ldots, U_p$ sur le sous-espace vectoriel de V engendré par $u_1, \ldots, u_p$. Ces deux sous-espaces ont donc même dimension. Or, par définition, la dimension du premier est le rang de la matrice A, et la dimension du second est le rang de $(u_1, \ldots, u_p)$. ∎

Ainsi la famille $(u_1, \ldots, u_p)$ est une base de V si, et seulement si, $p = n$ et si la matrice de cette famille dans la base $\mathcal{B}$ est inversible.

Définition 6. Soient E, F deux espaces vectoriels de dimension finie : $p := \dim(E)$ et $n := \dim(F)$. Soient $\mathcal{B} := (e_1, \ldots, e_p)$ une base de E, $\mathcal{C} := (f_1, \ldots, f_n)$ une base de F et $f : E \to F$ une application linéaire. On appelle *matrice de f dans les bases $\mathcal{B}$ et $\mathcal{C}$* la matrice de la famille $\big(f(e_1), \ldots, f(e_p)\big)$ dans la base $\mathcal{C}$. Cette matrice est notée $\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)$. Ses coefficients $a_{i,j}$ sont donc caractérisés par les formules suivantes :

$$f(e_j) \;=\; a_{1,j}f_1 + \cdots + a_{n,j}f_n \;=\; \sum_{i=1}^{n} a_{i,j}f_i \;\; (j = 1, \ldots, p)\,. \tag{8}$$

Nous dirons aussi que $\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)$ est la matrice représentant f dans les bases $\mathcal{B}$ et $\mathcal{C}$.

À propos de cette matrice, il est *essentiel* de retenir ceci :

1. le nombre de *lignes* de $\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)$ est la dimension de *l'espace d'arrivée* F de f, à savoir $\dim(F) = n$;
2. le nombre de *colonnes* de $\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)$ est la dimension de *l'espace de départ* E de f, à savoir $\dim(E) = p$.
3. pour tout indice $j \in [\![1, p]\!]$, la $j^{ème}$ colonne de $\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)$ est formée des composantes du vecteur $f(e_j)$ dans la base $\mathcal{C}$.

L'expression « la matrice de l'application linéaire f » ne veut rien dire : il est *impératif de préciser les deux bases choisies*.

Voici une application importante du théorème 21 de la page précédente :

Théorème 22. Soient E, F deux espaces vectoriels de dimension finie, $\mathcal{B}$ une base de E, $\mathcal{C}$ une base de F et $f : E \to F$ une application linéaire.
Posons $p := \dim(E)$, $n := \dim(F)$, et notons $A \in M_{n,p}(K)$ la matrice de f dans les bases $\mathcal{B}$ et $\mathcal{C}$. Alors *le rang de f est égal au rang de A.*

Démonstration. Si $\mathcal{B} := (e_1, \ldots, e_p)$, A est la matrice de la famille $w := \big(f(e_1), \ldots, f(e_p)\big)$ dans la base $\mathcal{C}$. Vu le théorème précédent, le rang de A est le rang de w, *i.e.* la dimension de $\mathrm{Vect}\big(f(e_1), \ldots, f(e_p)\big) = \mathrm{Im}(f)$, c'est-à-dire le rang de f. ■

Théorème 23.
Soient E, F deux espaces vectoriels de dimension finie p et n respectivement, $\mathcal{B} := (e_1, \ldots, e_p)$ une base de E et $\mathcal{C} := (f_1, \ldots, f_n)$ une base de F. L'application de $\mathcal{L}(E, F)$ dans $M_{n,p}(K)$ associant à toute application linéaire de E dans F sa matrice dans les bases $\mathcal{B}$ et $\mathcal{C}$ est un *isomorphisme* d'espaces vectoriels. Autrement dit, *une application linéaire de E dans F est entièrement déterminée par sa matrice dans les bases $\mathcal{B}$ et $\mathcal{C}$.*

Démonstration. Notons Φ l'application $f \mapsto \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)$. Soient d'abord $f, g \in \mathcal{L}(E, F)$ et $\lambda \in K$. Pour tout $j \in [\![1, p]\!]$, on a $(f+g)(e_j) = f(e_j) + g(e_j)$ et $(\lambda f)(e_j) = \lambda f(e_j)$. En considérant les composantes de ces différents vecteurs sur la base $\mathcal{C}$ et en faisant varier j, on en déduit les égalités :

$$\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f+g) \;=\; \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f) + \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(g) \quad \text{et} \quad \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(\lambda f) \;=\; \lambda\, \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)\,. \tag{9}$$

L'application Φ est donc linéaire. Considérons une matrice $A \in M_{n,p}(K)$, et notons $a_{i,j}$ ses coefficients. D'après le théorème 26 de la page 195, il existe une application $f \in \mathcal{L}(E, F)$ *unique* telle que, pour tout $j \in [\![1, p]\!]$, l'égalité (8) soit vérifiée, *i.e.* telle que $\Phi(f) = A$. Ainsi Φ est bijective, *i.e.* Φ est un isomorphisme. ■

Théorème 24. Conservons les hypothèses du théorème précédent. Soient $f : E \to F$ une application linéaire et $A \in M_{n,p}(K)$ sa matrice dans les bases $\mathcal{B}$ et $\mathcal{C}$. Soient $x \in E$ et $y := f(x)$. Notons

$$X := \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \in M_{p,1}(K) \quad \text{et} \quad Y := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in M_{n,1}(K) \tag{10}$$

les matrices colonnes donnant les composantes de x dans la base $\mathcal{B}$ et de y dans la base $\mathcal{C}$. La traduction matricielle de l'égalité $y = f(x)$ est :

$$Y \;=\; AX\,. \tag{11}$$

Démonstration. Notons $a_{i,j}$ les coefficients de A. Par définition de X,

$$x = \sum_{j=1}^{p} x_j e_j\,, \quad \text{d'où} \quad y = \sum_{j=1}^{p} x_j f(e_j)\,,$$

la deuxième égalité venant de la linéarité de f. Vu les égalités (8), il vient :

$$y = \sum_{j=1}^{p} x_j \left(\sum_{i=1}^{n} a_{i,j} f_i \right) = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} a_{i,j} x_j \right) f_i\,.$$

Cela montre que, pour tout $i \in [\![1,n]\!]$, la $i^{ème}$ composante y_i de y sur la base $\mathcal{C}$ vaut : $y_i = \sum\limits_{j=1}^{p} a_{i,j} x_j$. On en déduit l'égalité matricielle annoncée. ■

Voici un exemple important. Soient $A \in M_{n,p}(K)$ et $f : K^p \to K^n$ l'application linéaire $X \mapsto AX$ correspondante (*cf.* le module II.4). Dans le théorème précédent, prenons pour $\mathcal{B}$ et $\mathcal{C}$ les bases canoniques de K^p et K^n respectivement. Avec les notations de ce théorème, on a, pour tout $x \in K^p$, $X = x$ et $y = Y = AX$ $\big($en se rappelant que, pour tout $m \in \mathbb{N}^*$, K^m est identifié à $M_{m,1}(K)\big)$. On en déduit ceci, compte tenu du théorème 22 de la page ci-contre :

PROPRIÉTÉ. *Soit $A \in M_{n,p}(K)$. Alors A est la matrice de l'application linéaire $X \mapsto AX$ de K^p dans K^n, dans les bases canoniques de K^p et K^n respectivement. En particulier, le rang de A est égal au rang de cette application linéaire de K^p dans K^n, dite* canoniquement associée à A.

Exemple. Soit f l'application $(x,y,z) \mapsto (3x - 5y + z, 2x + y - 3z)$ de K^3 dans K^2. Notons $\mathcal{B} := (u_1, u_2, u_3)$ la base canonique de K^3 et $\mathcal{C} := (v_1, v_2)$ la base canonique de K^2. On vérifie que :

$$f(u_1) = 3v_1 + 2v_2\,, \quad f(u_2) = -5v_1 + v_2\,, \quad f(u_3) = v_1 - 3v_2\,.$$

Compte tenu de la définition 6 de la page 343 et de la propriété 3 suivant cette définition, la matrice de f dans les bases $\mathcal{B}$ et $\mathcal{C}$ est : $\begin{pmatrix} 3 & -5 & 1 \\ 2 & 1 & -3 \end{pmatrix}$. Le théorème 24 de la page ci-contre se traduit par l'égalité matricielle suivante, où $(x,y,z) \in K^3$ est arbitraire :

$$\begin{pmatrix} 3 & -5 & 1 \\ 2 & 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3x - 5y + z \\ 2x + y - 3z \end{pmatrix}.$$

Proposition 25. Soient E un espace vectoriel de dimension finie p et F un espace vectoriel de dimension finie n. L'espace $\mathcal{L}(E,F)$ est de dimension np.
Par exemple, le *dual* $E^* = \mathcal{L}(E,K)$ de E est de dimension p.

Démonstration. Une base $\mathcal{B}$ de E et une base $\mathcal{C}$ de F étant fixées, le théorème 23 de la page précédente fournit un isomorphisme de $\mathcal{L}(E,F)$ sur $M_{n,p}(K)$. D'où la conclusion, car $M_{n,p}(K)$ est de dimension np (exemple 4 de la page 334). ■

Théorème 26. Soient E, F, G trois espaces vectoriels de dimension finie, $\mathcal{B}$ une base de E, $\mathcal{C}$ une base de F, $\mathcal{D}$ une base de G.
Pour toutes applications linéaires $f \in \mathcal{L}(E, F)$ et $g \in \mathcal{L}(F, G)$, on a :

$$\mathrm{Mat}_{\mathcal{B}}^{\mathcal{D}}(g \circ f) \;=\; \mathrm{Mat}_{\mathcal{C}}^{\mathcal{D}}(g) \times \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)\,. \tag{12}$$

Ainsi *la composition des applications linéaires se traduit par la multiplication matricielle*. Bien noter l'ordre dans lequel est fait le produit matriciel.

Démonstration. Soient $p := \dim(E), n := \dim(F), s := \dim(G)$.
Posons $P := \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)$, $Q := \mathrm{Mat}_{\mathcal{C}}^{\mathcal{D}}(g)$ et $R := \mathrm{Mat}_{\mathcal{B}}^{\mathcal{D}}(g \circ f)$, de sorte que P, Q, R sont des matrices de type $(n, p), (s, n)$ et (s, p) respectivement. Il s'agit de montrer que $R = QP$.
Soit $x \in E$, posons $y := f(x)$ et $z := g(y)$. Notons :

- $X \in M_{p,1}(K)$ la matrice des composantes de x dans la base $\mathcal{B}$;
- $Y \in M_{n,1}(K)$ la matrice des composantes de y dans la base $\mathcal{C}$;
- $Z \in M_{s,1}(K)$ la matrice des composantes de z dans la base $\mathcal{D}$.

Vu le théorème 24 de la page 344, $Y = PX$ et $Z = QY$, donc $Z = Q(PX) = (QP)X$, mais aussi $Z = RX$, d'où $(R - QP)X = 0$. Cette égalité est vraie pour tout $X \in M_{p,1}(K)$ $\big($lorsque x décrit E, X décrit $M_{p,1}(K)\big)$. Ainsi $R - QP$ est nulle, car l'application linéaire de K^p dans K^s associée est nulle. ∎

Les applications linéaires *bijectives* se reconnaissent à leurs matrices :

Proposition 27. Soient E, F deux espaces vectoriels de *même* dimension finie $n \geqslant 1$, $\mathcal{B}$ une base de E, $\mathcal{C}$ une base de F, $f : E \to F$ une application linéaire et $A \in M_n(K)$ la matrice de f dans les bases $\mathcal{B}$ et $\mathcal{C}$. Alors f est un isomorphisme si, et seulement si, la matrice A est *inversible*, auquel cas l'inverse A^{-1} de A est la matrice de f^{-1} dans les bases $\mathcal{C}$ et $\mathcal{B}$: $\mathrm{Mat}_{\mathcal{C}}^{\mathcal{B}}(f^{-1}) \;=\; \big[\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)\big]^{-1}$.

Démonstration. Supposons f bijective.
D'après le théorème précédent, $\mathrm{Mat}_{\mathcal{C}}^{\mathcal{B}}(f^{-1}) \times \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f)$ est la matrice de $f^{-1} \circ f$ dans les bases $\mathcal{B}$ et $\mathcal{B}$ (au départ et à l'arrivée).
Comme $f^{-1} \circ f = \mathrm{Id}_E$, $\mathrm{Mat}_{\mathcal{C}}^{\mathcal{B}}(f^{-1}) \times \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f) = I_n$. En échangeant les rôles de f et f^{-1}, on voit que $\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f) \times \mathrm{Mat}_{\mathcal{C}}^{\mathcal{B}}(f^{-1}) = I_n$. Ainsi A est inversible et son inverse est la matrice $\mathrm{Mat}_{\mathcal{C}}^{\mathcal{B}}(f^{-1})$.
Si A est inversible, c'est-à-dire de rang n (*cf.* le module II.4), le théorème 22 de la page 344 montre que f est de rang n, *i.e.* est bijective (théorème 19 de la page 340). ∎

Intéressons-nous maintenant aux *endomorphismes* d'un espace vectoriel E de dimension finie $n \geqslant 1$, *i.e.* aux applications linéaires de E dans E. Rappelons que l'ensemble $\mathcal{L}(E)$ des endomorphismes est à la fois un anneau (dans lequel la multiplication est la *composition*), et un espace vectoriel. Lorsqu'on veut représenter un endomorphisme par une matrice, il est naturel de choisir la même base au départ et à l'arrivée :

Définition 7. Soient E un espace vectoriel de dimension finie $n \geqslant 1$, $f \in \mathcal{L}(E)$ et $\mathcal{B} := (e_1, \ldots, e_n)$ une base de E. On appelle *matrice de f dans la base $\mathcal{B}$* la matrice de l'application linéaire f dans les bases $\mathcal{B}$ au départ et $\mathcal{B}$ à l'arrivée. Cette matrice, notée $\mathrm{Mat}_{\mathcal{B}}(f)$, appartient à $M_n(K)$, et l'on a$\mathrm{Mat}_{\mathcal{B}}(f) := \mathrm{Mat}_{\mathcal{B}}^{\mathcal{B}}(f)$.

Nous dirons aussi que $\mathrm{Mat}_{\mathcal{B}}(f)$ est la matrice représentant f dans la base $\mathcal{B}$. Notant $a_{i,j}$ les coefficients de $\mathrm{Mat}_{\mathcal{B}}(f)$, on a donc :

$$f(e_j) = a_{1,j}e_1 + \cdots + a_{n,j}e_n = \sum_{i=1}^{n} a_{i,j}e_i \quad (j = 1, \ldots, n). \tag{13}$$

Traduisons, pour les endomorphismes, certains des résultats précédents.

Théorème 28. Soient E un espace vectoriel de dimension finie $n \geqslant 1$ et $\mathcal{B}$ une base de E fixée. L'application de $\mathcal{L}(E)$ dans $M_n(K)$ associant à tout endomorphisme de E sa matrice dans la base $\mathcal{B}$ est un isomorphisme à la fois *d'espaces vectoriels* et *d'anneaux*. Pour tous $f, g \in \mathcal{L}(E)$, on a donc :

$$\mathrm{Mat}_{\mathcal{B}}(g \circ f) = \mathrm{Mat}_{\mathcal{B}}(g) \times \mathrm{Mat}_{\mathcal{B}}(f). \tag{14}$$

Démonstration. Le théorème 23 de la page 344 montre que l'application Φ visée est un isomorphisme d'espaces vectoriels (prendre $F := E$ et $\mathcal{C} := \mathcal{B}$). L'égalité (14) vient du théorème 26 de la page précédente. Comme $\Phi(\mathrm{Id}_E) = I_n$ et $\Phi(f + g) = \Phi(f) + \Phi(g)$ pour tous $f, g \in \mathcal{L}(E)$ (parce que Φ est linéaire), Φ est un morphisme d'anneaux ; ce morphisme étant bijectif, c'est un isomorphisme d'anneaux. ∎

Exercice 5.

Montrer que la matrice A ci-dessous vérifie $A^n = I_n$ (A est carrée, d'ordre $n \geqslant 2$).

$$A := \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Solution.

Soit u l'endomorphisme de K^n de matrice A dans la base canonique $\mathcal{B} := (e_1, \ldots, e_n)$. L'examen des colonnes de A montre que $u(e_1) = e_2$, $u(e_2) = e_3, \ldots, u(e_{n-1}) = e_n$ et $u(e_n) = e_1$. Ainsi u permute circulairement les vecteurs $e_1, \ldots, e_n$. On en déduit que u^n laisse fixe chacun des e_i, donc vaut l'identité de K^n. D'après le théorème précédent, A^n est la matrice de u^n dans la base $\mathcal{B}$, c'est donc I_n.

Voici une conséquence immédiate de la proposition 27 de la page ci-contre :

Théorème 29. Soient E un espace vectoriel de dimension finie $n \geqslant 1$, $\mathcal{B}$ une base de E, f un endomorphisme de E et $A \in M_n(K)$ la matrice de f dans la base $\mathcal{B}$. Alors f est un *automorphisme* de E (*i.e.* f est bijectif) si, et seulement si, la matrice A est *inversible*, auquel cas l'inverse A^{-1} de A est la matrice de f^{-1} dans la base $\mathcal{B}$:

$$\left[\mathrm{Mat}_{\mathcal{B}}(f)\right]^{-1} = \mathrm{Mat}_{\mathcal{B}}(f^{-1}).$$

Le théorème 28 de la page précédente a la conséquence suivante :

Proposition 30. Soient E un espace vectoriel de dimension finie $n \geqslant 1$ et $\mathcal{B}$ une base de E fixée. L'application de $\mathcal{GL}(E)$ dans $GL_n(K)$ associant à tout $f \in \mathcal{GL}(E)$ sa matrice dans la base $\mathcal{B}$ est un isomorphisme de *groupes*.

Démonstration. D'après le théorème 28 de la page précédente, l'application $f \mapsto \mathrm{Mat}_{\mathcal{B}}(f)$ de $\mathcal{L}(E)$ dans $M_n(K)$ est un isomorphisme d'anneaux. Or $\mathcal{GL}(E)$ $\big($resp. $GL_n(K)\big)$ est le groupe multiplicatif des éléments inversibles de l'anneau $\mathcal{L}(E)$ (resp. $M_n(K)$). La conclusion résulte donc de la proposition 36 de la page 141. ∎

Exemples.
1. La matrice dans la base $\mathcal{B}$ de l'endomorphisme nul est la matrice nulle ; si $\lambda \in K$, la matrice dans la base $\mathcal{B}$ de l'homothétie $\lambda\,\mathrm{Id}_E$ est λI_n.
2. Soit $A \in M_n(K)$. Considérons l'endomorphisme $X \mapsto AX$ de K^n « canoniquement associé » à A. Sa matrice dans la base canonique de K^n est A elle-même (*cf.* la propriété de la page 345).
3. Soient V, W deux sous-espaces vectoriels supplémentaires de E et $p \in L(E)$ le *projecteur* de E sur V parallèlement à W. Alors $\mathrm{Im}\, p = V$ et $\mathrm{Ker}\, p = W$. Soient $(e_1, \ldots, e_r)$ une base de V et $(e_{r+1}, \ldots, e_n)$ une base de W. Alors $\mathcal{B} := (e_1, \ldots, e_n)$ est une base de E (proposition 34 de la page 197). Puisque $p(v) = v$ pour tout $v \in V$ et $p(w) = 0$ pour tout $w \in W$, il vient :

$$p(e_j) = e_j \;\; (j = 1, \ldots, r) \quad \text{et} \quad p(e_j) = 0 \;\; (j = r+1, \ldots, n)\,, \quad \text{d'où :}$$

$$\mathrm{Mat}_{\mathcal{B}}(p) \;=\; \mathrm{Diag}(\overbrace{1, 1, \ldots, 1}^{r \text{ fois } 1}, \overbrace{0, \ldots, 0}^{n-r \text{ fois } 0})\,. \qquad (*)$$

4. Gardant les hypothèses précédentes, supposons que la caractéristique de K ne soit pas égale à 2. Soit s la *symétrie* par rapport à V parallèlement à W. La restriction de s à V est Id_V, la restriction de s à W est $-\mathrm{Id}_W$. D'où :

$$\mathrm{Mat}_{\mathcal{B}}(s) \;=\; \mathrm{Diag}(\overbrace{1, 1, \ldots, 1}^{r \text{ fois } 1}, \overbrace{-1, \ldots, -1}^{n-r \text{ fois } -1})\,. \qquad (**)$$

Exercice 6.

Soient E un espace vectoriel de dimension finie n, $\mathcal{B}$ une base de E et $r \in [\![1, n-1]\!]$. Soient $p, s \in \mathcal{L}(E)$. Si l'égalité $(*)$ est vérifiée, montrer que p est un projecteur de E. Si l'égalité $(**)$ est vérifiée, montrer que s est une symétrie de E (ici, la caractéristique de K n'est pas 2).

Solution. Soient P et S les matrices indiquées. Un calcul immédiat montre que $P^2 = P$ et que $S^2 = I_n$; donc $p^2 = p$, ce qui, en toute caractéristique, caractérise les projecteurs ; et s^2 est l'identité, ce qui, en caractéristique différente de 2, caractérise les symétries (voir pour ces deux affirmations le chapitre sur les espaces vectoriels de dimension quelconque).

2.2 Changements de base

Soient E un espace vectoriel de dimension finie $p \geqslant 1$ ainsi que $\mathcal{B} := (e_1, \ldots, e_p)$ et $\mathcal{B}' := (e'_1, \ldots, e'_p)$ deux bases de E. Tout vecteur x de E est représenté par deux vecteurs colonnes : X dans la base $\mathcal{B}$ et X' dans la base $\mathcal{B}'$:

$$x = \sum_{j=1}^{p} x_j e_j = \sum_{j=1}^{p} x'_j e'_j \; ; \; X := \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} , \; X' := \begin{pmatrix} x'_1 \\ \vdots \\ x'_p \end{pmatrix} . \tag{15}$$

Comment X, X' sont-ils reliés ? Pour le savoir, introduisons une définition :

Définition 8. On appelle *matrice de passage* de la base $\mathcal{B}$ à la base $\mathcal{B}'$ la matrice de la famille $(e'_1, \ldots, e'_p)$ dans la base $\mathcal{B}$. C'est une matrice carrée d'ordre p, qui sera notée $\mathrm{Pass}(\mathcal{B}, \mathcal{B}')$.

Notons $P_{i,j}$ les coefficients de cette matrice. Vu la définition 5 de la page 342 on a :

$$e'_j = \sum_{i=1}^{p} P_{i,j} e_i \;\; (j = 1, \ldots, p) . \tag{16}$$

En comparant avec la définition 6 de la page 343, nous obtenons ceci : *la matrice de passage de la base $\mathcal{B}$ à la base $\mathcal{B}'$ est la matrice de l'application identique de E (de E dans E) dans les bases $\mathcal{B}'$ au départ et $\mathcal{B}$ à l'arrivée (*observer l'ordre*)* :

$$\mathrm{Pass}(\mathcal{B}, \mathcal{B}') = \mathrm{Mat}_{\mathcal{B}'}^{\mathcal{B}}(\mathrm{Id}_E) . \tag{17}$$

Considérons une troisième base $\mathcal{B}''$ de E. Le théorème 26 de la page 346 donne l'égalité :

$$\mathrm{Mat}_{\mathcal{B}''}^{\mathcal{B}}(\mathrm{Id}_E) = \mathrm{Mat}_{\mathcal{B}'}^{\mathcal{B}}(\mathrm{Id}_E) \times \mathrm{Mat}_{\mathcal{B}''}^{\mathcal{B}'}(\mathrm{Id}_E) .$$

D'où la propriété de transitivité suivante :

$$\mathrm{Pass}(\mathcal{B}, \mathcal{B}'') = \mathrm{Pass}(\mathcal{B}, \mathcal{B}') \times \mathrm{Pass}(\mathcal{B}', \mathcal{B}'') . \tag{18}$$

Proposition 31. Soient $\mathcal{B}$ et $\mathcal{B}'$ deux bases de E. La matrice de passage de $\mathcal{B}$ à $\mathcal{B}'$ est *inversible*, et son inverse est la matrice de passage de $\mathcal{B}'$ à $\mathcal{B}$.

Démonstration. Posons $P := \mathrm{Pass}(\mathcal{B}, \mathcal{B}')$ et $P' := \mathrm{Pass}(\mathcal{B}', \mathcal{B})$. D'après les deux propriétés ci-dessus, PP' est la matrice de l'application identique de E dans les bases $\mathcal{B}$ et $\mathcal{B}$, *i.e.* $PP' = I_p$, et de même $P'P = I_p$. ∎

Remarque. Soit $\mathcal{B}$ une base de E fixée. À toute base $\mathcal{B}'$ de E, associons la matrice $\mathrm{Pass}(\mathcal{B}, \mathcal{B}')$. Nous obtenons une application de l'ensemble des bases de E dans $GL_p(K)$. Cette application est *bijective*. L'injectivité vient des formules (16). Soit $P = (P_{i,j}) \in GL_p(K)$. Définissons $e'_1, \ldots, e'_p$ par les formules (16) et posons $\mathcal{B}' := (e'_1, \ldots, e'_p)$. Puisque P, matrice de la famille $(e'_1, \ldots, e'_p)$ dans la base $\mathcal{B}$, est inversible, le théorème 21 de la page 343 et la proposition 20 de la page 342 montrent que $\mathcal{B}'$ est une base de E. Vu les formules (16), $\mathrm{Pass}(\mathcal{B}, \mathcal{B}') = P$, d'où la surjectivité.

Ainsi, la base $\mathcal{B}$ de E étant fixée*, toute matrice $P \in GL_p(K)$ est la matrice de passage de $\mathcal{B}$ à une unique autre base de E.*

Répondons maintenant à la question posée au début de cette section :

Proposition 32 (Formule de changement de base pour un vecteur). Soit x un vecteur de E. Comme en (15), notons X (resp. X') le vecteur colonne formé des composantes de x dans « l'ancienne base » $\mathcal{B}$ (resp. dans la « nouvelle base » $\mathcal{B}'$). Alors :

$$X = PX', \quad \text{où} \quad P := \mathrm{Pass}(\mathcal{B}, \mathcal{B}')\,. \tag{19}$$

Attention : la matrice P donne $\mathcal{B}'$ en fonction de $\mathcal{B}$, mais la formule (19) *est $X = PX'$, et non $X' = PX$ comme on pourrait le penser.*

Démonstration. Écrivons $P := (P_{i,j})$. Les égalités (15) et (16) donnent :

$$x = \sum_{j=1}^{p} x'_j e'_j = \sum_{j=1}^{p} x'_j \left(\sum_{i=1}^{p} P_{i,j} e_i \right) = \sum_{i=1}^{p} \left(\sum_{j=1}^{p} P_{i,j} x'_j \right) e_i\,.$$

L'unicité de l'écriture de x comme combinaison linéaire des e_i donne $x_i = \sum_{j=1}^{p} P_{i,j} x'_j$, pour $i = 1, \ldots, p$, d'où :

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} P_{1,1} & P_{1,2} & \ldots & P_{1,p} \\ P_{2,1} & P_{2,2} & \ldots & P_{2,p} \\ \vdots & \vdots & & \vdots \\ P_{p,1} & P_{p,2} & \ldots & P_{p,p} \end{pmatrix} \begin{pmatrix} x'_1 \\ \vdots \\ x'_p \end{pmatrix} = PX'\,. \qquad \blacksquare$$

Théorème 33 (Changement de base pour une application linéaire). Soit f une application linéaire d'un espace vectoriel E de dimension finie $p \geqslant 1$ dans un espace vectoriel F de dimension finie $n \geqslant 1$. Considérons deux bases $\mathcal{B}, \mathcal{B}'$ de E et deux bases $\mathcal{C}, \mathcal{C}'$ de F. Posons :

$$A := \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f),\ A' := \mathrm{Mat}_{\mathcal{B}'}^{\mathcal{C}'}(f),\ P := \mathrm{Pass}(\mathcal{B}, \mathcal{B}'),\ Q := \mathrm{Pass}(\mathcal{C}, \mathcal{C}'). \text{ Alors :}$$

$$A' = Q^{-1}AP\,. \tag{20}$$

Ainsi *deux matrices représentant la même application linéaire dans des bases différentes sont équivalentes* (*cf.* le module II.4).

Démonstration. D'abord, P et Q sont inversibles.
Partant des égalités $f \circ \mathrm{Id}_E = f = \mathrm{Id}_F \circ f$, une double application du théorème 26 de la page 346 donne ceci :

$$\mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(f) \times \mathrm{Mat}_{\mathcal{B}'}^{\mathcal{B}}(\mathrm{Id}_E) = \mathrm{Mat}_{\mathcal{B}'}^{\mathcal{C}}(f) = \mathrm{Mat}_{\mathcal{C}'}^{\mathcal{C}}(\mathrm{Id}_F) \times \mathrm{Mat}_{\mathcal{B}'}^{\mathcal{C}'}(f)\,,$$

d'où $AP = QA'$. La formule (20) s'en déduit en multipliant à gauche par Q^{-1}, car Q est inversible. D'où la conclusion. $\blacksquare$

Donnons ici une preuve « géométrique » d'un résultat déjà démontré de façon « algorithmique » dans le module II.4. Il s'agit d'un exemple typique de traduction du langage matriciel dans le langage géométrique et *vice versa*.

Théorème 34. Soient A une matrice de $M_{n,p}(K)$ et r son rang. Alors A est équivalente à la matrice suivante (dont r coefficients valent 1) :

$$I_r^{(n,p)} := \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix} = \sum_{i=1}^{r} E_{i,i}. \tag{21}$$

Démonstration. C'est clair si $r = 0$, *i.e.* si $A = 0$. Si $r \geqslant 1$, soit $\mathcal{B}$ (resp. $\mathcal{C}$) la base canonique de K^p (resp. K^n). Soit $f : K^p \to K^n$ l'application linéaire $X \mapsto AX$, dont la matrice dans les bases $\mathcal{B}$ et $\mathcal{C}$ est A.

D'après le théorème 22 de la page 344, f est de rang r, et donc $\operatorname{Ker} f$ est de dimension $p-r$ (théorème du rang). Soit $(e'_{r+1}, \dots, e'_p)$ une base de $\operatorname{Ker} f$, complétons-la en une base $\mathcal{B}' := (e'_1, \dots, e'_p)$ de K^p. Le sous-espace vectoriel H de K^p engendré par $e'_1, \dots, e'_r$ est un supplémentaire de $\operatorname{Ker} f$ dans K^p.

Pour $j = 1, \dots, r$, posons $f'_j := f(e'_j)$. D'après le théorème 36 de la page 198, la restriction de f à H est un isomorphisme de H sur $\operatorname{Im} f$, donc $(f'_1, \dots, f'_r)$ est une base de $\operatorname{Im} f$; complétons-la en une base $\mathcal{C}' := (f'_1, \dots, f'_n)$ de K^n.

Si $j \in [\![1, r]\!]$, $f(e'_j) = f'_j$; si $r < j \leqslant p$, $e'_j \in \operatorname{Ker} f$, donc $f(e'_j) = 0$. Ainsi la matrice de f dans les bases $\mathcal{B}'$ et $\mathcal{C}'$ est $I_r^{(n,p)}$, et donc A est équivalente à la matrice $I_r^{(n,p)}$, vu la dernière assertion du théorème précédent. ∎

Pour un endomorphisme, la formule de changement de base donne ceci :

Théorème 35 (Changement de base pour un endomorphisme). Soient f un endomorphisme d'un espace vectoriel E de dimension finie $n \geqslant 1$ et $\mathcal{B}, \mathcal{B}'$ deux bases de E. Notons A (resp. A') la matrice de f dans la base $\mathcal{B}$ (resp. $\mathcal{B}'$) et P la matrice de passage de $\mathcal{B}$ à $\mathcal{B}'$. Alors :

$$A' = P^{-1}AP. \tag{22}$$

3 Déterminants

Les déterminants ont de nombreuses applications. Citons par exemple un critère pour qu'une matrice carrée soit inversible (théorème 45 de la page 361), l'étude théorique des systèmes linéaires (section 4.2), ou encore des applications géométriques (sections 3.3 et 4.2).

La définition du déterminant d'une matrice carrée est « calculatoire ». Elle fait appel au *groupe symétrique* (*cf.* le module II.2, section 2.3). Certaines démonstrations des sections 3.1 et 3.2 sont un peu techniques et peuvent être omises en première lecture ; en seconde lecture, elles permettront de mieux comprendre le groupe symétrique.

À titre d'introduction, soient $u := \begin{pmatrix} a \\ c \end{pmatrix}$ et $v := \begin{pmatrix} b \\ d \end{pmatrix}$ deux vecteurs de $\mathbb{R}^2$. Comment

voir si u et v sont colinéaires, à l'aide d'une condition simple portant sur a, b, c, d ? Voici la réponse : u et v sont colinéaires si, et seulement si, $ad - bc = 0$.
En effet, si $u = 0$, (u, v) est liée et $ad - bc = 0$. Supposons $u \neq 0$. Si (u, v) est liée, il existe $\lambda \in \mathbb{R}$ tel que $v = \lambda u$, soit $b = \lambda a$ et $d = \lambda c$, d'où $ad - bc = a(\lambda c) - (\lambda a)c = 0$.
Supposons inversement que $ad - bc = 0$. Si $a \neq 0$, on pose $\lambda := b/a$; alors $b = \lambda a$ et $a(d - \lambda c) = 0$, d'où $d = \lambda c$, donc $v = \lambda u$. Si $a = 0$, on pose de même $\lambda := d/c$.
L'expression $ad - bc$ sera notée $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ et appelée déterminant de la matrice $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, ou déterminant de la famille (u, v) dans la base canonique de $\mathbb{R}^2$. Si (u, v) est libre, $ad - bc$ présente un autre intérêt : comme nous le verrons, c'est l'*aire algébrique* du parallélogramme construit sur les vecteurs u, v : $|ad - bc|$ est l'aire *géométrique* de ce parallélogramme, et le signe de $ad - bc$ est $+$ si (u, v) est une base directe de $\mathbb{R}^2$ et $-$ sinon.
Dans le cas de n vecteurs en dimension n, les déterminants permettront entre autres de décider si n vecteurs d'un espace vectoriel de dimension n sont ou ne sont pas linéairement indépendants (*cf.* le théorème 54 de la page 366).

3.1 Déterminant d'une matrice carrée

Dans cette section uniquement, K désigne un anneau commutatif, pas nécessairement un corps : nous ne ferons pas de divisions.

Définition 9. Soit A une matrice carrée d'ordre $n \geqslant 1$, de coefficients $a_{i,j} \in K$. On appelle *déterminant de A* l'élément $\det(A)$ de K défini par la formule :

$$\det(A) := \sum_{s \in \mathfrak{S}_n} \varepsilon(s)\, a_{s(1),1} a_{s(2),2} \cdots a_{s(n),n}\,. \tag{23}$$

Dans cette formule, pour toute permutation s de $[\![1, n]\!]$, $\varepsilon(s) \in \{-1, 1\}$ est la *signature* de s (*cf.* le théorème et définition 26 de la page 129).

Le déterminant de la matrice des $a_{i,j}$ est noté comme la matrice elle-même, à ceci près que les parenthèses sont remplacées par des barres :

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{pmatrix} \quad \text{et} \quad \det(A) = \begin{vmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{vmatrix}.$$

Que donne la définition 9 pour des petites valeurs de n ? Si $n = 1$, le groupe $\mathfrak{S}_1$ est trivial, son seul élément est l'identité, de signature 1, d'où $\det(A) = a_{1,1}$. Si $n = 2$, les deux éléments de $\mathfrak{S}_2$ sont l'identité, de signature 1, et la transposition $(1\ 2)$ échangeant 1 et 2, de signature -1. D'où $\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1}a_{2,2} - a_{2,1}a_{1,2}$.
Si $n = 3$, les éléments de $\mathfrak{S}_3$ sont l'identité, les 3-cycles $(1\ 2\ 3), (1\ 3\ 2)$, de signature 1, et les transpositions $(1\ 2), (2\ 3), (1\ 3)$, de signature -1. D'où :

$$\begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{vmatrix} = \begin{aligned} &\big(a_{1,1}a_{2,2}a_{3,3} + a_{2,1}a_{3,2}a_{1,3} + a_{3,1}a_{1,2}a_{2,3}\big) \\ &-\big(a_{2,1}a_{1,2}a_{3,3} + a_{1,1}a_{3,2}a_{2,3} + a_{3,1}a_{2,2}a_{1,3}\big)\,. \end{aligned}$$

Le lecteur peut ainsi vérifier que $\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 11 \end{vmatrix} = -6$.

En général, $\mathfrak{S}_n$ étant de cardinal $n!$, $\det(A)$ est somme de $n!$ termes, chacun étant un produit de n facteurs. Le calcul de $\det(A)$ par la formule (23) nécessite donc $(n-1)n!$ multiplications et $n!$ additions. Si $n = 100$, cela fait plus de 10^{159} multiplications, c'est bien trop, même pour un ordinateur. Nous verrons plus loin une méthode de calcul beaucoup plus rapide. La formule (23) a surtout un intérêt théorique, comme nous allons le montrer.

Remarque. Supposons que K soit un *sous-anneau* d'un anneau L. Si $A \in M_n(K)$, $\det(A)$ est le même, que l'on considère A comme élément de $M_n(K)$ ou de $M_n(L)$, et c'est un élément de K. Ainsi, le déterminant d'une matrice carrée à coefficients entiers est un entier ; si les coefficients de A appartiennent à l'anneau $K[X]$, le déterminant de A appartient à $K[X]$.
Soient ensuite f un morphisme de K dans un anneau commutatif K' et $A \in M_n(K)$, de coefficients $a_{i,j}$. Notons $f(A) \in M_n(K')$ la matrice de coefficients $f(a_{i,j})$. Appliquons f aux deux membres de l'égalité (23) :

$$\det\big(f(A)\big) \;=\; f\big(\det(A)\big)\,. \tag{24}$$

Comme exemple, prenons $K = K' = \mathbb{C}$. Soit $f : \mathbb{C} \longrightarrow \mathbb{C}$ la conjugaison complexe. Si $A \in M_n(\mathbb{C})$, la matrice $f(A)$ est aussi notée $\bar{A}$ et appelée conjuguée de A. L'égalité (24) donne ici $\det(\bar{A}) = \overline{\det(A)}$.

Si $C_1, \ldots, C_n$ sont les colonnes de A, $\det(A)$ sera noté provisoirement $\det(C_1, \ldots, C_n)$ (*cf.* la définition 11 de la page 365). Voici des propriétés importantes de l'application $(C_1, \ldots, C_n) \longmapsto \det(C_1, \ldots, C_n)$ de $(K^n)^n$ dans K :

Théorème 36. Notons $C_1, \ldots, C_n$ les colonnes de la matrice $A \in M_n(K)$. Alors :

1. Soit $j \in [\![1, n]\!]$. Pour tout $\lambda \in K$, on a :
$$\det(C_1, \ldots, \lambda C_j, \ldots, C_n) \;=\; \lambda \det(C_1, \ldots, C_j, \ldots, C_n)\,. \tag{25}$$
Si C_j est somme de deux vecteurs colonnes C, C', on a :
$$\det(A) = \det(C_1, \ldots, C, \ldots, C_n) + \det(C_1, \ldots, C', \ldots, C_n)\,. \tag{26}$$
Autrement dit, $\det(C_1, \ldots, C_n)$ *dépend linéairement de* C_j lorsque les autres colonnes sont fixées.
2. Si l'une des colonnes C_j est nulle, $\det(A) = 0$.
3. Si deux des colonnes de A sont égales, $\det(A) = 0$.
4. L'échange de deux colonnes de A multiplie le déterminant par -1.
5. Si l'on effectue une permutation $s \in \mathfrak{S}_n$ sur les colonnes de A, le déterminant est multiplié par la *signature* $\varepsilon(s)$ de s :
$$\det\big(C_{s(1)}, \ldots, C_{s(n)}\big) \;=\; \varepsilon(s) \det(C_1, \ldots, C_n)\,. \tag{27}$$
6. Pour tout $\lambda \in K$, $\det(\lambda A) = \lambda^n \det(A)$.

Démonstration. Pour toute permutation s de $[\![1,n]\!]$, posons :

$$P_s \;:=\; P_s(C_1,\dots,C_n) \;=\; a_{s(1),1}a_{s(2),2}\cdots a_{s(n),n}\,. \tag{28}$$

Soit $j \in [\![1,n]\!]$. Multiplions C_j par $\lambda \in K$. Pour toute s, le facteur $a_{s(j),j}$ de P_s est multiplié par λ, et les autres facteurs ne changent pas. Ainsi P_s est multiplié par λ. Cela étant, la formule (23) s'écrit :

$$\det(A) \;=\; \det(C_1,\dots,C_n) \;=\; \sum_{s\in\mathfrak{S}_n} \varepsilon(s)\,P_s\,. \tag{29}$$

La formule (25) en résulte. La preuve de la formule (26) est analogue.

La propriété 2 résulte de la formule (25) : prendre $\lambda = 0$. La propriété 6 résulte aussi de (25), appliquée successivement à chacune des colonnes.

La propriété 4 résulte de 1 et 3. Soient en effet $i,j \in [\![1,n]\!]$ deux indices distincts avec, par exemple, $i<j$. Pour tous vecteurs $C,C' \in M_{n,1}(K)$, posons :

$$d(C,C') \;:=\; \det(C_1,\dots,\underset{\substack{\uparrow\\ i}}{C},\dots,\underset{\substack{\uparrow\\ j}}{C'},\dots,C_n)\,.$$

La propriété 1 donne :

$$\begin{aligned} d(C+C',C+C') &= d(C,C+C') + d(C',C+C') \\ &= \big[d(C,C)+d(C,C')\big] + \big[d(C',C)+d(C',C')\big]\,. \end{aligned}$$

Or $d(C+C',C+C') = d(C,C) = d(C',C') = 0$, en vertu de la propriété 3, d'où l'on déduit $d(C,C')+d(C',C)=0$. La propriété 4 s'obtient en prenant $C := C_i$ et $C' := C_j$.

La propriété 5 est une conséquence de la propriété 4, car toute permutation s de $[\![1,n]\!]$ est la composée d'un certain nombre r de transpositions, et l'on a de plus $\varepsilon(s) = (-1)^r$ (*cf.* le théorème et définition 26 de la page 129).

Pour la propriété 3, supposons que $C_i = C_j$, où $i,j \in [\![1,n]\!]$ et $i<j$ (donc $n \geqslant 2$). Soit t la transposition $(i\;j)$ échangeant i et j. Pour tout $s \in \mathfrak{S}_n$, $P_{s\circ t} = P_s$. En effet $a_{(s\circ t)(h),h} = a_{s(h),h}$ pour tout $h \in [\![1,n]\!] \setminus \{i,j\}$ et, puisque $C_i = C_j$, on a :

$$a_{(s\circ t)(i),i}\,a_{(s\circ t)(j),j} \;=\; a_{s(j),i}\,a_{s(i),j} \;=\; a_{s(j),j}\,a_{s(i),i}\,.$$

Comme $\varepsilon(s\circ t) = \varepsilon(s)\varepsilon(t) = -\varepsilon(s)$, on a $\varepsilon(s)P_s + \varepsilon(s\circ t)P_{s\circ t} = 0$. Posons ensuite :

$$S \;:=\; \big\{s \in \mathfrak{S}_n \mid s(i) < s(j)\big\} \quad\text{et}\quad S' \;:=\; \big\{s \in \mathfrak{S}_n \mid s(i) > s(j)\big\}\,.$$

Alors S et S' forment une partition de $\mathfrak{S}_n$, et $s \mapsto s\circ t$ est une bijection de S sur S'. La formule (29) s'écrit alors, par changement d'indice :

$$\begin{aligned} \det(A) &= \sum_{s\in\mathfrak{S}_n} \varepsilon(s)\,P_s \;=\; \sum_{s\in S}\varepsilon(s)\,P_s + \sum_{s\in S'}\varepsilon(s)\,P_s \\ &= \sum_{s\in S}\varepsilon(s)\,P_s + \sum_{s\in S}\varepsilon(s\circ t)\,P_{s\circ t} \;=\; \sum_{s\in S}\Big[\underbrace{\varepsilon(s)\,P_s + \varepsilon(s\circ t)\,P_{s\circ t}}_{=0}\Big] \;=\; 0\,. \end{aligned}$$

■

Théorème 37. Pour toute matrice $A \in M_n(K)$, on a $\det({}^tA) = \det(A)$.

Démonstration. L'application $s \mapsto s^{-1}$ est une bijection de $\mathfrak{S}_n$ sur $\mathfrak{S}_n$. Dans la formule (23), effectuons le changement d'indice $s := t^{-1}$.

Comme ε est un morphisme de $\mathfrak{S}_n$ dans $\{-1,1\}$, on a $\varepsilon(s^{-1}) = \varepsilon(s)^{-1} = \varepsilon(s)$ pour tout s, d'où :

$$\det(A) \;=\; \sum_{t\in\mathfrak{S}_n} \varepsilon(t)\, a_{t^{-1}(1),1} a_{t^{-1}(2),2} \cdots a_{t^{-1}(n),n} \;=\; \sum_{t\in\mathfrak{S}_n} \varepsilon(t) \left(\prod_{i=1}^{n} a_{t^{-1}(i),i}\right).$$

Soit $t \in \mathfrak{S}_n$ une permutation fixée. Effectuons le changement d'indice (bijectif) $j := t^{-1}(i)$, d'où $i = t(j)$:

$$\prod_{i=1}^{n} a_{t^{-1}(i),i} \;=\; \prod_{j=1}^{n} a_{j,t(j)}\,, \quad \text{puis} \quad \det(A) \;=\; \sum_{t\in\mathfrak{S}_n} \varepsilon(t)\, a_{1,t(1)} a_{2,t(2)} \cdots a_{n,t(n)}\,.$$

Notons $b_{i,j}$ les coefficients de ${}^tA$, de sorte que $b_{i,j} = a_{j,i}$ pour tous $i,j \in [\![1,n]\!]$.
L'égalité précédente s'écrit alors :

$$\det(A) \;=\; \sum_{t\in\mathfrak{S}_n} \varepsilon(t)\, b_{t(1),1} b_{t(2),2} \cdots b_{t(n),n} \;=\; \det({}^tA)\,.$$

■

Théorème 38.
Le théorème 36 de la page 353 reste vrai en remplaçant « colonne » par « ligne ».

Démonstration. Soient $L_1, \ldots, L_n$ les lignes de A. Les colonnes de ${}^tA$ sont donc ${}^tL_1, \ldots, {}^tL_n$. Si $h, i \in [\![1,n]\!]$ sont distincts et si $L_h = L_i$, ${}^tA$ a deux colonnes égales, donc $\det(A) = \det({}^tA) = 0$. D'où la propriété analogue à la propriété 3 du théorème 36 de la page 353. La preuve des autres propriétés est semblable. ■

Le théorème suivant est capital.

Théorème 39. Soient $A, B \in M_n(K)$, $n \geqslant 1$. Alors $\det(AB) = (\det A)(\det B)$. En particulier, si A est inversible, le déterminant de A^{-1} vaut $1/\det(A)$.

Démonstration. Notons $a_{i,j}, b_{i,j}, c_{i,j}$ les coefficients des matrices A, B, AB respectivement. Par définition, on a :

$$\det(AB) \;=\; \sum_{s\in\mathfrak{S}_n} \varepsilon(s)\, c_{s(1),1} c_{s(2),2} \cdots c_{s(n),n}\,. \qquad (*)$$

Par définition du produit de deux matrices on a, pour tous $i,j \in [\![1,n]\!]$:

$$c_{i,j} \;=\; \sum_{k=1}^{n} a_{i,k} b_{k,j}\,.$$

Dans la formule $(*)$, développons chaque terme $c_{s(i),i}$. On obtient ceci :

$$\det(AB) \;=\; \sum_{s\in\mathfrak{S}_n} \varepsilon(s) \left(\sum_{k_1=1}^{n} a_{s(1),k_1} b_{k_1,1}\right) \cdots \left(\sum_{k_n=1}^{n} a_{s(n),k_n} b_{k_n,n}\right).$$

On peut intervertir les différents signes somme :

$$\det(AB) \;=\; \sum_{k_1=1}^{n} \cdots \sum_{k_n=1}^{n} b_{k_1,1} \cdots b_{k_n,n} \left(\sum_{s\in\mathfrak{S}_n} \varepsilon(s)\, a_{s(1),k_1} \cdots a_{s(n),k_n}\right). \qquad (**)$$

Pour tout $k = (k_1, \ldots, k_n) \in [\![1, n]\!]^n$, posons $Q_k := \sum_{s \in \mathfrak{S}_n} \varepsilon(s)\, a_{s(1),k_1} \cdots a_{s(n),k_n}$.

Notant $C_1, \ldots, C_n$ les colonnes de A, on constate que $Q_k = \det(C_{k_1}, \ldots, C_{k_n})$. Si deux des indices $k_1, \ldots, k_n$ sont égaux, $Q_k = 0$, comme déterminant d'une matrice ayant deux colonnes égales. Sinon, $i \mapsto k_i$ est une permutation t de $[\![1, n]\!]$, et la propriété 5 du théorème 36 de la page 353 montre que $Q_k = \varepsilon(t) \det(C_1, \ldots, C_n)$, soit $Q_k = \varepsilon(t) \det(A)$. Le second membre de la formule $(**)$ ne change pas si, au lieu de faire varier $k_1, \ldots, k_n$ de 1 à n indépendamment, on fait varier la permutation t dans $\mathfrak{S}_n$ et l'on pose $k_i := t(i)$ pour tout i. Ce second membre devient :

$$\sum_{t \in \mathfrak{S}_n} b_{t(1),1} \cdots b_{t(n),n} \big[\varepsilon(t) \det(A)\big] = \det(A) \sum_{t \in \mathfrak{S}_n} \varepsilon(t)\, b_{t(1),1} \cdots b_{t(n),n} = \det(A) \det(B)\,.$$

Si A est inversible, le fait que $\det(A^{-1})$ soit l'inverse de $\det(A)$ vient de la formule que nous venons de démontrer (prendre $B := A^{-1}$), puisque $\det(I_n) = 1$. ■

Théorème 40. Le déterminant d'une matrice triangulaire, supérieure ou inférieure, est égal au produit des coefficients diagonaux de cette matrice.

Démonstration. Soit A une matrice triangulaire supérieure d'ordre n, de coefficients $a_{i,j}$. Son déterminant est donné par les formules (28) et (29). Il suffit de montrer que $P_s = 0$ pour toute permutation s de $[\![1, n]\!]$ autre que l'identité car, si s est l'identité, P_s est le produit des coefficients diagonaux $a_{i,i}$ de A. Soit donc $s \in \mathfrak{S}_n$ telle que $P_s \neq 0$. Pour tout $i \in [\![1, n]\!]$, le facteur $a_{s(i),i}$ n'est donc pas nul. Comme A est triangulaire supérieure, cela n'est possible que si $s(i) \leqslant i$. Ainsi $s(i) \leqslant i$ pour $i = 1, \ldots, n$. On en déduit, par exemple par récurrence sur i, que $s(i) = i$ pour tout i, donc s est l'identité. Si A est triangulaire inférieure, on remplace $s(i) \leqslant i$ par $s(i) \geqslant i$. ■

Théorème 41. Le déterminant d'une matrice ne change pas si l'on ajoute à une colonne une combinaison linéaire des autres colonnes. Cela reste vrai en remplaçant « colonne » par « ligne ».

Démonstration. Soient $A \in M_n(K)$ et $C_1, \ldots, C_n$ les colonnes de A.
Soit $j \in [\![1, n]\!]$ un indice fixé. Pour tout vecteur colonne $X \in M_{n,1}(K)$, posons $f(X) := \det(C_1, \ldots, C_{j-1}, X, C_{j+1}, \ldots, C_n)$; c'est le déterminant de la matrice obtenue en remplaçant C_j par X. Évidemment $\det(A) = f(C_j)$. La propriété 1 du théorème 36 de la page 353 montre que f est une application *linéaire* de $M_{n,1}(K)$ dans K. Par ailleurs, pour tout indice $i \neq j$, $f(C_i)$ est le déterminant d'une matrice ayant deux colonnes égales (celles d'indices i et j), donc $f(C_i) = 0$ (propriété 3 du théorème 36).
Considérons alors des scalaires λ_i, pour $i \in [\![1, n]\!] \setminus \{j\}$. Remplaçons C_j par

$$X \;:=\; C_j + \sum_{1 \leqslant i \leqslant n,\, i \neq j} \lambda_i C_i\,.$$

Le déterminant de la matrice obtenue est $f(X)$, soit, par linéarité de f :

$$f(X) \;=\; f(C_j) + \sum_{1 \leqslant i \leqslant n,\, i \neq j} \lambda_i\, f(C_i) \;=\; f(C_j) \;=\; \det(A)\,.$$

D'où la première assertion. On applique ensuite le théorème 37 de la page 354. ■

Dans le module II.4, ont été définies des opérations élémentaires sur les lignes (resp. colonnes) d'une matrice M, se traduisant par la multiplication à gauche (resp. droite) de M par une matrice carrée de l'un des types suivants : matrices d'échange, matrices de transvection, matrices de dilatation. Le théorème ci-après fait appel aux matrices d'échange et de transvection.

Exemples.

1. Soient $i, j \in [\![1, n]\!]$, $i \neq j$ et $S_{i,j}^{(n)} := I_n - E_{i,i}^{(n,n)} - E_{j,j}^{(n,n)} + E_{i,j}^{(n,n)} + E_{j,i}^{(n,n)}$ la matrice d'échange, déduite de I_n par échange des lignes d'indices i et j. La propriété 4 du théorème 36 de la page 353 et le théorème 38 de la page 355 montrent que $\det\big(S_{i,j}^{(n)}\big) = -1$.
2. Soient $i, j \in [\![1, n]\!]$, $i \neq j$, et $\lambda \in K$.
   Considérons la matrice de transvection $T_{i,j}^{(n)}(\lambda) := I_n + \lambda E_{i,j}^{(n,n)}$. Elle est triangulaire (supérieure si $j > i$, inférieure si $j < i$), et ses coefficients diagonaux valent 1. Son déterminant est donc 1 (théorème 40 de la page précédente).

Théorème 42. Soit $A \in M_n(K)$, $n \geqslant 1$. À l'aide d'échanges et de transvections sur les lignes, on peut transformer A en une matrice triangulaire supérieure M. De plus, si m est le nombre d'échanges effectués sur les lignes et si $d_1, \ldots, d_n$ sont les coefficients diagonaux de M, on a $\det(A) = (-1)^m d_1 d_2 \cdots d_n$.

Démonstration. À l'aide d'échanges et de transvections sur les lignes, on peut transformer A en une matrice $M \in M_n(K)$ échelonnée en lignes (section 2.2 du module II.4). Puisque M est carrée et échelonnée en lignes, elle est triangulaire supérieure, notons $d_1, \ldots, d_n$ ses coefficients diagonaux.
Les opérations élémentaires sur les lignes correspondent à la multiplication à gauche par des matrices élémentaires (*cf.* le module II.4). Il existe donc des matrices élémentaires $R_1, \ldots, R_s$ telles que $M = R_1 R_2 \cdots R_s A$. Parmi les R_i, m sont des matrices d'échange, de déterminant -1, les autres sont des transvections, de déterminant 1 (exemples 1 et 2 ci-dessus). D'où $\det(M) = (-1)^m \det(A)$, et l'on conclut grâce au théorème 40 de la page ci-contre. ∎

Remarque. Soit $A \in M_n(K)$. La preuve ci-dessus suggère un algorithme pour calculer $\det(A)$. Plus précisément, en analysant la démonstration donnée dans la section 2.2 du module II.4, et en ne tenant compte que des multiplications et divisions, on obtient ceci :

- une transvection, définie par une matrice $T_{i,j}^{(n)}(\lambda) = I_n + \lambda E_{i,j}^{(n)}$ demande une division (pour calculer λ) et $n-1$ multiplications ;
- une fois un pivot choisi sur la première colonne et amené sur la première ligne, $n-1$ transvections permettent d'annuler les autres coefficients de cette colonne ;
- en opérant ainsi successivement sur les colonnes de A, on transforme A en une matrice triangulaire supérieure.

Le nombre de multiplications ou divisions pour y parvenir est au plus (le vérifier) :

$$n(n-1)+(n-1)(n-2)+\cdots+2\times 1 \;=\; \frac{n(n^2-1)}{3}\,.$$

Le calcul de $\det(A)$ par cet algorithme nécessite ainsi moins de n^3 multiplications, cet algorithme est donc rapide : si $n=100$, cela fait moins de 10^6 multiplications, à comparer avec 10^{159} si l'on applique la formule (23).

3.2 Mineurs d'une matrice

Soit $A \in M_{m,n}(K)$ une matrice de coefficients $a_{i,j}$. Considérons une partie non vide I de $[\![1,m]\!]$ et une partie non vide J de $[\![1,n]\!]$. On écrit :

$$I := \{i_1,\ldots,i_p\}\,,\;\; J := \{j_1,\ldots,j_q\}\,,$$

avec $1 \leqslant i_1 < i_2 < \cdots < i_p \leqslant m$ et $1 \leqslant j_1 < j_2 < \cdots < j_q \leqslant n$.
Notons $A_{I,J}$ la matrice de type (p,q) dont le coefficient d'indices (h,k) est a_{i_h,j_k}, ceci pour tout $(h,k) \in [\![1,p]\!]\times[\![1,q]\!]$. Nous dirons que $A_{I,J}$ est la matrice *extraite* de A obtenue en prenant les lignes d'indices appartenant à I et les colonnes d'indices appartenant à J. Dans le cas où $p=q$, le déterminant d'une telle matrice *carrée* extraite de A sera appelé *mineur d'ordre* p de A.
Voici un cas particulier qui va s'avérer important.
On suppose que A est carrée ($m=n\geqslant 2$). Deux indices $i,j \in [\![1,n]\!]$ étant donnés, on prend $I := [\![1,n]\!]\setminus\{i\}$ et $J := [\![1,n]\!]\setminus\{j\}$. La matrice extraite correspondante $A_{I,J}$, obtenue en *supprimant* dans A la $i^{\text{ème}}$ ligne et la $j^{\text{ème}}$ colonne, sera notée simplement $A_{i,j}$, et son déterminant sera noté $D_{i,j}$. Ainsi :

$$A \;:=\; \begin{pmatrix} 1 & -1 & 3 \\ 2 & 4 & 1 \\ 3 & 5 & 8 \end{pmatrix}\,,\;\; A_{3,2} \;:=\; \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix}\,.$$

Les $D_{i,j}$ sont donc les mineurs d'ordre $n-1$ de A. Le théorème ci-après fournit une méthode « récursive » de calcul d'un déterminant : le calcul d'un déterminant d'ordre n se ramène au calcul de n déterminants d'ordre $n-1$:

Théorème 43. Soit A une matrice carrée d'ordre $n \geqslant 2$ à coefficients dans K et $k \in [\![1,n]\!]$ un indice fixé. Alors :

1. On peut développer $\det(A)$ par rapport à la $k^{\text{ème}}$ ligne :

$$\det(A) \;=\; \sum_{j=1}^{n} (-1)^{k+j} a_{k,j}\, D_{k,j}\,. \tag{30}$$

2. On peut de même développer $\det(A)$ par rapport à la $k^{\text{ème}}$ colonne :

$$\det(A) \;=\; \sum_{i=1}^{n} (-1)^{i+k} a_{i,k}\, D_{i,k}\,. \tag{31}$$

Démonstration. Pour tout $j \in [\![1,n]\!]$, posons $S_j := \{s \in \mathfrak{S}_n \mid s(j) = k\}$. Les S_j forment une partition de $\mathfrak{S}_n$, donc la formule (23) peut être écrite ainsi :

$$\det A = \sum_{s\in\mathfrak{S}_n} \varepsilon(s) \left(\prod_{h=1}^{n} a_{s(h),h}\right) = \sum_{j=1}^{n} a_{k,j} \left(\sum_{s\in S_j} \varepsilon(s) \left(\prod_{1\leqslant h\leqslant n,\, h\neq j} a_{s(h),h}\right)\right).$$

Fixons $j \in [\![1,n]\!]$. Notons φ la permutation de $[\![1,n]\!]$ définie par $\varphi(k) := n$, $\varphi(h) := h$ si $h < k$ et $\varphi(h) := h-1$ si $h > k$. Soit ψ la permutation analogue, où j remplace k. Les permutations φ et ψ ont pour nombres d'inversions $n-k$ et $n-j$ respectivement, donc pour signatures $(-1)^{n-k}$ et $(-1)^{n-j}$. Si $t \in \mathfrak{S}_{n-1}$, notons $\hat{t} \in \mathfrak{S}_n$ le prolongement de t défini par $\hat{t}(n) := n$, puis posons $s := \varphi^{-1} \circ \hat{t} \circ \psi$. Il est clair que $\varepsilon(\hat{t}) = \varepsilon(t)$, d'où $\varepsilon(s) = (-1)^{k+j}\varepsilon(t)$. De plus $t \mapsto s$ est une bijection de $\mathfrak{S}_{n-1}$ sur S_j. La formule de changement d'indice donne donc :

$$\sum_{s\in S_j} \varepsilon(s) \left(\prod_{1\leqslant h\leqslant n,\, h\neq j} a_{s(h),h}\right) = (-1)^{k+j} \sum_{t\in\mathfrak{S}_{n-1}} \varepsilon(t) \left(\prod_{1\leqslant h\leqslant n,\, h\neq j} a_{s(h),h}\right).$$

Notons $B := (b_{u,v})$ la matrice A privée de sa $k^{\text{ème}}$ ligne et de sa $j^{\text{ème}}$ colonne. Par définition, $\det B = D_{k,j}$. Si $v \in [\![1,n-1]\!]$, $h := \psi^{-1}(v) \in [\![1,n]\!]$ est distinct de j, et $s(h) = \varphi^{-1}\big(t(v)\big)$; de plus, on vérifie que $b_{t(v),v} = a_{s(h),h}$. Par changement d'indice, $v \mapsto h$ étant une bijection de $[\![1,n-1]\!]$ sur $[\![1,n]\!] \setminus \{j\}$, on obtient alors :

$$\prod_{1\leqslant h\leqslant n,\, h\neq j} a_{s(h),h} = \prod_{v=1}^{n-1} b_{t(v),v}\,.$$

La formule (30) en résulte, compte tenu des formules établies précédemment.

La formule (31) se déduit de la formule (30) appliquée à ${}^tA$, vu le théorème 37 de la page 354. ■

Voici par exemple ce que donne le développement d'un déterminant d'ordre 3 par rapport à sa première ligne :

$$\begin{aligned}\begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{vmatrix} &= (-1)^{1+1}a_{1,1}D_{1,1} + (-1)^{1+2}a_{1,2}D_{1,2} + (-1)^{1+3}a_{1,3}D_{1,3} \\ &= a_{1,1}(a_{2,2}a_{3,3} - a_{2,3}a_{3,2}) - a_{1,2}(a_{2,1}a_{3,3} - a_{2,3}a_{3,1}) \\ &\qquad + a_{1,3}(a_{2,1}a_{3,2} - a_{2,2}a_{3,1})\,.\end{aligned}$$

Exercice 7.

Calculer le déterminant $D := \begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 11 \end{vmatrix}$.

Solution. Par linéarité du déterminant par rapport à sa dernière ligne, il vient :

$$D = \underbrace{\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix}}_{D_1} + \underbrace{\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 0 & 0 & 2 \end{vmatrix}}_{D_2}.$$

En notant L_1, L_2, L_3 les lignes de D_1, on a $L_1 + L_3 = 2L_2$, donc $D_1 = 0$.
Développons D_2 par rapport à sa troisième ligne :

$$D = D_2 = 2\begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix} = 2 \times \big[1 \times 5 - 2 \times 4\big] = -6\,.$$

Donnons un exemple plus général de développement d'un déterminant.

Exercice 8 (Déterminant de Vandermonde).

Soient $x_1, \ldots, x_n$ des éléments de K $(n \geqslant 1)$. Le déterminant suivant :

$$V(x_1, \ldots, x_n) := \begin{vmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} \end{vmatrix} \tag{32}$$

est appelé *déterminant de Vandermonde de* $x_1, \ldots, x_n$ (exercice II.4.10 de la page 239).
La $i^{ème}$ ligne est formée des puissances x_i^j de x_i, $j = 0, 1, \ldots, n-1$.
Noter que $V(x_1, \ldots, x_n) = 0$ si deux des x_i sont égaux, car deux des lignes de ce déterminant sont alors égales. Si $n = 1$, $V(x_1) = 1$. Si $n = 2$, $V(x_1, x_2) = \begin{vmatrix} 1 & x_1 \\ 1 & x_2 \end{vmatrix} = x_2 - x_1$.
Dans le cas général, prouver la formule :

$$V(x_1, \ldots, x_n) = \prod_{1 \leqslant i < j \leqslant n} (x_j - x_i)\,. \tag{33}$$

Solution. Raisonnons par récurrence sur n. Le résultat est vrai lorsque n vaut 1 ou 2. Supposons $n \geqslant 3$ et la formule vraie pour $n-1$. Prouvons la formule (33), pour des éléments $x_1, \ldots, x_n$ de K donnés. La conclusion est claire si deux des x_i sont égaux car les deux membres de la formule sont alors nuls. Supposons donc les x_i distincts. Considérons le déterminant :

$$V(x_1, \ldots, x_{n-1}, X) := \begin{vmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \ldots & x_{n-1}^{n-1} \\ 1 & X & X^2 & \ldots & X^{n-1} \end{vmatrix}.$$

C'est un polynôme en X (*cf.* la remarque de la page 353), notons-le P. Pour tout $i \in [\![1, n-1]\!]$, on a $P(x_i) = V(x_1, \ldots, x_{n-1}, x_i) = 0$, *i.e.* x_i est racine de P. Comme $x_1, \ldots, x_{n-1}$ sont distincts, P est multiple du polynôme :

$$Q := (X - x_1)(X - x_2) \cdots (X - x_{n-1})\,.$$

Développons le déterminant donnant P par rapport à sa dernière ligne :

$$P = \sum_{j=1}^{n} (-1)^{n+j} X^{j-1} D_{n,j}.$$

Ainsi P est un polynôme de degré au plus $n-1$; le coefficient de X^{n-1} dans P est $D_{n,n}$, qui vaut $V(x_1, \ldots, x_{n-1})$. L'hypothèse de récurrence donne :

$$D_{n,n} = V(x_1, \ldots, x_{n-1}) = \prod_{1 \leqslant i < j \leqslant n-1} (x_j - x_i) \neq 0\,.$$

Ainsi P est de degré exactement $n-1$, comme Q.

Or Q est unitaire et divise P, donc $P = D_{n,n}Q$. Prenons les valeurs en x_n :

$$P(x_n) = Q(x_n)D_{n,n} = (x_n - x_1)\cdots(x_n - x_{n-1}) \prod_{1\leqslant i<j\leqslant n-1} (x_j - x_i)\,.$$

D'où la formule à l'ordre n, car $P(x_n) = V(x_1, \ldots, x_n)$.

Définition 10. Soit $A \in M_n(K)$, $n \geqslant 2$. Pour tous $i, j \in [\![1, n]\!]$, $(-1)^{i+j}D_{i,j}$ est appelé *cofacteur d'indice* (i, j) de A. La matrice carrée d'ordre n ayant pour coefficients ces cofacteurs est appelée *comatrice* de A.

L'intérêt de la comatrice réside dans le théorème suivant :

Théorème 44. Soient $A \in M_n(K)$ et $\widetilde{A}$ sa *comatrice*. Alors :

$$A \times {}^t\widetilde{A} \;=\; {}^t\widetilde{A} \times A \;=\; \det(A)\, I_n\,. \tag{34}$$

Démonstration. Notons $b_{i,j}$ les coefficients de $B := A \times {}^t\widetilde{A}$. Alors :

$$b_{k,i} \;=\; \sum_{j=1}^{n} a_{k,j}\,({}^t\widetilde{A})_{j,i} \quad \text{pour tous} \quad k, i,$$

où $({}^t\widetilde{A})_{j,i}$ est le coefficient d'indices (j, i) de la matrice ${}^t\widetilde{A}$, c'est-à-dire le coefficient d'indices (i, j) de la matrice $\widetilde{A}$, qui est par définition $(-1)^{i+j}D_{i,j}$. Ainsi :

$$b_{k,i} \;=\; \sum_{j=1}^{n} (-1)^{i+j} a_{k,j} D_{i,j}\,.$$

Si $i = k$, la formule (30) montre que $b_{k,k} = \det(A)$. Si $i \neq k$, la même formule montre que $b_{k,i}$ est le déterminant de la matrice obtenue en remplaçant la $i^{\text{ème}}$ ligne de A par la $k^{\text{ème}}$; cette matrice ayant deux lignes égales (!), son déterminant $b_{k,i}$ est nul. Ainsi, pour tous $k, i \in [\![1, n]\!]$, $b_{k,i}$ vaut $\det A$ si $k = i$ et 0 sinon. D'où $B = \det(A)I_n$. Pour la deuxième égalité, on utilise la formule (31). ∎

Une conséquence essentielle du théorème précédent est une caractérisation des matrices carrées inversibles :

Théorème 45 (Critère d'inversibilité d'une matrice carrée).
Pour qu'une matrice $A \in M_n(K)$ soit inversible, il faut, et il suffit, que son déterminant soit *non nul*. Dans ce cas, l'inverse A^{-1} de A est donné par la formule ci-dessous, où $\widetilde{A}$ est la *comatrice* de A :

$$A^{-1} \;=\; \frac{1}{\det(A)}\; {}^t\widetilde{A}\,. \tag{35}$$

Démonstration. Supposons A inversible, et soit A^{-1} son inverse. L'égalité $A \times A^{-1} = I_n$ donne, vu le théorème 39 de la page 355 : $1 = \det(I_n) = \det(A)\det(A^{-1})$.
Ainsi $\det(A) \neq 0$, et la deuxième formule (35) est vérifiée. Supposons inversement que $\det(A)$ soit non nul. Considérons la matrice $M := (\det A)^{-1}\,{}^t\widetilde{A}$. D'après le théorème 44, $AM = MA = I_n$, ce qui montre que A est inversible et a pour inverse M. ∎

Proposition 46. Étant donné un entier $n \geqslant 1$, l'application $\det : A \mapsto \det(A)$ est un morphisme surjectif (de groupes multiplicatifs) de $GL_n(K)$ dans K^*.

Démonstration. D'après le théorème précédent, $\det(M) \in K^*$ pour toute matrice $M \in GL_n(K)$. Le théorème 39 de la page 355 montre ensuite que l'application déterminant est un morphisme de groupes de $GL_n(K)$ dans K^*.
Si $\lambda \in K^*$ et si $A = \mathrm{Diag}(1, 1, \ldots, \lambda)$, on a $\det(A) = \lambda$, d'où la surjectivité. ■

Remarque. Supposons que K ne soit pas un corps, mais seulement un anneau commutatif. Tous les résultats de cette section *avant* le théorème 45 de la page précédente restent vrais. En modifiant la preuve précédente, on voit que A est inversible si, et seulement si, $\det(A)$ est un élément *inversible* de l'anneau K, auquel cas les formules (35) doivent être remplacées par les suivantes :

$$A^{-1} = \big[\det(A)\big]^{-1}\, {}^t\widetilde{A} \quad \text{et} \quad \det(A^{-1}) = \big[\det(A)\big]^{-1}.$$

Ainsi, soit $A \in M_n(\mathbb{Z})$ une matrice à coefficients entiers, de déterminant non nul. Pour que $A^{-1} \in M_n(\mathbb{Q})$ soit *à coefficients entiers*, il faut, et il suffit, que $\det(A)$ soit un élément inversible de $\mathbb{Z}$, *i.e.* soit égal à ± 1.

Exercice 9.
Inverser la matrice suivante, à coefficients dans $\mathbb{Q}$:

$$A := \begin{pmatrix} 1 & -1 & 1 \\ 2 & 4 & 9 \\ 3 & -4 & 2 \end{pmatrix}.$$

Solution. Calculons les mineurs $D_{i,j}$ de A : $D_{1,1} = \begin{vmatrix} 4 & 9 \\ -4 & 2 \end{vmatrix} = 8 + 36 = 44$, de même $D_{1,2} = -23$, $D_{1,3} = -20$, $D_{2,1} = 2$, $D_{2,2} = -1$, $D_{2,3} = -1$, $D_{3,1} = -13$, $D_{3,2} = 7$, $D_{3,3} = 6$. Développons $\det(A)$ par rapport à sa première ligne :

$$\det(A) = (-1)^{1+1} \times 1 \times 44 + (-1)^{1+2} \times (-1) \times (-23) + (-1)^{1+3} \times 1 \times (-20),$$

soit $\det(A) = 44 - 23 - 20 = 1$. Ainsi A est inversible, et $A^{-1} = {}^t\widetilde{A}$:

$$\widetilde{A} = \begin{pmatrix} 44 & 23 & -20 \\ -2 & -1 & 1 \\ -13 & -7 & 6 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 44 & -2 & -13 \\ 23 & -1 & -7 \\ -20 & 1 & 6 \end{pmatrix}.$$

Le lecteur pourra vérifier que le produit de A par la matrice précédente, censée être égale à A^{-1}, est bien I_3.

Les mineurs d'une matrice A déterminent le rang de A, nous allons le voir. Introduisons une définition. Soit $A_{I,J}$ une matrice extraite de A. On appelle *matrice bordante* de $A_{I,J}$ toute matrice extraite $A_{I',J'}$ de A, où I' est de la forme $I \cup \{i\}$, $i \in [\![1, n]\!] \setminus I$, et J' de la forme $J \cup \{j\}$, $j \in [\![1, p]\!] \setminus J$.

Théorème 47. Soit $A \in M_{n,p}(K)$. Considérons une matrice carrée *inversible* B d'ordre r extraite de A. Alors $\mathrm{rang}(A) \geqslant r$. Le rang de A est égal à r si, et seulement si, B ne possède aucune matrice *bordante* inversible.

Démonstration. Notons $C_1, \dots, C_p$ les colonnes de A. On écrit $B := A_{I,J}$, où

$$I := \{i_1, \dots, i_r\} \quad \text{et} \quad J := \{j_1, \dots, j_r\} \qquad \text{avec} \qquad \begin{array}{c} 1 \leqslant i_1 < \cdots < i_r \leqslant n \\ 1 \leqslant j_1 < \cdots < j_r \leqslant p. \end{array}$$

La matrice $A_{I,J}$ est inversible, donc ses colonnes sont linéairement indépendantes. Ces colonnes sont les images de $C_{j_1}, \dots, C_{j_r}$ par la projection $(x_1, \dots, x_n) \mapsto (x_{i_1}, \dots, x_{i_r})$ de K^n sur K^r, donc $C_{j_1}, \dots, C_{j_r}$ sont linéairement indépendantes (proposition 22 de la page 190), d'où $\operatorname{rang}(A) \geqslant r$. Si B' est une matrice bordante inversible de B, c'est une matrice carrée inversible d'ordre $r+1$ extraite de A, donc $\operatorname{rang}(A) \geqslant r+1$, nous venons de le voir.

Supposons inversement $\operatorname{rang}(A) \geqslant r+1$. Les colonnes de B étant linéairement indépendantes, la famille $(C_{j_1}, \dots, C_{j_r})$ de vecteurs de K^n est libre.

Comme $\dim\big(\operatorname{Vect}(C_1, \dots, C_n)\big) = \operatorname{rang}(A) > r$, le théorème 8 de la page 335 entraîne l'existence d'un indice $j \in [\![1, p]\!] \setminus J$ tel que $(C_{j_1}, \dots, C_{j_r}, C_j)$ soit libre.

La matrice $M := [C_{j_1}; \dots; C_{j_r}; C_j] \in M_{n,r+1}(K)$ est de rang $r+1$. Comme ci-dessus, les lignes de M d'indices $i_1, \dots, i_r$ sont linéairement indépendantes, et il existe un $i \in [\![1, n]\!] \setminus I$ tel que les lignes d'indices $i_1, \dots, i_r, i$ de M forment une base de K^{r+1}. En posant $I' := I \cup \{i\}$ et $J' := J \cup \{j\}$, on voit donc que la matrice $A_{I',J'}$ est une matrice bordante *inversible* de B. ■

Corollaire 48. Le rang d'une matrice A est le maximum des ordres des matrices carrées *inversibles* extraites de A.

Démonstration. Soient t ledit maximum et $A_{I,J}$ une matrice carrée d'ordre t extraite de A et inversible. Elle n'a pas de matrice bordante inversible, par maximalité de t. Ainsi $t = \operatorname{rang}(A)$, vu le théorème précédent. ■

Exercice 10.

Trouver le rang de la matrice A suivante, à coefficients dans $\mathbb{Q}$:

$$A := \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 8 & 9 \end{pmatrix}.$$

Solution. Puisque $a_{1,1} \neq 0$, $\operatorname{rang}(A) \geqslant 1$. Pour voir si $\operatorname{rang}(A) \geqslant 2$, vu le corollaire, il faut voir si A possède un mineur non nul d'ordre 2, il y a 18 tels mineurs (pourquoi ?). Le théorème 47 de la page ci-contre est plus puissant : il suffit de regarder les 6 matrices bordantes $A_{\{1,i\},\{1,j\}}$, où $i \in \{2, 3\}$ et $j \in \{2, 3, 4\}$.

L'une d'elles est la matrice B, de déterminant $1 \times 8 - 3 \times 3 = -1$, d'où $\operatorname{rang}(A) \geqslant 2$.

$$A := \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 8 & 9 \end{pmatrix} \quad \text{et} \quad B := \begin{pmatrix} 1 & 3 \\ 3 & 8 \end{pmatrix}.$$

On peut de même chercher une bordante inversible de B. Il n'y en a pas : notant L_1, L_2, L_3 les lignes de A, on a $L_2 = 2L_1$, d'où $\operatorname{rang}(A) = 2$.

Application : orientation d'un espace vectoriel réel.
Soit E un espace vectoriel *réel* de dimension finie $n \geqslant 1$ (le corps de base est donc $\mathbb{R}$). Étant données deux bases $\mathcal{B}, \mathcal{B}'$ de E, le déterminant de la matrice de passage de $\mathcal{B}$ à $\mathcal{B}'$ est un réel non nul ; nous dirons que $\mathcal{B}, \mathcal{B}'$ sont *de même sens* si $\det\big(\text{Pass}(\mathcal{B}, \mathcal{B}')\big) > 0$ et de sens opposé si $\det\big(\text{Pass}(\mathcal{B}, \mathcal{B}')\big) < 0$. Sur l'ensemble des bases de E, la relation « être de même sens » est une relation d'équivalence : la réflexivité est évidente, la symétrie vient de la proposition 31 de la page 349, la transitivité vient de l'égalité (18).
En pratique, choisissons une base $\mathcal{B}_0$ de E, par exemple la base canonique si $E = \mathbb{R}^n$. Une base $\mathcal{B}$ est dite *directe* si elle est de même sens que $\mathcal{B}_0$ et *indirecte* sinon ; on dit que E est *orienté* par le choix de la base $\mathcal{B}_0$. Sur l'ensemble des bases de E, il y a donc deux classes pour la relation d'équivalence « être de même sens », ce sont par définition les deux *orientations* de E. Orienter E, c'est choisir l'une d'elles.

3.3 Déterminant d'un endomorphisme, déterminant d'une famille de n vecteurs

À partir de la notion de déterminant d'une matrice carrée, nous allons définir deux autres notions de déterminant : le déterminant d'un endomorphisme et le déterminant de n vecteurs relativement à une base.

Théorème et définition 49. Soit f un endomorphisme d'un espace vectoriel E de dimension finie $n \geqslant 1$. Il existe un unique élément de K, noté $\det(f)$, tel que, pour *toute* base $\mathcal{B}$ de E, l'égalité suivante soit vérifiée :

$$\det(f) = \det\big(\text{Mat}_{\mathcal{B}}(f)\big). \tag{36}$$

Ce scalaire $\det(f)$ est appelé *déterminant* de f.

Démonstration. Il s'agit de voir que, si $\mathcal{B}$ est une base de E, $\det\big(\text{Mat}_{\mathcal{B}}(f)\big)$ est *indépendant* du choix de $\mathcal{B}$, *i.e.* ne dépend que de f.
Soient donc $\mathcal{B}$ et $\mathcal{B}'$ deux bases de E, $P := \text{Pass}(\mathcal{B}, \mathcal{B}') \in GL_n(K)$ et $A := \text{Mat}_{\mathcal{B}}(f)$, $A' := \text{Mat}_{\mathcal{B}'}(f)$. La formule (22) de la page 351 donne : $A' = P^{-1}AP$. Ainsi, d'après la proposition 46 de la page 362 et la commutativité de K, $\det(A') = \det\big(P^{-1}AP\big)$ vaut :

$$\det(P)^{-1}\det(A)\det(P) = \big[\det(P)^{-1}\det(P)\big]\det(A) = \det(A). \qquad \blacksquare$$

Pour calculer le déterminant d'un endomorphisme f de E, nous pouvons donc choisir de façon appropriée une base de E, alors $\det(f)$ est le déterminant de la matrice représentant f dans cette base.
Exemples.

1. Soit $\alpha \in K$. Le déterminant de $\alpha\,\text{Id}_E$ est α^n : dans toute base de E, la matrice de $\alpha\,\text{Id}_E$ est αI_n, de déterminant α^n (théorème 36 de la page 353).
2. Reprenons les hypothèses et notations de l'exemple 4 de la page 348. Vu l'égalité $(**)$ de cet exemple : $\det(s) = \det\big(\text{Diag}(\underbrace{1, \ldots, 1}_{r}, \underbrace{-1, \ldots, -1}_{n-r})\big) = (-1)^{n-r}$.

   Dans $\mathbb{R}^2$, une symétrie par rapport à une droite est de déterminant -1. Dans $\mathbb{R}^3$, une symétrie par rapport à un plan (resp. une droite) est de déterminant -1 (resp. 1).

Le résultat suivant montre l'importance de cette notion de déterminant :

Théorème 50. Soit f un endomorphisme d'un espace vectoriel E de dimension finie $n \geqslant 1$. Pour que f soit bijectif (*i.e.* soit un automorphisme de E), il faut, et il suffit, que son déterminant soit non nul. Lorsque c'est le cas, on a :

$$\det\big(f^{-1}\big) \;=\; \big[\det(f)\big]^{-1}. \tag{37}$$

Démonstration. Soient $\mathcal{B}$ une base et $A := \mathrm{Mat}_{\mathcal{B}}(f) \in M_n(K)$. D'après le théorème 29 de la page 347, f est bijectif si, et seulement si, A est inversible, *i.e.* si $\det(A) \neq 0$, auquel cas $\mathrm{Mat}_{\mathcal{B}}(f^{-1}) = A^{-1}$. Vu le théorème 45 de la page 361, A est inversible si, et seulement si, $\det(A) \neq 0$, soit $\det(f) \neq 0$, auquel cas :

$$\det\big(f^{-1}\big) \;=\; \det\big(\mathrm{Mat}_{\mathcal{B}}(f^{-1})\big) \;=\; \det(A^{-1}) \;=\; \big[\det(A)\big]^{-1} \;=\; \big[\det(f)\big]^{-1}. \qquad \blacksquare$$

Proposition 51. Soit E un espace vectoriel de dimension finie $n \geqslant 1$.

1. Pour tous $f, g \in \mathcal{L}(E)$, on a :
$$\det(f \circ g) \;=\; \det(f)\det(g). \tag{38}$$
2. Pour tous $f \in \mathcal{L}(E)$ et $\alpha \in K$, on a : $\det(\alpha f) = \alpha^n \det(f)$.
3. L'application $\det : f \mapsto \det(f)$ de $\mathcal{GL}(E)$ dans K^* est un morphisme de groupes multiplicatifs.

Démonstration. En passant par les matrices dans une base de E fixée, 1 et 3 résultent de la proposition 30 de la page 348 et de la proposition 46 de la page 362. L'assertion 2 résulte de 1, puisque $\det(\alpha\ \mathrm{Id}_E) = \alpha^n$. $\blacksquare$

Passons au déterminant de n vecteurs dans une base donnée.

Définition 11. Soient E un espace vectoriel de dimension finie $n \geqslant 1$ et $\mathcal{B}$ une base de E. Étant donnés n vecteurs $x_1, \ldots, x_n$ de E, on appelle *déterminant de* $x_1, \ldots, x_n$ *dans la base* $\mathcal{B}$ le déterminant de la matrice de la famille $(x_1, \ldots, x_n)$ dans la base $\mathcal{B}$. Ce déterminant est noté $\det_{\mathcal{B}}(x_1, \ldots, x_n)$:

$$\det_{\mathcal{B}}(x_1, \ldots, x_n) \;:=\; \det\big(\mathrm{Mat}_{\mathcal{B}}(x_1, \ldots, x_n)\big).$$

Voyons quelques propriétés de ce déterminant.

Proposition 52. Soient E un espace vectoriel de dimension finie $n \geqslant 1$ et P la matrice de passage d'une base $\mathcal{B}$ à une base $\mathcal{B}'$ de E. Si $x_1, \ldots, x_n \in E$, on a :

$$\det_{\mathcal{B}}(x_1, \ldots, x_n) \;=\; \det(P) \times \det_{\mathcal{B}'}(x_1, \ldots, x_n).$$

Démonstration. Pour tout j, soit X_j (resp. X'_j) la matrice colonne des composantes de x_j dans la base $\mathcal{B}$ (resp. $\mathcal{B}'$). Par définition, on a :

$$\mathrm{Mat}_{\mathcal{B}}(x_1, \ldots, x_n) \;=\; [X_1; \ldots; X_n] \quad \text{et} \quad \mathrm{Mat}_{\mathcal{B}'}(x_1, \ldots, x_n) = [X'_1; \ldots; X'_n].$$

La proposition 32 de la page 350 donne $X_j = PX'_j$ pour tout j, d'où :

$$\mathrm{Mat}_{\mathcal{B}}(x_1, \ldots, x_n) \;=\; P \times \mathrm{Mat}_{\mathcal{B}'}(x_1, \ldots, x_n).$$

Il suffit alors de prendre les déterminants des deux membres. $\blacksquare$

La proposition suivante résulte immédiatement du théorème 36 de la page 353 :

Proposition 53. Soient E un espace vectoriel de dimension finie $n \geqslant 1$, $\mathcal{B}$ une base de E et $x_1, \ldots, x_n$ des vecteurs de E. Posons $d := \det_{\mathcal{B}}(x_1, \ldots, x_n)$. Alors :

1. Soit $j \in [\![1, n]\!]$. Lorsque les x_i pour $i \neq j$ sont fixés, d dépend linéairement de x_j.
2. Si l'un des vecteurs x_j est nul, ou si deux des vecteurs x_j sont égaux, on a $d = 0$.
3. Pour toute permutation $s \in \mathfrak{S}_n$, on a :
$$\det_{\mathcal{B}}(x_{s(1)}, \ldots, x_{s(n)}) \;=\; \varepsilon(s)\det_{\mathcal{B}}(x_1, \ldots, x_n).$$
Si l'on échange deux des vecteurs x_j, d est donc remplacé par $-d$.

Théorème 54. Soient E un espace vectoriel de dimension finie $n \geqslant 1$, $\mathcal{B}$ une base de E et $x_1, \ldots, x_n$ des vecteurs de E. Pour que $(x_1, \ldots, x_n)$ soit une base de E, il faut, et il suffit, que $\det_{\mathcal{B}}(x_1, \ldots, x_n)$ soit non nul.

Démonstration. Soit $A := \mathrm{Mat}_{\mathcal{B}}(x_1, \ldots, x_n)$. Le rang r de A est égal au rang de la famille $(x_1, \ldots, x_n)$ (théorème 21 de la page 343). D'après la proposition 20 de la page 342, $(x_1, \ldots, x_n)$ est une base de E si, et seulement si, $r = n$, *i.e.* si A est de rang n, c'est-à-dire inversible. Comme $\det_{\mathcal{B}}(x_1, \ldots, x_n) = \det(A)$, la conclusion vient du théorème 45 de la page 361. ∎

Exercice 11.
Soient $a_1, \ldots, a_n$ des scalaires deux à deux distincts. Pour tout $i \in [\![1, n]\!]$, posons $x_i := (1, a_i, a_i^2, \ldots, a_i^{n-1}) \in K^n$. Montrer que $(x_1, \ldots, x_n)$ est une base de K^n.
Solution. Soit en effet $\mathcal{B}$ la base canonique de K^n. La matrice de $(x_1, \ldots, x_n)$ dans la base $\mathcal{B}$ est la matrice $A \in M_n(K)$ de coefficients a_i^{j-1}, $(i, j) \in [\![1, n]\!]^2$. Le déterminant de A est le déterminant de Vandermonde $V(a_1, \ldots, a_n)$, il n'est donc pas nul (*cf.* l'exercice 8 de la page 360). Il suffit alors d'appliquer le théorème précédent.

Voici enfin un lien entre les notions de déterminant d'un endomorphisme et de déterminant de n vecteurs.

Théorème 55. Soit f un endomorphisme d'un espace vectoriel E de dimension finie $n \geqslant 1$. Pour toute base $\mathcal{B}$ de E et tous vecteurs $x_1, \ldots, x_n$ de E, on a :
$$\det_{\mathcal{B}}\big(f(x_1), \ldots, f(x_n)\big) \;=\; \det(f) \times \det_{\mathcal{B}}(x_1, \ldots, x_n)\,. \tag{39}$$

Démonstration. Soit $A := \mathrm{Mat}_{\mathcal{B}}(f)$. Par ailleurs, pour tout $j \in [\![1, n]\!]$, soit X_j (resp. Y_j) le vecteur colonne des composantes de x_j $\big($resp. $f(x_j)\big)$ dans la base $\mathcal{B}$. Le théorème 24 de la page 344 montre que $Y_j = AX_j$ pour tout j, donc :
$$\begin{aligned}\mathrm{Mat}_{\mathcal{B}}\big(f(x_1), \ldots, f(x_n)\big) &= [Y_1; \ldots; Y_n] = [AX_1; \ldots; AX_n] \\ &= A[X_1; \ldots; X_n] \;=\; A\,\mathrm{Mat}_{\mathcal{B}}(x_1, \ldots, x_n)\,.\end{aligned}$$
D'où la conclusion, en prenant les déterminants des deux membres. ∎

Terminons par une illustration géométrique des déterminants. Notons $\mathcal{B} := (\vec{\imath}, \vec{\jmath}, \vec{k})$ la base canonique de $\mathbb{R}^3$. Soient u, u' deux vecteurs de $\mathbb{R}^3$, considérés comme vecteurs colonnes :

$$u := \begin{pmatrix} a \\ b \\ c \end{pmatrix} \qquad \text{et} \qquad u' := \begin{pmatrix} a' \\ b' \\ c' \end{pmatrix}.$$

Notons $(u \mid u')$ le produit scalaire de u et u' : $(u \mid u') := aa' + bb' + cc'$.
Ainsi, $\|u\|$ désignant la longueur de u, $\|u\|^2 = (u \mid u) = a^2 + b^2 + c^2$. On définit aussi le *produit vectoriel* $u \wedge u'$ de u et u', utile en physique notamment :

$$u \wedge u' := \begin{pmatrix} bc' - cb' \\ ca' - ac' \\ ab' - ba' \end{pmatrix} \quad \textit{i.e.} \quad \begin{pmatrix} a \\ b \\ c \end{pmatrix} \wedge \begin{pmatrix} a' \\ b' \\ c' \end{pmatrix} := \begin{pmatrix} bc' - cb' \\ ca' - ac' \\ ab' - ba' \end{pmatrix}.$$

Remarquons que les composantes de $u \wedge u'$ sont, dans l'ordre, $\begin{vmatrix} b & b' \\ c & c' \end{vmatrix}$, $-\begin{vmatrix} a & a' \\ c & c' \end{vmatrix}$ et $\begin{vmatrix} a & a' \\ b & b' \end{vmatrix}$. Par exemple, $\vec{\imath} \wedge \vec{\jmath} = \vec{k}$. On vérifie immédiatement *l'identité de Lagrange* suivante (en développant consciencieusement) :

$$\begin{aligned} (a^2 + b^2 + c^2)(a'^2 + b'^2 + c'^2) &= (aa' + bb' + cc')^2 + (bc' - b'c)^2 \\ &\quad + (ca' - ac')^2 + (ab' - ba')^2, \\ \text{soit :} \qquad \|u\|^2 \|u'\|^2 &= (u \mid u')^2 + \|u \wedge u'\|^2 . \end{aligned} \tag{40}$$

Supposons u, u' non nuls, et notons $\theta \in [\![0, \pi]\!]$ l'angle de u et u'.
On sait que $(u|u') = \|u\| \, \|u'\| \cos\theta$, d'où $\|u \wedge u'\| = \|u\| \, \|u'\| \sin\theta$, vu l'égalité ci-dessus. Ainsi $u \wedge u' = 0$ si, et seulement si, u, u' sont colinéaires ($\theta = 0$ ou $\theta = \pi$). Si (u, u') est libre et si $c = c' = 0$, on peut assimiler u, u' à des vecteurs de $\mathbb{R}^2$.
L'égalité $\|u \wedge u'\| = \|u\| \, \|u'\| \sin\theta$ montre alors que $|ab' - ba'|$ est *l'aire géométrique* du parallélogramme construit sur u, u' (le vérifier en exercice).
En général, l'égalité (40) impliquant $\big|(u|u')\big| \leqslant \|u\| \, \|u'\|$, on peut d'ailleurs définir « rigoureusement » l'angle θ par la formule :

$$\theta \ := \ \arccos\left(\frac{(u \mid u')}{\|u\| \, \|u'\|}\right).$$

Définition 12. Soit $u'' = \begin{pmatrix} a'' \\ b'' \\ c'' \end{pmatrix}$ un troisième vecteur de $\mathbb{R}^3$. On appelle *produit mixte* de u, u', u'' (dans cet ordre) et l'on note $[u, u', u'']$ le déterminant :

$$[u, u', u''] \ := \ \det_{\mathcal{B}}(u, u', u'') \ = \ \begin{vmatrix} a & a' & a'' \\ b & b' & b'' \\ c & c' & c'' \end{vmatrix}. \tag{41}$$

Indiquons quelques propriétés simples du produit mixte.

1. Le produit mixte $[u, u', u'']$ est nul si, et seulement si, les vecteurs u, u', u'' sont *liés* : cela résulte du théorème 54 de la page 366.
2. Si $[u, u', u''] \neq 0$, $\mathcal{C} = (u, u', u'')$ est une base de $\mathbb{R}^3$, elle est directe si $[u, u', u''] > 0$ et indirecte si $[u, u', u''] < 0$.
3. Si l'on échange deux des trois vecteurs u, u', u'', $[u, u', u'']$ change de signe. Par contre $[u, u', u'']$ est *invariant par permutation circulaire* :
$$[u, u', u''] = [u', u'', u] = [u'', u, u'].$$
Cela résulte de la propriété 5 du théorème 36 de la page 353, puisqu'une transposition est de signature -1 et le 3-cycle $(1\ 2\ 3)$ est de signature $+1$.
4. Si deux des vecteurs u, u', u'' sont fixés, $[u, u', u'']$ dépend linéairement du troisième (propriété 1 du théorème 36 de la page 353).
5. Développons le déterminant ci-après par rapport à sa troisième colonne :
$$\begin{vmatrix} a & a' & a'' \\ b & b' & b'' \\ c & c' & c'' \end{vmatrix} = a'' \begin{vmatrix} b & b' \\ c & c' \end{vmatrix} - b'' \begin{vmatrix} a & a' \\ c & c' \end{vmatrix} + c'' \begin{vmatrix} a & a' \\ b & b' \end{vmatrix}.$$
Du calcul des composantes de $u \wedge u'$, on déduit la formule suivante, reliant produit scalaire, produit vectoriel et produit mixte :
$$[u, u', u''] = \big(u \wedge u' \mid u''\big). \tag{42}$$
En prenant $u'' := u$ ou $u'' := u'$, on voit que $u \wedge u'$ est orthogonal à u et u'. En prenant $u'' := u \wedge u'$, il vient $\|u \wedge u'\|^2 = [u, u', u \wedge u']$. Si (u, u') est libre, $(u, u', u \wedge u')$ est donc une base *directe* de $\mathbb{R}^3$.
6. Supposons $[u, u', u''] \neq 0$. On montre que $\big|[u, u', u'']\big|$ est le *volume* du parallélépipède construit sur les vecteurs u, u', u''.

3.4 Valeurs propres et vecteurs propres

Cette section constitue une introduction aux notions de valeur propre et de vecteur propre d'un endomorphisme ou d'une matrice carrée. Ces notions feront l'objet d'une étude plus approfondie en L2.

Soient E un K-espace vectoriel non trivial et u un endomorphisme de E. Pour l'instant, E n'est pas nécessairement de dimension finie. Un sous-espace vectoriel V de E est dit *stable* ou *invariant* par u si $u(V) \subset V$. Cette notion sera utile en L2, à propos de la réduction des endomorphismes.

Exemples.

1. Soit u l'endomorphisme $(x, y, z) \mapsto (y, z, x)$ de K^3.
Le plan d'équation $x + y + z = 0$ est stable par u, la droite engendrée par $(1, 1, 1)$ aussi.

2. Soient V, W deux sous-espaces vectoriels supplémentaires de E et $p \in \mathcal{L}(E)$ le projecteur sur V parallèlement à W.
   Alors V et W sont stables par p, puisque $\forall x \in V,\ p(x) = x$ et $\forall x \in W,\ p(x) = 0$.

Lorsque V est une droite, V est stable par u si, et seulement si, pour tout vecteur $x \in V$, $u(x)$ est colinéaire à x. D'où la définition suivante :

Définition 13.

1. Soit x un vecteur *non nul* de E. On dit que x est un *vecteur propre* de u si $u(x)$ est colinéaire à x.
2. Soit $\lambda \in K$. On dit que λ est une *valeur propre* de u si l'endomorphisme $u - \lambda\ \mathrm{Id}_E$ de E n'est pas injectif. Dans ce cas, le noyau de $u - \lambda\ \mathrm{Id}_E$ est appelé *sous-espace propre de u associé à λ*.

Soit x un vecteur propre de u. *Il est essentiel de se rappeler que x est alors non nul*, par définition. Puisque $u(x)$ est colinéaire à x et $x \neq 0$, il existe un *unique* $\lambda \in K$ tel que $u(x) = \lambda x$. Ainsi $x \in \mathrm{Ker}(u - \lambda\ \mathrm{Id}_E)$, en particulier $\mathrm{Ker}(u - \lambda\ \mathrm{Id}_E) \neq \{0\}$, autrement dit λ est valeur propre de u ; nous dirons que x est un *vecteur propre de u associé à la valeur propre λ*.
Soit λ une valeur propre de u. Notons $E_\lambda(u)$ l'espace propre de u associé à λ. Par définition, $u(x) = \lambda x$ pour tout $x \in E_\lambda(u)$. Ainsi $E_\lambda(u)$ est formé de 0 et des vecteurs propres de u associés à λ. Noter qu'un vecteur propre de u donné appartient à un unique sous-espace propre de u, à savoir $E_\lambda(u)$, où λ est l'unique scalaire tel que $u(x) = \lambda x$.

Exemples.

1. Supposons que u soit une homothétie : $u := \alpha\ \mathrm{Id}_E$, où $\alpha \in K$. Par définition même, α est l'unique valeur propre de u, et l'espace propre correspondant est E : tout vecteur non nul de E est vecteur propre de u. Cette propriété *caractérise* les homothéties. En effet, la preuve du théorème 42 de la page 200 montre que, si $u \in \mathcal{L}(E)$ est tel que tout vecteur non nul de E soit vecteur propre de u, alors u est une homothétie.
2. Pour tout $a \in \mathbb{R}$, notons f_a la fonction $x \mapsto e^{ax}$ de $\mathbb{R}$ dans $\mathbb{R}$. Elle est dérivable, et $f_a' = a f_a$. Soit E le sous-espace vectoriel de $\mathcal{F}(\mathbb{R}, \mathbb{R})$ engendré par les f_a, $a \in \mathbb{R}$. Il est clair que toute fonction $f \in E$ est dérivable et que $f' \in E$. L'application $u : f \mapsto f'$ est donc un endomorphisme de E. Si $a \in \mathbb{R}$, f_a est vecteur propre de u associé à la valeur propre a. Ainsi tous les réels sont des valeurs propres de u.
3. Soient V, W deux sous-espaces vectoriels supplémentaires non triviaux de E et $p \in \mathcal{L}(E)$ la projection sur V parallèlement à W. On a $p(v) = v$ pour tout $v \in V$ et $p(w) = 0$ pour tout $w \in W$. Ainsi 1 et 0 sont valeurs propres de p, les sous-espaces propres correspondants étant V et W : $V = \mathrm{Ker}(p - \mathrm{Id}_E)$ et $W = \mathrm{Ker}(p)$.
   Si $\mathrm{car}(K) \neq 2$, soit $s \in \mathcal{L}(E)$ la symétrie par rapport à V parallèlement à W. Ici, $V = \mathrm{Ker}(s - \mathrm{Id}_E)$ et $W = \mathrm{Ker}(s + \mathrm{Id}_E)$, donc 1 et -1 sont valeurs propres de s. Le lecteur vérifiera que $1, 0$ (respectivement $1, -1$) sont les seules valeurs propres de p (respectivement de s).

Les *éléments propres*, *i.e.* les valeurs propres et vecteurs propres, peuvent aussi être définis pour une matrice *carrée* :

Définition 14. Soit $A \in M_n(K)$, $n \geqslant 1$.

1. On appelle *vecteur propre de* A tout vecteur colonne non nul $X \in M_{n,1}(K)$ tel que AX soit colinéaire à X.
2. Notons u l'endomorphisme $X \mapsto AX$ de K^n $\big($on identifie K^n à $M_{n,1}(K)\big)$.
   On appelle *valeur propre de* A toute valeur propre de u.

Étant donné un endomorphisme u d'un espace vectoriel $E \neq \{0\}$, ou bien une matrice $A \in M_n(K)$, deux problèmes naturels se posent. Comment déterminer les valeurs propres de u (ou A) ? D'autre part, une telle valeur propre λ étant fixée, comment déterminer l'espace propre correspondant ?
En dimension finie, nous verrons que le premier problème se ramène à la recherche des racines d'un certain polynôme, le second problème se ramenant à la résolution d'un système linéaire. Voici notre point de départ :

Proposition 56.

1. Soit u un endomorphisme d'un espace vectoriel E de dimension finie $n \geqslant 1$.
   Un scalaire λ est valeur propre de u si, et seulement si :
$$\det(u - \lambda \operatorname{Id}_E) = 0. \tag{43}$$
2. Soit de même $A \in M_n(K)$, $n \geqslant 1$. Un scalaire λ est valeur propre de A si, et seulement si :
$$\det(A - \lambda I_n) = 0. \tag{44}$$

Démonstration. Prouvons 1. Par définition, λ est valeur propre de u si, et seulement si, $u - \lambda \operatorname{Id}_E \in \mathcal{L}(E)$ n'est pas injectif. D'après le théorème 50 de la page 365, cela revient à dire que le déterminant de $u - \lambda \operatorname{Id}_E$ est nul.
Pour 2, soit u l'endomorphisme $X \mapsto AX$ de K^n. Par définition, λ est valeur propre de A si, et seulement s'il est valeur propre de u, ce qui revient, d'après 1, à dire que $\det(u - \lambda \operatorname{Id}_{K^n}) = 0$. Or $A - \lambda I_n$ est la matrice de $u - \lambda \operatorname{Id}_{K^n}$ dans la base canonique de K^n, donc le théorème et définition 49 de la page 364 donne :
$$\det(u - \lambda \operatorname{Id}_{K^n}) = \det(A - \lambda I_n).$$
L'assertion 2 en résulte. ∎

Remarque. Soient E un K-espace vectoriel de dimension finie $n \geqslant 1$, $\mathcal{B} := (e_1, \ldots, e_n)$ une base de E, $u \in \mathcal{L}(E)$ et $A := \operatorname{Mat}_{\mathcal{B}}(u)$. Soit $\lambda \in K$. La matrice de $u - \lambda \operatorname{Id}_E$ dans la base $\mathcal{B}$ est $A - \lambda I_n$. Vu la proposition précédente, λ est valeur propre de u si, et seulement s'il est valeur propre de A. Un vecteur colonne $X := {}^t(x_1, \ldots, x_n)$ est vecteur propre de A associé à λ si, et seulement si, $x = \sum_{j=1}^{n} x_j e_j$ est vecteur propre de u associé à λ.

L'assertion 2 de la proposition 56 de la page précédente motive la définition suivante :

Définition 15. Soit $A \in M_n(K)$, $n \geqslant 1$. On appelle *polynôme caractéristique de* A le polynôme $\chi_A \in K[X]$ défini par la formule :

$$\chi_A := \det(XI_n - A)\,. \tag{45}$$

La matrice carrée $XI_n - A$ est à coefficients dans l'anneau des polynômes $K[X]$. Ainsi le déterminant de cette matrice est bien défini, et il appartient à $K[X]$, d'après la remarque de la page 353. Cela justifie la définition. En notant $a_{i,j}$ les coefficients de A, le polynôme χ_A s'écrit :

$$\chi_A(X) = \begin{vmatrix} X - a_{1,1} & -a_{1,2} & \dots & -a_{1,n} \\ -a_{2,1} & X - a_{2,2} & \dots & -a_{2,n} \\ \vdots & \vdots & & \vdots \\ -a_{n,1} & -a_{n,2} & \dots & X - a_{n,n} \end{vmatrix}. \tag{46}$$

Certains prennent comme définition $\chi_A := \det(A - XI_n)$, ce qui revient à multiplier par $(-1)^n$ le polynôme χ_A défini ci-dessus.
Voici des propriétés simples du polynôme caractéristique :

Proposition 57. Soient $A \in M_n(K)$ et χ_A son polynôme caractéristique. Alors :

1. Pour tout $\lambda \in K$, on a : $\chi_A(\lambda) = \det(\lambda I_n - A)$. En particulier, le terme constant de χ_A est : $\chi_A(0) = (-1)^n \det(A)$.
2. Le polynôme χ_A est de degré n et unitaire : son terme de plus haut degré est X^n.
3. Le coefficient de X^{n-1} dans χ_A est : $-\big(a_{1,1} + a_{2,2} + \cdots + a_{n,n}\big)$, en notant $a_{i,j}$ les coefficients de A.

Démonstration. Les coefficients $b_{i,j}$ de la matrice $B := XI_n - A$ appartiennent à $K[X]$. Par définition, $\chi_A = \det(B)$. Soit $\lambda \in K$. L'application $P \mapsto P(\lambda)$ est un morphisme d'anneaux f de $K[X]$ dans K. Avec les notations de la remarque de la page 353, il vient :

$$\chi_A(\lambda) = f(\chi_A) = f\big(\det(B)\big) = \det\big(f(B)\big).$$

Mais $f(B) = \lambda I_n - A$, d'où la première formule en 1.
Pour l'assertion 2, partons de la définition d'un déterminant :

$$\chi_A = \det(B) = \sum_{s \in \mathfrak{S}_n} \varepsilon(s)\, b_{s(1),1} b_{s(2),2} \cdots b_{s(n),n}\,. \tag{$*$}$$

Dans la somme donnant $\det(B)$, mettons à part le terme c correspondant à $s = \mathrm{Id}_{[\![1,n]\!]}$: $c := (X - a_{1,1})(X - a_{2,2}) \cdots (X - a_{n,n})$. Comme polynôme en X, c est unitaire de degré n. Si $s \neq \mathrm{Id}_{[\![1,n]\!]}$, il existe un indice i tel que $s(i) \neq i$, alors $b_{s(i),i} = -a_{s(i),i}$ est constant, *i.e.* de degré au plus 0. Le terme correspondant à s est donc de degré au plus $n-1$. D'où l'assertion 2, χ_A étant la somme de X^n et de termes de degré au plus $n-1$.
L'assertion 3 utilise la formule (28) du module II.2 (*cf.* l'exercice II.7.23 de la page 386). ■

Théorème 58. Soient $A \in M_n(K)$, $n \geqslant 1$ et χ_A le polynôme caractéristique de A.

1. Les valeurs propres de A sont les racines de χ_A dans K.
2. La matrice A possède *au plus* n valeurs propres.
3. Si $K = \mathbb{C}$, la matrice A possède *au moins* une valeur propre.

Démonstration. D'après la proposition 56 de la page 370, un scalaire λ est valeur propre de A si, et seulement si, $\det(A - \lambda I_n) = 0$, ce qui revient à dire que $\chi_A(\lambda) = 0$, *i.e.* que λ est racine de χ_A. Cela prouve l'assertion 1.

Un polynôme de degré n à coefficients dans K a au plus n racines dans K ; l'assertion 2 résulte donc de l'assertion 1, car le degré de χ_A est n.

Prouvons de même 3 : $\chi_A \in \mathbb{C}[X]$ étant de degré $n \geqslant 1$, le théorème de d'Alembert-Gauß montre qu'il a au moins une racine dans $\mathbb{C}$. ■

Exemples.

1. Soit $A \in M_n(K)$ une matrice triangulaire et $a_{i,j}$ ses coefficients.

La matrice $XI_n - A \in M_n\big(K[X]\big)$ est encore triangulaire. D'après le théorème 40 de la page 356, son déterminant est le produit de ses coefficients diagonaux :

$$\chi_A \;=\; (X - a_{1,1})(X - a_{2,2}) \cdots (X - a_{n,n})\,.$$

En particulier, si $A := \alpha\, I_n$ est une matrice d'homothétie, $\chi_A = (X - \alpha)^n$.

2. Soit $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(K)$. La proposition 57 de la page précédente donne :

$$\chi_A \;=\; X^2 - (a + d)X + (ad - bc)\,.$$

Exercice 12.

Chercher les valeurs propres de la matrice $A \;:=\; \begin{pmatrix} 3 & -2 & -3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \end{pmatrix} \in M_3(\mathbb{C})$.

Solution. Développons $\chi_A = \begin{vmatrix} X-3 & 2 & 3 \\ -2 & X-1 & 3 \\ 0 & 1 & X-1 \end{vmatrix}$ par rapport à sa dernière ligne :

$$\begin{aligned}
\chi_A &= -\begin{vmatrix} X-3 & 3 \\ -2 & 3 \end{vmatrix} + (X-1)\begin{vmatrix} X-3 & 2 \\ -2 & X-1 \end{vmatrix} \\
&= -\big[3(X-3)+6\big] + (X-1)\big[(X-3)(X-1)+4\big] \\
&= -3(X-1) + (X-1)\big[X^2 - 4X + 7\big] = (X-1)\big[(X^2-4X+7)-3\big] \\
&= (X-1)(X^2-4X+4) \;=\; (X-1)(X-2)^2\,.
\end{aligned}$$

Ainsi A a deux valeurs propres, à savoir 1 et 2.

Pour passer aux endomorphismes, nous avons besoin d'un résultat :

Proposition 59. Soient $A \in M_n(K)$ et $P \in GL_n(K)$, $n \geqslant 1$.
Posons $A' := P^{-1}AP$. Les matrices A et A' ont le même polynôme caractéristique.

Démonstration. Considérons les matrices $B := XI_n - A$, $B' := XI_n - A'$, à coefficients dans $K[X]$. La matrice d'homothétie XI_n commute avec P, d'où :

$$P^{-1}BP = P^{-1}(XI_n - A)P = P^{-1}(XI_n)P - P^{-1}AP = (P^{-1}P)(XI_n) - A',$$

soit $P^{-1}BP = XI_n - A' = B'$. On en déduit :

$$\chi_{A'} = \det(B') = \det(P^{-1}BP) = \det(P)^{-1}\det(B)\det(P) = \det(B) = \chi_A. \qquad \blacksquare$$

Théorème et définition 60. Soient E un espace vectoriel de dimension finie $n \geqslant 1$ et $u \in \mathcal{L}(E)$. Soient $\mathcal{B}$ une base de E et A la matrice de u dans la base $\mathcal{B}$. Le polynôme caractéristique de A est *indépendant* du choix de la base $\mathcal{B}$, il ne dépend que de u. On l'appelle *polynôme caractéristique de* u et on le note χ_u.

Démonstration. Soient en effet $\mathcal{B}'$ une autre base de E et A' la matrice de u dans cette base. Notant P la matrice de passage de $\mathcal{B}$ à $\mathcal{B}'$, on sait que $A' = P^{-1}AP$. La proposition précédente montre donc que $\chi_{A'} = \chi_A$. $\blacksquare$

Retenons de ce théorème la propriété suivante :

PROPRIÉTÉ. *Pour* toute *matrice A représentant u dans une certaine base, le polynôme caractéristique de u est égal au polynôme caractéristique de A.*

Le polynôme caractéristique d'un endomorphisme a les mêmes propriétés que le polynôme caractéristique d'une matrice. Ainsi :

Proposition 61. Soient u un endomorphisme d'un espace vectoriel E de dimension finie $n \geqslant 1$ et χ_u son polynôme caractéristique. Alors :

1. Pour tout $\lambda \in K$, on a : $\chi_u(\lambda) = \det(\lambda\,\mathrm{Id}_E - u)$.
   En particulier, le terme constant de χ_u est : $\chi_u(0) = (-1)^n \det(u)$.
2. Le polynôme χ_u est unitaire de degré n.

Démonstration. Soit $A \in M_n(K)$ la matrice représentant u dans une certaine base $\mathcal{B}$ de E. Nous savons que $\chi_u = \chi_A$. La conclusion vient alors de la proposition 57 de la page 371 et du théorème et définition 49 de la page 364, à condition de noter que, si $\lambda \in K$, $\lambda I_n - A$ est la matrice de $\lambda\,\mathrm{Id}_E - u$ dans la base $\mathcal{B}$. $\blacksquare$

Théorème 62. Soient u un endomorphisme d'un espace vectoriel E de dimension finie $n \geqslant 1$ et χ_u son polynôme caractéristique. Alors :

1. Les valeurs propres de u sont les racines de χ_u dans K.
2. L'endomorphisme u possède *au plus* n valeurs propres.
3. Si $K = \mathbb{C}$, l'endomorphisme u possède *au moins* une valeur propre.

Démonstration. Soit A comme dans la preuve précédente, de sorte que $\chi_u = \chi_A$. D'après la proposition 56 de la page 370, u et A ont les mêmes valeurs propres. Il suffit donc d'appliquer le théorème 58 de la page précédente pour conclure. $\blacksquare$

Voyons maintenant comment déterminer les vecteurs propres d'une matrice $A \in M_n(K)$ de coefficients $a_{i,j}$. Le théorème 58 de la page 372 montre que les valeurs propres de A sont exactement les racines du polynôme χ_A dans K, il y en a au plus n. Soit λ une telle valeur propre. L'espace propre de A associé à λ est le noyau de l'endomorphisme $v : X \longmapsto (\lambda I_n - A)X$ de $K^n = M_{n,1}(K)$, il est donc formé des vecteurs colonnes X tels que $(\lambda I_n - A)X = 0$. Si l'on écrit $X := {}^t(x_1, \ldots, x_n)$, l'égalité précédente se traduit par le système :

$$\begin{cases} (\lambda - a_{1,1})x_1 - a_{1,2}x_2 - \cdots - a_{1,n}x_n = 0 \\ -a_{2,1}x_1 + (\lambda - a_{2,2})x_2 - \cdots - a_{2,n}x_n = 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ -a_{n,1}x_1 - a_{n,2}x_2 - \cdots + (\lambda - a_{n,n})x_n = 0\,. \end{cases} \tag{47}$$

Les systèmes linéaires feront l'objet de la section suivante. Nous nous contenterons donc ici d'un résultat théorique simple et d'un exemple.

Proposition 63. Soit $\lambda \in K$ une valeur propre de la matrice $A \in M_n(K)$.
La dimension de l'espace propre de A associé à λ est $n - \mathrm{rang}(\lambda I_n - A)$.

Démonstration. Soit $V \subset K^n$ cet espace propre. C'est le noyau de l'endomorphisme $v : X \longmapsto (\lambda I_n - A)X$ de K^n. D'après le théorème du rang, $\dim(V) = n - \mathrm{rang}(v)$. La matrice de v dans la base canonique de K^n étant $\lambda I_n - A$, le théorème 22 de la page 344 montre que le rang de v est égal au rang de $\lambda I_n - A$, et la proposition en résulte. ■

Exercice 13.

Déterminer les espaces propres de la matrice :

$$A := \begin{pmatrix} 3 & -2 & -3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \end{pmatrix} \in M_3(\mathbb{C})\,.$$

Solution. Dans l'exercice 12 de la page 372, nous avons montré que A a deux valeurs propres : 1 et 2. Cherchons les espaces propres V_1 et V_2 correspondants.

$$I_3 - A = \begin{pmatrix} -2 & 2 & 3 \\ -2 & 0 & 3 \\ 0 & 1 & 0 \end{pmatrix}, \quad 2I_3 - A = \begin{pmatrix} -1 & 2 & 3 \\ -2 & 1 & 3 \\ 0 & 1 & 1 \end{pmatrix}.$$

Chacune de ces matrices est de rang 2 : son rang est au plus 2 par définition d'une valeur propre, et ses deux dernières lignes ne sont pas colinéaires. Chacun des espaces V_1, V_2 est donc une droite. Écrivons ici $X := {}^t(x, y, z)$. Pour V_1, l'égalité $(I_3 - A)X = 0$ équivaut au système : $-2x + 3z = 0$ et $y = 0$. On en déduit que V_1 est la droite $\mathrm{Vect}(X_1)$, où $X_1 := {}^t(3, 0, 2)$.
Pour V_2, l'égalité $(2I_3 - A)X = 0$ équivaut au système : $-2x + y + 3z = 0$ et $y + z = 0$. On en déduit que V_2 est la droite engendrée par le vecteur $X_2 := {}^t(1, -1, 1)$.

Soit $A \in M_n(\mathbb{R})$. Par valeur propre de A, nous entendons en général une valeur propre *complexe* de A, ce qui a un sens puisque $A \in M_n(\mathbb{C})$. Par contre, soient E un espace vectoriel réel de dimension finie $n \geqslant 1$ et $f \in \mathcal{L}(E)$. Une valeur propre de f est *par définition* un réel. Si besoin est, on peut parler de racines complexes du polynôme caractéristique $\chi_f \in \mathbb{R}[X]$. On peut aussi (et c'est la même chose !) parler de valeur propre complexe de la matrice de f dans une base donnée de E.

4 Systèmes linéaires

4.1 Équations linéaires

Soient E, F deux espaces vectoriels sur un corps K et f une application linéaire de E dans F. Étant donné un élément b de F, l'équation suivante :

$$f(x) \;=\; b\,, \tag{48}$$

est dite *équation linéaire* (parce que f est linéaire). Dans cette équation, l'application linéaire f est donnée, ainsi que $b \in F$, appelé *second membre* ; si $b = 0$, l'équation est dite *homogène*. Le vecteur $x \in E$ est *l'inconnue*. Notons $\mathcal{S} := \{x \in E \mid f(x) = b\}$ l'ensemble des solutions de l'équation (48). Résoudre cette équation, c'est déterminer $\mathcal{S}$. Donnons d'abord des propriétés simples de $\mathcal{S}$. La définition d'une application injective (surjective, bijective) a la conséquence utile suivante (*cf.* le module II.3) :

PROPRIÉTÉS.

1. *L'application f est surjective si, et seulement si,* pour tout $b \in F$, *l'équation* (48) *a* au moins *une solution.*
2. *L'application f est injective si, et seulement si,* pour tout $b \in F$, *l'équation* (48) *a* au plus *une solution.*
3. *L'application f est bijective si, et seulement si,* pour tout $b \in F$, *l'équation* (48) *a* exactement *une solution.*

Avant de poursuivre, donnons une définition (*cf.* le module III.1).

Définition 16. Une partie $\mathcal{V}$ de E est appelée *sous-espace affine* s'il existe un sous-espace vectoriel V de E et un point a de E tels que :

$$\mathcal{V} \;=\; a + V\,, \quad \text{où} \quad a + V \;:=\; \{a + v \mid v \in V\}\,. \tag{49}$$

On dit alors que $\mathcal{V}$ est le sous-espace affine de *direction* V passant par a.
Si V est de dimension finie m, $\mathcal{V}$ est appelé *sous-espace affine de dimension* m de E.

Soit $\mathcal{V}$ un sous-espace affine de E : il existe un point $a \in E$ et un sous-espace vectoriel V de E tels que $\mathcal{V} = a + V$. D'abord $\mathcal{V}$ passe bien par a : $0 \in V$, donc $a = a + 0 \in \mathcal{V}$.
Si $a' \in \mathcal{V}$, on a $\mathcal{V} = a' + V$. En effet, il existe $v_0 \in V$ tel que $a' = a + v_0$. Soit $v \in V$.
D'abord $a' + v = a + (v_0 + v)$, et $v_0 + v \in V$, donc $a' + V \subset a + V$.
De même $a + v = a' + (v - v_0)$, et $v - v_0 \in V$, donc $a + V \subset a' + V$.
Ainsi $a' + V = a + V = \mathcal{V}$: l'égalité (49) est en fait vraie *pour tout* $a \in \mathcal{V}$.

Le sous-espace vectoriel V est unique : si $a \in \mathcal{V}$, l'égalité (49) donne :

$$V \;=\; \mathcal{V} - a \;:=\; \big\{x - a \mid x \in \mathcal{V}\big\}\,. \tag{50}$$

Soit $\mathcal{V}$ un sous-espace affine de E de dimension m.
Si $m = 0$, $\mathcal{V}$ est un point (un singleton).
Si $m = 1$ (resp. 2), $\mathcal{V}$ est appelé *droite affine* (resp. *plan affine*) de E.
Si la direction V de $\mathcal{V}$ est un hyperplan vectoriel (*cf.* la définition 12 de la page 181) de E, $\mathcal{V}$ est appelé *hyperplan affine* de E.

Exemples.

1. Un sous-espace affine de E passant par 0 est simplement un sous-espace *vectoriel* de E.

2. Soient $x, y \in E$, $x \neq y$. Posons $\mathcal{D} := x + K(y - x)$. Alors $\mathcal{D}$ est une droite affine passant par x et y, sa direction est la droite vectorielle $D := K(y - x)$. En fait $\mathcal{D}$ est la seule droite affine passant par x et y : si Δ est la direction d'une telle droite, l'égalité (50) montre que $y - x \in \Delta$, d'où $\Delta = K(y - x)$ puisque $\dim(\Delta) = 1$.

3. Soit $(\mathcal{V}_i)_{i \in I}$ une famille de sous-espaces affines de E, d'intersection $\mathcal{V}$. Si $i \in I$, soit V_i la direction de $\mathcal{V}_i$. Alors, ou bien $\mathcal{V}$ est vide, ou bien c'est un sous-espace affine de E, de direction $V := \bigcap\limits_{i \in I} V_i$.

   En effet, supposons que $a \in \mathcal{V}$. Un point $x \in E$ appartient à $\mathcal{V}$ si, et seulement si, pour tout $i \in I$, $x \in \mathcal{V}_i = a + V_i$, soit $x - a \in V_i$. D'où $\mathcal{V} = a + V$.

4. Vu la définition 12 de la page 181, tous les hyperplans affines de E s'obtiennent ainsi : partant d'une *forme linéaire* non nulle $\varphi : E \to K$ et d'un point $a \in E$, on pose :

   $$\mathcal{V} \;:=\; a + \operatorname{Ker}\varphi\,. \tag{51}$$

   Supposons E de dimension finie $n \geqslant 1$, et soit $\mathcal{B} := (e_1, \ldots, e_n)$ une base de E. Posons $a_i := \varphi(e_i)$ pour $i = 1, \ldots, n$ et $b := \varphi(a)$.

   Soit maintenant $x \in E$, notons $x_1, \ldots, x_n$ ses composantes dans la base $\mathcal{B}$. Alors $x \in \mathcal{V}$ si, et seulement si, $x - a \in \operatorname{Ker}\varphi$, soit $\varphi(x) = b$. Or :

   $$\varphi(x) = \varphi\Big(\sum_{i=1}^{n} x_i e_i\Big) = \sum_{i=1}^{n} x_i \varphi(e_i) = \sum_{i=1}^{n} a_i x_i.$$

   Ainsi $x \in \mathcal{V}$ si, et seulement si, l'on a :

   $$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \;=\; b\,, \quad \text{soit} \quad \sum_{i=1}^{n} a_i x_i \;=\; b\,. \tag{52}$$

   En d'autres termes, (52) est une *équation de* $\mathcal{V}$.

Concernant l'équation linéaire (48), voici un théorème important.

Théorème 64. Soit $\mathcal{S} \subset E$ l'ensemble des solutions de l'équation (48).

1. Ou bien $\mathcal{S}$ est vide, ou bien c'est un sous-espace affine de E, la direction de $\mathcal{S}$ étant alors le noyau de f.
2. Supposons E de dimension finie n. Si $\mathcal{S}$ n'est pas vide, c'est un sous-espace affine de E de dimension $n - \operatorname{rang}(f)$.

Démonstration. Soit donc $a \in \mathcal{S}$. Pour tout $x \in E$, on a les équivalences :

$$(x \in \mathcal{S}) \iff \big(f(x) = b\big) \iff \big(f(x) = f(a)\big) \iff \big(f(x-a) = 0\big)$$
$$\iff \big(x - a \in \operatorname{Ker}(f)\big) \iff \big(x \in a + \operatorname{Ker}(f)\big)\,.$$

La troisième équivalence vient de la linéarité de f.
Ainsi $\mathcal{S} = a + \operatorname{Ker}(f)$, ce qui établit l'assertion 1.
L'assertion 2 vient alors du théorème du rang, puisque $\dim(\operatorname{Ker} f) = n - \operatorname{rang}(f)$. ■

4.2 Systèmes linéaires

La résolution pratique des systèmes linéaires a été détaillée dans le module II.4, section 2.2. Certains résultats théoriques, concernant la notion de dimension, avaient alors été admis. Ces résultats étant acquis, reprenons l'étude de ces systèmes, d'un point de vue théorique plutôt qu'algorithmique.
Rappelons que K est un corps commutatif fixé. Considérons un système :

$$(S) : \begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,p}x_p = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,p}x_p = b_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,p}x_p = b_n\,. \end{cases} \tag{53}$$

Un tel système est appelé *système linéaire de n équations à p inconnues* $(n, p \geqslant 1)$.
Les scalaires $a_{i,j}, (i,j) \in [\![1,n]\!] \times [\![1,p]\!]$ et $b_1, \ldots, b_n$ sont donnés, ce sont les coefficients du système (ils appartiennent à K). Le système est dit *homogène* (ou « sans second membre ») si tous les b_j sont nuls. Le système :

$$(S_0) : \begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,p}x_p = 0 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,p}x_p = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,p}x_p = 0 \end{cases} \tag{54}$$

est appelé système homogène *associé* au système (S).
Une *solution* du système (S) est un p-uplet $(x_1, \ldots, x_p) \in K^p$ vérifiant les n équations (53). On note $\mathcal{S} \subset K^p$ l'ensemble des solutions du système (S). *Résoudre* ce système, c'est déterminer $\mathcal{S}$. L'ensemble $\mathcal{S}_0$ des solutions du système homogène (S_0) contient toujours $0 = (0, 0, \ldots, 0) \in K^p$, c'est la solution *triviale*, les autres solutions sont dites *non triviales*.

La matrice $A \in M_{n,p}(K)$ de coefficients $a_{i,j}$ est la *matrice* du système, notons $C_1, \ldots, C_p$ ses colonnes et $L_1, \ldots, L_n$ ses lignes. Le rang r de A est le *rang* du système ; on a $r \leqslant \min(n, p)$. Le vecteur colonne $B := {}^t(b_1, \ldots, b_n)$ est appelé *second membre* du système. En notant $X := {}^t(x_1, \ldots, x_p)$ (vecteur colonne inconnu), le système (S) prend la forme matricielle :

$$AX = B. \tag{55}$$

C'est là une *équation linéaire* : en notant $f : X \mapsto AX$ l'application linéaire de K^p dans K^n canoniquement associée à A, l'équation (55) devient :

$$f(X) = B. \tag{56}$$

Ainsi les n équations (53) sont remplacées par la *seule* équation linéaire (vectorielle) (56) ; le langage géométrique apparaît dans l'équation (56), le langage matriciel apparaît dans l'équation (55).
En vertu du théorème 22 de la page 344, le rang de f est égal à r (rang de A). Le théorème 64 de la page précédente a donc la conséquence importante suivante :

Théorème 65.

1. L'ensemble $\mathcal{S}_0$ des solutions du système homogène (S_0) est un sous-espace vectoriel de dimension $p - r$ de K^p, r étant le rang de A.
2. L'ensemble $\mathcal{S}$ des solutions du système (S) est soit vide, soit est un sous-espace affine de dimension $p - r$ de K^p, de direction $\mathcal{S}_0$.

Il est bon de retenir un cas particulier de ce théorème :

Théorème 66. Un système linéaire *homogène* de n équations à n inconnues possède une solution *non triviale* si, et seulement si, sa matrice $A \in M_n(K)$ n'est pas inversible, *i.e.* si, et seulement si, $\det(A) = 0$.

Démonstration. Soit r le rang du système, *i.e.* le rang de A. D'après le théorème précédent, $\mathcal{S}_0$ est un sous-espace vectoriel de dimension $n - r$ de K^n. Dire que le système possède une solution non triviale signifie que l'espace $\mathcal{S}_0$ n'est pas trivial, *i.e.* que $\dim(\mathcal{S}_0) > 0$, soit $r < n$. Cela revient à dire que A n'est pas inversible (*cf.* le corollaire 25 dans le module II.4), ou encore que $\det(A) = 0$ (en vertu du théorème 45 de la page 361). ■

Exercice 14.

Dans le plan $\mathbb{R}^2$, montrer que trois points $M_i := (x_i, y_i)$ $(i = 1, 2, 3)$ sont alignés si, et seulement si, le déterminant suivant est nul :

$$\Delta := \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}.$$

Solution. Voici l'idée : on se demande s'il existe une droite (affine) D, d'équation $ax + by + c = 0$, où $(a, b, c) \in \mathbb{R}^3$ et $(a, b) \neq 0$, passant par les trois points donnés,

ce qui revient à dire que (a, b, c) est solution du système homogène :

$$\begin{cases} ax_1 + by_1 + c = 0 \\ ax_2 + by_2 + c = 0 \\ ax_3 + by_3 + c = 0\,. \end{cases}$$

D'après le théorème précédent, ce système possède une solution $(a, b, c) \neq (0, 0, 0)$ si, et seulement si, son déterminant Δ est nul. Pour conclure, il reste à observer que, pour une telle solution, nécessairement $(a, b) \neq (0, 0)$.

Revenons à un système (S) quelconque. Le fait que l'ensemble $\mathcal{S}$ puisse être non vide ou vide appelle une définition :

Définition 17. Le système linéaire (53) est dit *compatible* s'il possède au moins une solution (*i.e.* si $\mathcal{S}$ n'est pas vide) et *incompatible* sinon (*i.e.* si $\mathcal{S}$ est vide).

Il est évidemment important de savoir si un système linéaire donné est compatible ou non. Voici une première réponse à cette question.

Proposition 67. Les conditions suivantes sont équivalentes :

1. Le système (S) est compatible.
2. Le second membre B est combinaison linéaire des colonnes de A.
3. Notons $[A; B] \in M_{n,p+1}(K)$ la matrice $[C_1; \ldots; C_p; B]$ obtenue en rajoutant B aux colonnes de A. Alors $[A; B]$ est de rang r.

Démonstration. Le système (S) se traduit par l'égalité (56), il est compatible si, et seulement si, $B \in \mathrm{Im}(f)$. Cette image est engendrée par les colonnes de A, images par f des vecteurs de la base canonique de K^p. Ainsi $\mathrm{Im}\, f = \mathrm{Vect}(C_1, \ldots, C_p)$, d'où l'équivalence des conditions 1 et 2.
La matrice $[A; B]$ ayant pour colonnes $C_1, \ldots, C_p, B$, son rang est la dimension de $\mathrm{Vect}(C_1, \ldots, C_p, B) = \mathrm{Im}\, f + \mathrm{Vect}(B) \subset K^n$. Comme $\mathrm{Im}\, f$ est de dimension r, le rang de $[A; B]$ vaut r si, et seulement si, $B \in \mathrm{Im}\, f$, d'où l'équivalence des conditions 1 et 3. ∎

Appliquons la propriété 3 de la page 375 à l'application linéaire $f : K^p \to K^n$ associée à la matrice A. L'équation $f(X) = B$ a une solution unique, ceci *quel que soit* $B \in K^n$, si, et seulement si, f est bijective, *i.e.* si $p = n$ et si A est *inversible*. Les systèmes de ce type méritent une attention particulière.

Définition 18. Le système linéaire (S) (ou sa forme matricielle (55)) est appelé *système de Cramer* si $p = n$ et si la matrice A du système est inversible.

Le fait que le système (S) soit de Cramer ne dépend pas de B, il ne dépend que de A. Le résultat essentiel concernant les systèmes de Cramer est le suivant (rappelons que $C_1, \ldots, C_n$ sont les colonnes de A) :

Théorème 68 (Formules de Cramer). Supposons que le système (55) soit de Cramer. Il possède alors une solution $X \in K^n$ unique, à savoir $X := A^{-1}B$. Si l'on écrit $X := {}^t(x_1, \ldots, x_n)$, les x_j sont donnés par les *formules de Cramer* :

$$x_j = \frac{\det\big([C_1; \ldots; C_{j-1}; B; C_{j+1}; \ldots; C_n]\big)}{\det(A)} \qquad (j = 1, \ldots, n)\,. \tag{57}$$

Démonstration. Le numérateur du second membre est le déterminant de la matrice obtenue en remplaçant la $j^{\text{ème}}$ colonne de A par B. Soit $X \in K^n = M_{n,1}(K)$.

Puisque A est inversible, l'égalité $AX = B$ équivaut à $A^{-1}(AX) = A^{-1}B$, soit $X = A^{-1}B$. D'où la première assertion.

Posons $X := A^{-1}B = {}^t(x_1, \ldots, x_n)$. D'après le théorème 44 de la page 361, si $\widetilde{A}$ désigne la comatrice de A, on a $(\det A)X = \big({}^t\widetilde{A}\big)B$. D'où, pour tout $j \in [\![1, n]\!]$:

$$(\det A)\, x_j = \sum_{i=1}^{n} \big({}^t\widetilde{A}\big)_{j,i} b_i = \sum_{i=1}^{n} \widetilde{A}_{i,j} b_i = \sum_{i=1}^{n} (-1)^{i+j} D_{i,j}\, b_i\,.$$

Si $i \in [\![1, n]\!]$, $D_{i,j}$ est le déterminant de la matrice déduite de A par suppression de la $j^{\text{ème}}$ colonne et de la $i^{\text{ème}}$ ligne, c'est aussi le déterminant de la matrice déduite de $[C_1; \ldots; C_{j-1}; B; C_{j+1}; \ldots]$ par suppression de la $j^{\text{ème}}$ colonne et de la $i^{\text{ème}}$ ligne. En développant le déterminant de la matrice $[C_1; \ldots; C_{j-1}; B; C_{j+1}; \ldots; C_n]$ par rapport à sa $j^{\text{ème}}$ colonne (théorème 43 de la page 358), on obtient l'égalité

$$(\det A)\, x_j = \det\big([C_1; \ldots; C_{j-1}; B; C_{j+1}; \ldots; C_n]\big).$$

■

Exercice 15.

Le système ci-dessous, où $a, b, c, d, u, v \in K$ sont donnés, est-il de Cramer ?

$$\begin{cases} ax + by = u \\ cx + dy = v\,. \end{cases}$$

Solution. C'est un système de Cramer si, et seulement si, $ad - bc \neq 0$.

Exercice 16.

Résoudre le système $\begin{cases} x - \ \ y + \ \ z = 1 \\ 2x + 4y + 9z = 1 \\ 3x - 4y + 2z = 1\,. \end{cases}$

Solution. La matrice de ce système est : $A = \begin{pmatrix} 1 & -1 & 1 \\ 2 & 4 & 9 \\ 3 & -4 & 2 \end{pmatrix}$.

Dans l'exercice 9 de la page 362, nous avions montré que $\det(A) = 1$.

Le système est donc de Cramer. Son unique solution (x, y, z) peut être calculée par les formules de Cramer (le lecteur vérifiera les calculs de déterminants) :

$$x = \begin{vmatrix} 1 & -1 & 1 \\ 1 & 4 & 9 \\ 1 & -4 & 2 \end{vmatrix} = 29,\ y = \begin{vmatrix} 1 & 1 & 1 \\ 2 & 1 & 9 \\ 3 & 1 & 2 \end{vmatrix} = 15,\ z = \begin{vmatrix} 1 & -1 & 1 \\ 2 & 4 & 1 \\ 3 & -4 & 1 \end{vmatrix} = -13\,.$$

Dans l'exercice précité, nous avions aussi calculé A^{-1}. Vérifions :

$$A^{-1}B = \begin{pmatrix} 44 & -2 & -13 \\ 23 & -1 & -7 \\ -20 & 1 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 29 \\ 15 \\ -13 \end{pmatrix}.$$

Remarques.

1. L'exercice ci-dessus met en lumière le fait suivant. Soit $A \in GL_n(K)$. Si l'on connaît A^{-1}, chaque système de Cramer $AX = B$ se résout aussitôt : sa seule solution est $X := A^{-1}B$. En fait, notons $(E_1, \ldots, E_n)$ la base canonique de $K^n = M_{n,1}(K)$. Pour tout $j \in [\![1, n]\!]$, le système $AX = E_j$ a une solution unique, notons-la C'_j. Alors :
$$A[C'_1; \ldots ; C'_n] = [AC'_1; \ldots ; AC'_n] = [E_1; \ldots ; E_n] = I_n .$$
Il en résulte que $[C'_1; \ldots ; C'_n] = A^{-1}$, *i.e.* les C'_j sont les colonnes de A^{-1}. Ainsi le calcul de A^{-1} est *équivalent* à la résolution des n systèmes de Cramer $AX = E_j$, $j = 1, \ldots, n$.
2. Les formules de Cramer ont surtout un intérêt théorique. Ainsi, soit à calculer A^{-1}, où $A \in GL_n(K)$ et n est grand, par exemple $n = 100$. Il est beaucoup plus rapide d'utiliser la méthode par « opérations élémentaires » exposée dans le module II.4 que de résoudre n systèmes de Cramer par les formules de Cramer !

Revenons au cas d'un système linéaire (S) quelconque, de rang r. Le cas $r = 0$ (c'est-à-dire $A = 0$) est trivial : ou bien $B \neq 0$, auquel cas le système est incompatible, ou bien $B = 0$, auquel cas $\mathcal{S} = K^p$. Supposons désormais $r \geqslant 1$. Nous allons d'abord préciser la proposition 67 de la page 379.

D'après le théorème 47 de la page 362 (ou son corollaire), il existe une matrice carrée d'ordre r extraite de A et inversible, notons-la $A_{I,J}$, où $I \subset [\![1, n]\!]$ et $J \subset [\![1, p]\!]$ sont des parties à r éléments *fixées*. Soit $i \in [\![1, n]\!]$. La $i^{ème}$ équation (53) est dite *équation principale* si $i \in I$ et non principale si $i \notin I$. De même, soit $j \in [\![1, p]\!]$. L'inconnue x_j est dite *inconnue principale* si $j \in J$ et non principale si $j \notin J$. Quitte à renuméroter les n équations (53), ce qui revient à permuter les lignes de A (et les composantes de B), nous pouvons supposer que $I := [\![1, r]\!]$. De même, quitte à renuméroter les inconnues, ce qui revient à permuter les colonnes de A, nous pouvons supposer que $J := [\![1, r]\!]$. Sous ces hypothèses, on a le résultat suivant :

Théorème 69. Supposons que les r premières équations (53) soient principales et que les inconnues $x_1, \ldots, x_r$ soient principales. Alors le système (S) est compatible si, et seulement si, les $n - r$ déterminants ci-dessous sont nuls :

$$\begin{vmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,r} & b_1 \\ a_{2,1} & a_{2,2} & \ldots & a_{2,r} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{r,1} & a_{r,2} & \ldots & a_{r,r} & b_r \\ a_{i,1} & a_{i,2} & \ldots & a_{i,r} & b_i \end{vmatrix} \qquad (i = r+1, \ldots, n).$$

Ces déterminants sont appelés *déterminants caractéristiques* du système.

Démonstration. D'après la proposition 67 de la page 379, le système est compatible si, et seulement si, la matrice $M' := [A; B] \in M_{n,p+1}(K)$ est de rang r. Appliquons le théorème 47 de la page 362 à la matrice M'. La matrice M, formée des r premières lignes et colonnes de M', c'est-à-dire de A, est inversible. Ainsi M' est de rang r si, et seulement si, le déterminant de toute matrice bordante de M (comme matrice extraite de M') est nul. Ces déterminants sont de l'un des types ci-dessous, où $i \in [\![r+1, n]\!]$ et $j \in [\![r+1, p]\!]$:

$$\begin{vmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,r} & a_{1,j} \\ a_{2,1} & a_{2,2} & \dots & a_{2,r} & a_{2,j} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{r,1} & a_{r,2} & \dots & a_{r,r} & a_{r,j} \\ a_{i,1} & a_{i,2} & \dots & a_{i,r} & a_{i,j} \end{vmatrix} \quad \text{ou} \quad \begin{vmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,r} & b_1 \\ a_{2,1} & a_{2,2} & \dots & a_{2,r} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{r,1} & a_{r,2} & \dots & a_{r,r} & b_r \\ a_{i,1} & a_{i,2} & \dots & a_{i,r} & b_i \end{vmatrix}.$$

Puisque A est de rang r, le théorème 47 de la page 362 montre que les déterminants du premier type sont nuls. Les déterminants du deuxième type sont les déterminants caractéristiques, et le théorème en résulte. ■

Lorsqu'on sait qu'un système est compatible, le théorème ci-dessous fournit une méthode pour le résoudre à l'aide de certains systèmes de Cramer. Pour formuler ce théorème il est commode de supposer, comme pour le théorème précédent, que les équations principales (resp. inconnues principales) choisies sont les r premières. Autrement dit la matrice M formée des r premières lignes et colonnes de A est inversible.

Théorème 70 (Rouché-Fontené). Avec ces hypothèses, supposons de plus le système (S) compatible. On obtient toutes les solutions $(x_1, \dots, x_p)$ de ce système ainsi :

1. On donne des valeurs arbitraires aux inconnues non principales $x_{r+1}, \dots, x_p$ (si $r < p$).
2. Ces valeurs étant choisies, on considère le système (C) suivant en les inconnues $x_1, \dots, x_r$:

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,r}x_r = b_1 - (a_{1,r+1}x_{r+1} + \dots + a_{1,p}x_p) \\ a_{2,1}x_1 + a_{2,2}x_2 + \dots + a_{2,r}x_r = b_2 - (a_{2,r+1}x_{r+1} + \dots + a_{2,p}x_p) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{r,1}x_1 + a_{r,2}x_2 + \dots + a_{r,r}x_r = b_r - (a_{r,r+1}x_{r+1} + \dots + a_{r,p}x_p)\,. \end{cases}$$

   C'est un système de Cramer, dont la résolution détermine les valeurs des inconnues principales $x_1, \dots, x_r$.

Démonstration. La matrice $M' := [A; B]$ est de rang r (proposition 67 de la page 379), et ses r premières lignes sont linéairement indépendantes (car les lignes de M le sont).

Soit $i \in [\![r+1, n]\!]$. La $i^{ème}$ ligne de M' est combinaison linéaire des r premières lignes de M'. Il existe donc $\beta_1, \dots, \beta_r \in K$ (dépendant de i) tels que :

$$(a_{i,1}, \dots, a_{i,p}, b_i) = \sum_{k=1}^{r} \beta_k (a_{k,1}, \dots, a_{k,p}, b_k).$$

On en déduit, pour tous $x_1, \ldots, x_p \in K$, l'égalité suivante (*cf.* les coefficients de $x_1, \ldots, x_p$) :

$$(a_{i,1}x_1 + \cdots + a_{i,p}x_p) - b_i \;=\; \sum_{k=1}^{r} \beta_k \big[(a_{k,1}x_1 + \cdots + a_{k,p}x_p) - b_k\big]\,.$$

Ainsi, parmi les égalités (53), la $i^{ème}$ est une conséquence des r premières. Il en résulte que le système (S) est équivalent au système formé par les r premières équations (53), *i.e.* ces deux systèmes ont même ensemble de solutions $\mathcal{S}$. Ainsi, un p-uplet $(x_1, \ldots, x_p) \in K^p$ appartient à $\mathcal{S}$ si, et seulement si, $(x_1, \ldots, x_r)$ est solution du système (C) associé à $(x_{r+1}, \ldots, x_p)$. Ce système (C) est de Cramer : sa matrice M est inversible. Ce qui précède montre que l'application $(x_1, \ldots, x_p) \mapsto (x_{r+1}, \ldots, x_p)$ de $\mathcal{S}$ dans K^{p-r} est bijective : le seul antécédent de $(x_{r+1}, \ldots, x_p)$ par cette application est $(x_1, \ldots, x_r, x_{r+1}, \ldots, x_p)$, où $(x_1, \ldots, x_r)$ est l'unique solution du système (C) associé à $(x_{r+1}, \ldots, x_p)$. Le théorème en résulte. ■

Exercice type corrigé

II.7.0 Soient E un espace vectoriel de dimension finie n et $V_1, \ldots, V_p$ des sous-espaces vectoriels de E $(p \geqslant 1)$. Démontrer l'inégalité :

$$\dim(V_1) + \dim(V_2) + \cdots + \dim(V_p) \leqslant (p-1)n + \dim(V_1 \cap V_2 \cap \cdots \cap V_p)\,,$$

par récurrence sur p. En déduire que, si chaque V_i est un hyperplan de E et si $p < n$, l'intersection des V_i n'est pas triviale.
Retrouver ce résultat en interprétant chaque V_i comme noyau d'une forme linéaire.
Solution. L'inégalité $(*)$ à démontrer est évidente (c'est une égalité) si $p = 1$. Supposons $p \geqslant 2$, le résultat étant vrai à l'ordre $p-1$. Soient donc $V_1, \ldots, V_p$ des sous-espaces vectoriels de E. Soit W l'intersection de $V_1, \ldots, V_{p-1}$. L'hypothèse de récurrence donne l'inégalité :

$$\sum_{i=1}^{p-1} \dim(V_i) \;\leqslant (p-2)n + \dim(W).$$

Par ailleurs, le théorème 16 de la page 339 donne, puisque $\dim(V_p + W) \leqslant n$:

$$\dim(V_p) + \dim(W) \;=\; \dim(V_p + W) + \dim(V_p \cap W) \;\leqslant\; n + \dim(V_p \cap W)\,.$$

Additionnons membre à membre les deux inégalités obtenues. Il vient :

$$\sum_{i=1}^{p} \dim(V_i) + \dim(W) \;\leqslant\; (p-1)n + \dim(W) + \dim(V_p \cap W)\,.$$

En simplifiant par $\dim W$, on obtient l'inégalité $(*)$ au rang p, puisque $V_p \cap W$ est l'intersection de $V_1, \ldots, V_p$.
Désormais $p < n$, et chaque V_i est un hyperplan. Notons V l'intersection de $V_1, \ldots, V_p$. L'inégalité $(*)$ s'écrit : $p(n-1) \leqslant (p-1)n + \dim(V)$, c'est-à-dire $\dim(V) \geqslant p(n-1) - (p-1)n = n - p$. Mais $n - p > 0$, et donc $\dim(V) > 0$, *i.e.* $V \neq \{0\}$.

Voici une autre solution. Pour tout i, il existe une forme linéaire non nulle φ_i sur E telle que $\mathrm{Ker}(\varphi_i) = V_i$. Considérons alors l'application $f : x \longmapsto \big(\varphi_1(x), \ldots, \varphi_p(x)\big)$ de E dans K^p. Elle est linéaire (car chaque φ_i l'est), et son noyau est formé des $x \in E$ tels que $\varphi_i(x) = 0$ pour tout i, *i.e.* $\mathrm{Ker}(f) = V$. Comme $\dim(E) = n > p = \dim(K^p)$, f n'est pas injective. Autrement dit $\mathrm{Ker}(f) = V$ n'est pas réduit à $\{0\}$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

II.7.1 Déduire directement la proposition 13 de la page 338 du théorème 12 de la page 337.

II.7.2 Soient E, E', F, F' des espaces vectoriels, et $u : E' \to E$, $f : E \to F$, $v : F \to F'$ trois applications linéaires. On suppose que u est surjective et que v est injective. Montrer que $v \circ f \circ u$ est de rang fini si, et seulement si, f l'est, auquel cas $v \circ f \circ u$ et f ont même rang.

II.7.3 $*$ Soient E un espace vectoriel de dimension finie n et $f \in \mathcal{L}(E)$ tel que $f^2 = 0$. Démontrer l'inégalité $\mathrm{rang}(f) \leqslant n/2$.

II.7.4 $*$ Soient E, F, G trois espaces vectoriels et $f : E \to F$, $g : F \to G$ deux applications linéaires.

1. On suppose que f et g sont de rang fini. Montrer que $g \circ f : E \to G$ est de rang fini, et prouver l'inégalité $\mathrm{rang}(g \circ f) \leqslant \min(\mathrm{rang}\, f, \mathrm{rang}\, g)$.
2. On suppose F de dimension finie n. Montrer que $g \circ f$ est de rang fini, et prouver l'inégalité $\mathrm{rang}(g \circ f) \geqslant \mathrm{rang}\, f + \mathrm{rang}\, g - n$. On pourra étudier la restriction de g à $\mathrm{Im}(f)$. Lien avec l'exercice précédent ?

II.7.5 $*$ Soient E un espace vectoriel de dimension finie et $f \in \mathcal{L}(E)$. Pour tout $k \in \mathbb{N}$, posons $V_k := \mathrm{Ker}(f^k)$ et $d_k := \dim(V_k)$.

1. Montrer que la suite $(V_k)_{k \geqslant 0}$ est croissante (pour l'inclusion), en déduire que la suite $(d_k)_{k \geqslant 0}$ est stationnaire.
2. Montrer que la suite $(d_{k+1} - d_k)_{k \geqslant 0}$ est décroissante. Si $k \in \mathbb{N}$, on étudiera la restriction de f à un supplémentaire de V_{k+1} dans V_{k+2}.

II.7.6 Soient E, F deux espaces vectoriels et $f \in \mathcal{L}(E, F)$. Montrer que f est de rang 1 si, et seulement s'il existe un vecteur non nul $b \in F$ et une forme linéaire non nulle α sur E tels que $f(x) = \alpha(x) \cdot b$ pour tout $x \in E$. Décrire alors tous les couples (α, b) répondant à la question.

II.7.7 Soient E un espace vectoriel de dimension finie n et $f \in \mathcal{L}(E)$. On suppose que f n'est pas une homothétie. Montrer qu'il existe une base $(e_1, \ldots, e_n)$ de E telle que $f(e_1) = e_2$ (d'où en particulier $n \geqslant 2$).

II.7.8 Soient E un espace vectoriel de dimension finie et $u, v \in \mathcal{L}(E)$. On suppose que $u \circ v = 0$ et $u + v \in \mathcal{GL}(E)$. Montrer que $E = \mathrm{Ker}(u) \oplus \mathrm{Im}(u)$. Inversement, si $u \in \mathcal{L}(E)$ et si $E = \mathrm{Ker}(u) \oplus \mathrm{Im}(u)$, construire $v \in \mathcal{L}(E)$ de sorte que $u \circ v = 0$ et $u + v \in \mathcal{GL}(E)$.

II.7.9 Soit E un $\mathbb{C}$-espace vectoriel. On peut le considérer comme $\mathbb{R}$-espace vectoriel, par restriction de la loi externe de $\mathbb{C} \times E$ à $\mathbb{R} \times E$. Montrer que E est de dimension finie sur $\mathbb{C}$ si, et seulement s'il est de dimension finie sur $\mathbb{R}$, auquel cas la dimension de E sur $\mathbb{R}$ est le double de la dimension de E sur $\mathbb{C}$.

II.7.10 Cet exercice généralise le précédent. Soient L un corps et K un sous-corps de L. Étant donné un L-espace vectoriel E, la restriction de la loi externe $L \times E \to E$ à $K \times E$ fait de E un K-espace vectoriel. Ainsi, L est un K-espace vectoriel pour la loi externe $K \times L \to L$ définie par $(\lambda, a) \mapsto \lambda \cdot a = \lambda a$ (λa est le produit de λ et a dans L). On suppose désormais que L est de dimension finie p comme K-espace vectoriel.
Soit maintenant E un L-espace vectoriel de dimension finie n. Montrer que E est un K-espace vectoriel de dimension np.
Étant données une base $(a_1, \ldots, a_p)$ de L sur K et une base $(e_1, \ldots, e_n)$ de E sur L, on montrera que la famille $(a_i \cdot e_j)_{(i,j) \in [\![1,p]\!] \times [\![1,n]\!]}$ est une base de E sur K. Que dire si l'on suppose seulement que E est de dimension finie sur K ?

II.7.11 Soit K un corps (commutatif) ayant un nombre fini q d'éléments. Par la proposition 54 de la page 152, la caractéristique de K est un nombre premier p, et d'après l'exercice II.2.39 de la page 159, K peut être considéré comme espace vectoriel sur $\mathbb{Z}/p\mathbb{Z}$. En déduire que q est une puissance de p (ainsi, par exemple, il n'existe pas de corps commutatif à 6 éléments).

II.7.12 $*$ Considérons un entier $n \geqslant 1$. Notons R_n l'ensemble des matrices $A \in M_n(K)$ ayant la propriété suivante : il existe un scalaire $\ell(A)$ (dépendant de A) tel que la somme des termes de chaque ligne (resp. colonne) de A soit égale à $\ell(A)$.

1. Montrer que R_n est une sous-algèbre de $M_n(K)$, *i.e.* est à la fois un sous-K-espace vectoriel et un sous-anneau de $M_n(K)$.
2. Montrer que $\ell : R_n \to K$ est un morphisme d'algèbres, *i.e.* est à la fois un morphisme d'anneaux et une application K-linéaire.
3. Montrer que $\dim(R_n) = n^2 - 2n + 2$ $\big($établir un isomorphisme entre R_n et K^{n^2-2n+2}, en s'inspirant au besoin les cas $n = 2$ et $n = 3\big)$.

Exercices sur la section 2

II.7.13 Dans $\mathbb{R}^3$, soient P le plan d'équation $x + y + z = 0$ et D la droite dirigée par $(1, 1, 1)$. Montrer que $\mathbb{R}^3 = P \oplus D$. On désigne par $s \in \mathcal{L}(\mathbb{R}^3)$ la symétrie par rapport à P parallèlement à D. Déterminer la matrice de s dans la base canonique de $\mathbb{R}^3$.

II.7.14 Soient E un espace vectoriel de dimension 3 et $f \in \mathcal{L}(E)$, $f \neq 0$. Montrer que $f^2 = 0$ si, et seulement s'il existe une base $\mathcal{B}$ de E dans laquelle f ait pour matrice $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

II.7.15 Soit $n \in \mathbb{N}$. Étant donné un élément a de K fixé, déterminer la matrice A de l'endomorphisme $P \mapsto P(X + a)$ de $K_n[X]$ dans la base $(1, X, \ldots, X^n)$. Montrer que A est inversible et calculer son inverse.

II.7.16 Soient n un entier strictement positif et $P := X^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0 \in K[X]$ un polynôme unitaire de degré n. Notons $f : K_{n-1}[X] \to K_{n-1}[X]$ l'application associant à tout polynôme $Q \in K_{n-1}[X]$ le reste dans la division euclidienne de XQ par P. Montrer que f est un endomorphisme de $K_{n-1}[X]$. Expliciter la matrice de f dans la base $(1, X, \ldots, X^{n-1})$ de $K_{n-1}[X]$.

II.7.17 Considérons un entier $n \geqslant 2$. Soit $\mathcal{B} := (e_1, \ldots, e_n)$ la base canonique de K^n. Pour tout $k \in [\![1, n]\!]$, posons $V_k := \mathrm{Vect}(e_1, \ldots, e_k)$. Rappelons (*cf.* le module II.4, section 1.2) que $T_n(K)$ est l'ensemble des matrices triangulaires supérieures de $M_n(K)$ et que $T_n^*(K)$ est formé des matrices de $T_n(K)$ dont les coefficients diagonaux sont nuls.

1. Soient $A \in M_n(K)$ et $f \in \mathcal{L}(K^n)$ tels que $\mathrm{Mat}_{\mathcal{B}}(f) = A$. Montrer que $A \in T_n(K)$ si, et seulement si, $f(V_k) \subset V_k$ pour $k = 1, \ldots, n$.
2. En déduire, sans calculs matriciels, que $T_n(K)$ est un sous-anneau et un sous-espace vectoriel de $M_n(K)$, puis que $T_n(K) \cap GL_n(K)$ est un sous-groupe multiplicatif de $GL_n(K)$.
3. Soient A et f comme en 1. Montrer que $A \in T_n^*(K)$ si, et seulement si, $f(V_k) \subset V_{k-1}$ pour $k = 2, \ldots, n$ et $f(V_1) = (0)$. En déduire que, si $A_1, \ldots, A_n \in T_n^*(K)$, le produit $A_1 A_2 \cdots A_n$ est nul.

II.7.18 $*$ Soient $m, n, p \in \mathbb{N}^*$ et $A \in M_{n,p}(K)$ une matrice fixée. Calculer, en fonction du rang de A, la dimension du sous-espace vectoriel de $M_{m,n}(K)$ formé des matrices M telles que $MA = 0$ (donner deux solutions). Même question si M est fixée et A varie.

Exercices sur la section 3

II.7.19 Calculer le déterminant $\begin{vmatrix} 1 & 3 & 5 \\ -2 & 4 & 1 \\ 3 & -2 & 7 \end{vmatrix}$ par la formule (23) de la page 352, puis par développement par rapport à une ligne ou colonne, enfin à l'aide d'opérations élémentaires.

II.7.20 Prouver l'identité suivante :

$$\begin{vmatrix} x & a & b & c \\ a & x & c & b \\ b & c & x & a \\ c & b & a & x \end{vmatrix} = (x+a+b+c)(x+a-b-c)(x-a+b-c)(x-a-b+c)\,.$$

II.7.21 Montrer que $\begin{vmatrix} 3 & 5 & 7 \\ 4 & -1 & 2 \\ -2 & 2 & 5 \end{vmatrix}$ est un entier se terminant par un 5, sans calculer ce déterminant (*cf.* la remarque de la page 353).

II.7.22 Dans le plan, soient $a, b, c \in]0, \pi[$ les angles d'un triangle. Prouver l'égalité suivante :

$$\begin{vmatrix} 1 & \cos a & \cos b \\ \cos a & 1 & \cos c \\ \cos b & \cos c & 1 \end{vmatrix} = 4 \cos a \, \cos b \, \cos c\,.$$

II.7.23 $*$ Soit $A \in M_n(K)$, $n \geqslant 1$. Pour toute partie non vide I de $[\![1, n]\!]$, de cardinal k, notons D_I le déterminant de la matrice $A_{I,I} \in M_k(K)$ extraite de A (formée des lignes ou colonnes dont les indices appartiennent à I). En utilisant la formule (28) du module II.2, montrer que, pour tout $k \in [\![1, n]\!]$, le coefficient de X^{n-k} dans χ_A, polynôme caractéristique de A, est la somme :

$$(-1)^k \sum_I D_I\,,$$

I décrivant l'ensemble des parties à k éléments de $[\![1, n]\!]$. En déduire que, si $K := \mathbb{R}$ et si tous les D_I sont positifs, A n'a pas de valeur propre réelle strictement négative.

II.7.24 Soient $A \in M_n(K)$, $n \geqslant 2$, et $\widetilde{A}$ la comatrice de A.
Calculer le rang et le déterminant de $\widetilde{A}$ en fonction de $\operatorname{rang}(A)$ et $\det(A)$.

II.7.25 Soit E un espace vectoriel réel de dimension finie $n \geqslant 1$.

1. Soit $u \in \mathcal{GL}(E)$. Montrer que $\det(u) > 0$ si, et seulement si, pour toute base $\mathcal{B} := (e_1, \ldots, e_n)$ de E, la base $\big(u(e_1), \ldots, u(e_n)\big)$ est de même sens que $\mathcal{B}$ (voir l'application de la page 364). Si c'est le cas, on dit que u *conserve l'orientation*, sinon on dit que u *change l'orientation*.
2. Une homothétie de rapport non nul conserve-t-elle l'orientation ?
On séparera les cas selon la parité de n.
3. Soient V, W deux sous-espaces vectoriels supplémentaires de E et s la symétrie par rapport à V parallèlement à W. À quelle condition s conserve-t-elle l'orientation ?

II.7.26 Cet exercice utilise les dérivées de fonctions vectorielles (*cf.* le module IV.6).
Soient I un intervalle ouvert de $\mathbb{R}$, n un entier strictement positif. Pour tout $(i,j) \in [\![1,n]\!]^2$, soit $a_{i,j} : I \to \mathbb{R}$ une fonction dérivable. Pour tout $t \in I$, notons $A(t)$ la matrice de coefficient général $a_{i,j}(t)$, $C_1(t), \ldots, C_n(t)$ les colonnes de $A(t)$, et $D(t) := \det\big(C_1(t), C_2(t), \ldots, C_n(t)\big)$ le déterminant de $A(t)$, *i.e.* le déterminant de $\big(C_1(t), \ldots, C_n(t)\big)$ dans la base canonique de $\mathbb{R}^n$.
Montrer que la fonction $t \mapsto D(t)$ est dérivable sur I, sa dérivée étant donnée par la formule suivante :

$$\begin{aligned} D'(t) = {} & \det\big(C_1'(t), C_2(t), \ldots, C_n(t)\big) + \det\big(C_1(t), C_2'(t), \ldots, C_n(t)\big) \\ & + \cdots + \det\big(C_1(t), \ldots, C_{n-1}'(t), C_n(t)\big) + \det\big(C_1(t), C_2(t), \ldots, C_n'(t)\big) . \end{aligned}$$

II.7.27 Soit $\mathcal{H}$ l'ensemble des matrices de la forme $\begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix}$, où (a,b) décrit $\mathbb{C}^2$.

1. Montrer que $\mathcal{H}$ est un sous-anneau de $M_2(\mathbb{C})$. Montrer que $\mathcal{H}$ est un sous-$\mathbb{R}$-espace vectoriel de $M_2(\mathbb{C})$, et expliciter une base de $\mathcal{H}$ sur $\mathbb{R}$. Vérifier que $\mathcal{H}$ n'est pas un sous-$\mathbb{C}$-espace vectoriel de $M_2(\mathbb{C})$.
2. Montrer que l'anneau $\mathcal{H}$ est un corps. Donner deux solutions, l'une utilisant l'exercice 4 de la page 341, l'autre étant plus explicite.
3. Comparer $\mathcal{H}$ avec le corps des quaternions $\mathbf{H}$, (*cf.* l'exercice II.5.4). On pourra considérer les matrices suivantes :

$$\mathbf{i} := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} , \ \mathbf{j} := \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} , \ \mathbf{k} := \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} .$$

II.7.28 $*$ Soient $I = \left]0, +\infty\right[$ et n un entier strictement positif.

1. Soient $t_1, \ldots, t_n \in \mathbb{R}^*$ et $b_1, \ldots, b_n \in \mathbb{R}$ tels que $b_1 > b_2 > \cdots > b_n$. Pour tout $x \in I$, posons :
$$f(x) \ := \ t_1 x^{b_1} + t_2 x^{b_2} + \cdots + t_n x^{b_n} .$$
Montrer que l'équation $f(x) = 0$ a au plus $n-1$ solutions dans I (raisonner par récurrence sur n et penser au théorème de Rolle).

2. Soient $x_1, \ldots, x_n, a_1, \ldots, a_n$ des réels tels que :
$$a_1 < a_2 < \cdots < a_n \quad \text{et} \quad 0 < x_1 < x_2 < \cdots < x_n.$$
Soit D le déterminant de la matrice $A \in M_n(\mathbb{R})$ de coefficients $x_i^{a_j}$, $(i,j) \in [\![1,n]\!]^2$. On veut prouver l'inégalité $D > 0$. Si $x \in I$, soit $g(x)$ le déterminant obtenu en remplaçant x_n par x dans D. Montrer que l'équation $g(x) = 0$ a exactement $n-1$ solutions dans I. Conclure en raisonnant par récurrence sur n (examiner le comportement de $g(x)$ lorsque x tend vers $+\infty$).

II.7.29 Soit E l'espace vectoriel des suites complexes.

1. Pour tout $z \in \mathbb{C}$, notons $u(z)$ la suite géométrique $(z^n)_{n \geqslant 0}$.

   Montrer que la famille $\big(u(z)\big)_{z \in \mathbb{C}}$ de vecteurs de E est libre (on pourra penser aux déterminants de Vandermonde).

2. Soit t un entier strictement positif. Une suite $(u_n)_{n \geqslant 0}$ est dite t-périodique si $u_{n+t} = u_n$ pour tout $n \in \mathbb{N}$, et l'on note F_t l'ensemble des suites t-périodiques. Montrer que F_t est un sous-espace vectoriel de E, de dimension t ; on construira un isomorphisme de F_t sur $\mathbb{C}^t$.

3. Une suite est dite périodique si elle est t-périodique pour un certain entier $t \geqslant 1$, et l'on note F l'ensemble des suites périodiques. Montrer que F est un sous-espace vectoriel de E.

4. Pour tout $q \in \mathbb{Q}$, posons $v(q) := u\big(\exp(2\pi i q)\big)$.

   Montrer que la famille $\big(v(q)\big)_{q \in T}$ est une base de F, en posant $T := \mathbb{Q} \cap \big[0, 1\big[$.

II.7.30 Soit $J \in M_n(\mathbb{R})$ la matrice dont tous les coefficients valent 1 ($n \geqslant 2$).

1. Donner deux preuves de l'égalité $J^2 = nJ$. Quel est le rang de J ?

2. Montrer qu'il existe $P \in GL_n(\mathbb{R})$ telle que les $n-1$ premières colonnes de $P^{-1}JP$ soient nulles. En déduire le polynôme caractéristique de J.

II.7.31 Soit $A \in M_n(\mathbb{R})$ une matrice à coefficients positifs ou nuls. On suppose que la somme des coefficients de chaque ligne de A vaut 1.
Montrer que, si $\lambda \in \mathbb{C}$ est une valeur propre de A, on a $|\lambda| \leqslant 1$.

II.7.32 Soient E, F deux espaces vectoriels de dimension finie et $f \in \mathcal{L}(E,F)$, $g \in \mathcal{L}(F,E)$. Montrer qu'un scalaire *non nul* $\lambda \in K$ est valeur propre de $g \circ f$ si, et seulement s'il est valeur propre de $f \circ g$.

II.7.33 Soient E un espace vectoriel réel de dimension finie $n \geqslant 1$ et f un endomorphisme de E tel que $f^2 = -Id_E$.

1. Montrer que f n'a aucune valeur propre. En déduire, par l'absurde, que n est pair (penser au théorème des valeurs intermédiaires).

2. Si $a, b \in \mathbb{R}$ et $x \in E$, posons $(a+ib) \cdot x := ax + bf(x)$. Montrer que cette loi externe $\mathbb{C} \times E \to E$ fait de E un $\mathbb{C}$-espace vectoriel. Retrouver alors le fait que n soit pair, à l'aide de l'exercice II.7.9.

Exercices sur la section 4

II.7.34 Résoudre de plusieurs façons le système suivant sur $\mathbb{Q}$:

$$\begin{cases} 2x + 3y + 5z = 10 \\ 3x + 7y + 4z = 3 \\ x + 2y + 2z = 3\,. \end{cases}$$

II.7.35 Résoudre le système suivant, où $t \in K$ est un paramètre :

$$\begin{cases} x + ty + 2z = t \\ -2x + y + (t-2)z = 1 \\ tx + y + 2z = 2t - 1\,. \end{cases}$$

Discuter suivant les valeurs de t, en précisant les valeurs de t pour lesquelles le système est compatible, la dimension du sous-espace affine de K^3 formé des solutions du système, et en détaillant le cas où le système est de Cramer.

II.7.36 Soit $A \in M_n(K)$ une matrice de rang $n-1$ ($n \geqslant 2$). Montrer qu'il existe $k \in [\![1, n]\!]$ tel que, en posant $y_j := (-1)^{k+j} D_{k,j}$ (*cf.* la section 3.2 pour la notation $D_{i,j}$) pour $j = 1, \ldots, n$, l'espace des solutions du système $AX = 0$ soit la droite engendrée par le vecteur ${}^t(y_1, \ldots, y_n)$.

II.7.37 Déterminer les valeurs propres et vecteurs propres des matrices suivantes de $M_3(\mathbb{C})$:

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

II.7.38 Soit $n \geqslant 3$.
Déterminer les valeurs propres et vecteurs propres de la matrice suivante de $M_n(\mathbb{R})$:

$$\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \end{pmatrix}.$$

II.7.39 Montrer que, dans l'espace $\mathbb{R}^3$, quatre points $M_i := (x_i, y_i, z_i)$ ($i = 1, 2, 3, 4$) appartiennent à un même plan affine si, et seulement si :

$$\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x_4 & y_4 & z_4 & 1 \end{vmatrix} = 0\,.$$

II.7.40 Dans le plan $\mathbb{R}^2$, considérons trois droites D_i d'équations $a_i x + b_i y + c_i = 0$, où $(a_i, b_i) \neq (0, 0)$ ($i = 1, 2, 3$). Montrer que ces droites sont concourantes (*i.e.* passent par un même point) ou toutes parallèles si, et seulement si, le déterminant suivant est nul :

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}.$$

II.7.41 Dans le plan $\mathbb{R}^2$, soient $M_i := (x_i, y_i)$ $(i = 1, 2, 3)$ trois points non alignés . Montrer qu'il existe un unique cercle passant par ces trois points, et déterminer son centre $A := (a, b)$ et son rayon R (étudier le système ci-après en les inconnues a, b, c et appliquer l'exercice 14 de la page 378).

$$\begin{cases} x_1^2 + y_1^2 - 2ax_1 - 2by_1 + c \;=\; 0 \\ x_2^2 + y_2^2 - 2ax_2 - 2by_2 + c \;=\; 0 \\ x_3^2 + y_3^2 - 2ax_3 - 2by_3 + c \;=\; 0\,. \end{cases}$$

Application numérique : $M_1 := (-7, 5)$, $M_2 := (1, 1)$, $M_3 := (7, -3)$.

II.7.42 Soit le système linéaire (53) à coefficients dans K, K étant un sous-corps d'un corps L. Montrer que ce système possède une solution sur K si, et seulement s'il en possède une sur L.

II.7.43 $*$ Soient $A, B \in \mathbb{C}[X]$ deux polynômes non constants, de degrés respectifs n, p. Avec les notations de l'exercice II.6.4 de la page 326, montrer que A et B ont une racine commune si, et seulement si, le déterminant de la matrice $S(A, B)$ est nul. Montrer en général que le rang de la matrice $S(A, B)$ vaut $n + p - \deg(A \wedge B)$.

II.8

Initiation à l'algorithmique et au calcul formel

Si le cœur des mathématiques est la forme pure, axiomatique et déductive née en Grèce il y a 2 500 ans, la résolution *effective* de problèmes a toujours fait partie de la compétence professionnelle des mathématiciens. Nous en avons rencontré des exemples classiques au module II.1 (algorithme d'Euclide de recherche du pgcd de deux entiers naturels), au module II.4 (méthode du pivot de Gauß pour la résolution de systèmes d'équations linéaires), au module II.5 (résolution selon Cardan de l'équation du 3ème degré) et au module II.6 (schéma de Horner pour l'évaluation d'un polynôme, algorithme d'Euclide de recherche du pgcd de deux polynômes et de calcul des coefficients de Bézout).

L'avènement et l'expansion rapide de l'informatique ont accentué cet aspect des mathématiques et donné naissance à de nouveaux domaines des mathématiques appliquées. Nous ignorerons ici la programmation (qui ne relève pas d'un manuel de mathématiques) et l'analyse numérique (qui sera abordée en L2), pour nous concentrer sur le dernier-né de ces domaines, le *calcul formel*. Son nom anglo-saxon de *computer algebra* indique clairement qu'il se relie principalement à l'algèbre.

Notre but, dans ce module, est d'étudier quelques algorithmes fondamentaux en des termes qui fournissent une passerelle entre leur aspect mathématique (tel qu'il est apparu dans les exemples cités) et leur aspect informatique (tel qu'on l'expérimente lorsque l'on programme des ordinateurs). Il ne s'agit pas d'une introduction à l'informatique. Après avoir dégagé et motivé la plupart des notions essentielles sur des exemples (section 1), on formulera (section 2) des rudiments du vocabulaire de l'algorithmique (description et analyse des algorithmes) sans définitions formelles et sans énoncés généraux. Puis, à la section 3, on appliquera ces notions à quelques algorithmes élémentaires mais non triviaux.

1 Exemple introductif : l'addition en base b

Il est chaudement recommandé au lecteur d'exécuter sur des exemples tous les algorithmes présentés ici.

Soit b un entier naturel supérieur ou égal à 2. Comme on l'a vu au module II.1 (théorème 31 de la page 81), tout entier naturel $n \in \mathbb{N}$ admet une unique écriture en base b :

$$n = \sum_{i \geqslant 0} c_i b^i,$$

dans laquelle les c_i sont des chiffres : $c_i \in [\![0, b-1]\!]$, nuls à partir d'un certain rang (sans quoi la somme ci-dessus n'aurait pas de sens). Les exemples les plus importants de bases de numération sont dix (numération décimale, utilisée dans la vie quotidienne) et deux et seize (numérations binaire et hexadécimale, utilisées en informatique). Si $n \neq 0$, on peut préciser l'écriture :

$$n = c_0 + \cdots + c_k b^k \ , \ c_k \neq 0,$$

ce qui permet de *coder* l'entier n par la suite de ses chiffres écrite « poids forts à gauche » :

$$n = (c_k \ldots c_0)_b.$$

Exercice 1.

Préciser la *taille* de ce codage, autrement dit, le nombre $|n|_b$ de chiffres nécessaires à l'écriture de n.

Solution. Dans l'écriture précédente, on a l'encadrement :

$$0 + \cdots + 0\, b^{k-1} + b^k \leqslant n \leqslant (b-1) + \cdots + (b-1)b^k = b^{k+1} - 1,$$

autrement dit, $b^k \leqslant n < b^{k+1}$, d'où $k \leqslant \log_b n < k+1$. Le nombre de chiffres de n est donc $|n|_b = k+1 = 1 + \lfloor \log_b n \rfloor$. Lorsque $n \to +\infty$, son *ordre de grandeur* est $\Theta(\log n)$; cela signifie qu'il existe des constantes $a, b > 0$ telles que ce nombre de chiffres est compris entre $a \log n$ et $b \log n$ quel que soit n. Remarquons d'ailleurs que *cette estimation est indépendante du type de logarithme utilisé* (népérien, décimal ou de base 2).

1.1 L'algorithme d'addition

L'intérêt du codage en base b est qu'il permet d'effectuer les quatre opérations arithmétiques fondamentales, et bien d'autres encore, par des *successions automatiques d'opérations élémentaires*. De telles séquences sont appelées *algorithmes* depuis le moyen âge, époque où les mathématiciens Arabes ont dominé et revitalisé notre discipline : ce mot est en effet dérivé du nom du mathématicien Perse Abu Jaʼfar Mohammed ibn Mûsâ al-Khowârizmî, qui a vécu aux alentours de l'an 825. Nous allons détailler l'exemple apparemment innocent de l'addition des entiers naturels : nous verrons ainsi ce que sont les « opérations élémentaires » et sur quoi elles portent, puis en quoi consiste une « succession automatique » de telles opérations.

Les opérations élémentaires sont ici résumées par la table d'addition de l'école primaire ! Le principe est que la somme de deux nombres à un chiffre (*i.e.* compris entre 0 et $b-1$) est soit un nombre à un chiffre soit un nombre à deux chiffres, dont le chiffre de poids fort est 1 : si $c+d \leqslant b-1$, son écriture en base b est $(c+d)_b$; sinon, $b \leqslant c+d \leqslant 2b-2$ et son écriture en base b est $(1e)_b$, avec $e := c+d-b \in [\![0, b-1]\!]$. Un écolier apprend ces opérations

par cœur et en restitue le résultat en temps constant (et faible). Un ordinateur *tabule* ces résultats, c'est à dire qu'il les garde dans une zone mémoire ; ou bien il possède un circuit spécialisé dans les additions à un chiffre. Dans tous les cas, ces opérations élémentaires s'effectuent en temps constant (et faible). De plus, elles concernent des données et des résultats de taille fixe : l'écolier n'a pas besoin de papier et crayon et le programmeur n'a pas besoin de demander l'allocation d'une zone mémoire pour les effectuer.

Exercice 2.
Décrire de manière analogue la table de multiplication.
Solution. Puisque $0 \leqslant c, d \leqslant b-1$, on a $0 \leqslant cd \leqslant (b-1)^2 = b(b-2)+1$. Le produit cd est donc un nombre à un chiffre (si $cd \leqslant b-1$) ou bien (si $b \leqslant cd \leqslant b(b-2)+1$) à deux chiffres : $cd = (fe)_b$, avec de plus $f \leqslant b-2$.

Pour simplifier, nous considérerons des entiers naturels non nuls $n := (c_k \dots c_0)_b$ et $p := (d_\ell \dots d_0)_b$. Par convention, on note $c_i := 0$ lorsque $i > k$ et $d_i := 0$ lorsque $i > \ell$. Les deux nombres jouant un rôle symétrique, nous supposerons que $k \leqslant \ell$. La somme des entiers n et p est l'entier q dont l'écriture $q = (e_m \dots e_0)_b$ en base b est définie de la manière suivante. On pose $r_0 := 0$, puis, par récurrence : $e_i := (c_i + d_i + r_i) \bmod b$ et $r_{i+1} := (c_i + d_i + r_i) \operatorname{div} b$. Autrement dit, on a effectué une division euclidienne :

$$c_i + d_i + r_i = r_{i+1} b + e_i.$$

On vérifie par récurrence que les *retenues* r_i valent 0 ou 1 : c'est évident pour $i = 0$; puis, si $r_i \in \{0, 1\}$, on a $0 \leqslant c_i + d_i + r_i \leqslant (b-1) + (b-1) + 1 = 2b - 1$ et le quotient r_{i+1} de la division euclidienne par b vaut 0 ou 1. On voit enfin que, soit $c_\ell + d_\ell + r_\ell \leqslant b - 1$, et l'on a alors $m = \ell$ et $e_m = c_\ell + d_\ell + r_\ell$, soit $b \leqslant c_\ell + d_\ell + r_\ell \leqslant 2b - 1$, et l'on a alors $m = \ell + 1$ et $e_m = 1$.

Exercice 3.
Vérifier que q est bien la somme de n et de p.
Solution. Si l'on pose $r_{m+1} := 0$, on a, pour $0 \leqslant i \leqslant m$:

$$c_i + d_i + r_i = r_{i+1} b + e_i \qquad \text{d'où l'on tire} \qquad c_i b^i + d_i b^i + r_i b^i = r_{i+1} b^{i+1} + e_i b^i,$$

puis, en sommant pour $0 \leqslant i \leqslant m$: $n + p + \sum_{i=0}^{m} r_i b_i = \sum_{i=1}^{m+1} r_i b_i + q$, d'où la conclusion.

1.1.1 Des mathématiques à l'algorithme

La description précédente est écrite en style algorithmique des mathématiciens. Le style algorithmique des algorithmiciens est assez différent ! Il faut d'abord décrire les opérations élémentaires, ici, la table d'addition. Celle-ci prend en entrée deux chiffres $c, d \in [\![0, b-1]\!]$ et une retenue antérieure $r \in \{0, 1\}$ et rend comme résultat un couple (s, e) formé d'une retenue $s \in \{0, 1\}$ et d'un chiffre $e \in [\![0, b-1]\!]$ tels que $c + d + r = sb + e$. En style mathématique, $s := (c + d + r) \operatorname{div} b$, $e := (c + d + r) \bmod b$. En style algorithmique :

```
si c + d + r < b
  alors (s := 0 ; e := c + d + r)
  sinon (s := 1 ; e := c + d + r - b)
```

Il faut lire : « **s** reçoit la valeur **0** », etc. les *affectations* de la *variable* **s** par diverses valeurs ainsi calculées. Notons au passage la différence de traitement typographique entre les objets algorithmiques, comme les variables **c**, **d**, **r** et leurs contreparties mathématiques, les *valeurs* c, d, r que peuvent *contenir* ces variables. Si l'on veut éviter de calculer deux fois **c + d + r**, on peut introduire une *variable auxiliaire* ou *locale* **x** qui ne sert qu'à cet endroit :

```
x := c + d + r ;
si x < b
  alors (s := 0 ; e := x)
  sinon (s := 1 ; e := x - b)
```

Pratiquement, le fait qu'il s'agit d'une opération élémentaire signifie que l'on n'a pas à en connaître les détails d'exécution. On la considérera donc comme une *fonction* ou une *procédure* que l'algorithme principal pourra *appeler* :

```
(e,s) := add(c,d,r)              (* addition avec retenue *)
```

Remarquons qu'il y a ici une *affectation simultanée* des variables **s** et **e**. La petite phrase à droite est un *commentaire* qui rappelle au lecteur (plus ou moins en détail) ce dont il s'agit.

Il faut ensuite décrire les objets sur lesquels opère l'algorithme :

1. Les *données* sont les chiffres $c_k, \ldots, c_0$ et $d_\ell, \ldots, d_0$. On veut les conserver jusqu'au bout, elles sont donc rangées dans des *tableaux* respectivement notés **C[0..k]** et **D[0..l]**. Pour dénoter la valeur contenue dans la $i^{ème}$ case du tableau, c'est à dire c_i, on écrit simplement **C[i]**. On peut de plus supposer que $k = \ell$ quitte à rajouter au plus court des deux nombres des chiffres 0 en poids forts.
2. Le *résultat* sera rangé dans un tableau **E[0..m]**. Comme on ne connaît pas exactement son poids le plus fort (k ou $k+1$), on convient de le ranger dans le tableau **E[0..k+1]**.
3. Le calcul fournit des résultats auxiliaires, les retenues r_i. Mais chacune ne sert qu'une fois. L'algorithmicien est économe et décide donc de ne pas encombrer la mémoire de toutes ces retenues : il déclare une seule variable **r** qui sera successivement affectée de toutes ces valeurs.

Nous sommes maintenant en mesure d'écrire l'algorithme.

```
r := 0 ;
pour i de 0 a k faire
   (E[i],r) := add(C[i],D[i],r) ;
E[k+1] := r ;;
```

Pour la clarté des explications qui suivent, nous choisissons d'en écrire la version complète ; exceptionnellement, nous numérotons les lignes pour simplifier les commentaires.

```
1. r := 0 ;
2. pour i de 0 a k faire
3.   (x := C[i] + D[i] + r ;
4.    si x < b
5.       alors (E[i] := x ; r := 0)
6.       sinon (E[i] := x - b ; r := 1)) ;
7. E[k+1] := r ;;
```

1.1.2 Anatomie de cet algorithme

En principe, les objets qui nous intéressent sont des entiers naturels. Ici, l'écriture en base b nous permet de représenter les entiers par des suites finies de chiffres. Les données ont donc été rangées dans les tableaux **`C[0..k]`** et **`D[0..k]`** et le résultat sera récupéré dans le tableau **`E[0..k+1]`**. Cela sous-entend *l'occupation* ou *l'allocation* de diverses zones de la mémoire. L'espace mémoire nécessaire au stockage d'un tel tableau est en gros proportionnel au nombre de cases de ce tableau (ici, $k+1$, $k+1$ et $k+2$). L'espace mémoire nécessaire au stockage de l'une de ces cases est une constante faible qui ne dépend que de la base b et du codage des chiffres $0, \ldots, b-1$. Par exemple, le codage **`BCD`** (pour « binary coded decimal », en français : « décimal codé binaire ») correspond à la base dix, et chaque chiffre de 0 à 9 est représenté en binaire sur un demi octet (donc 4 bits). Dans le contexte de notre algorithme, on considère donc comme *taille* d'un entier le nombre $|n|_b$ des chiffres de son écriture en base b, l'unité de taille étant une telle case.

Les seuls autres objets de base ici manipulés sont la variable **`r`**, qui doit contenir un entier (ici, 0 ou 1) ; et la variable **`x`**, qui doit contenir un entier de $[\![0, 2b-1]\!]$. Le « compteur » **`i`** relève plutôt du « contrôle » du déroulement de l'algorithme (voir plus bas).

Les *instructions* de l'algorithme reposent sur des opérations de base qui sont les suivantes :

1. Opérations arithmétiques sur des nombres à un ou deux chiffres.
2. Divers *tests* : ici, tests de comparaison **`x < b`**.
3. Affectations, comme **`r := 0`** ou **`E[i] := x`**.

Il faut mentionner aussi des opérations « invisibles » comme *l'accès* aux composantes d'un tableau. Elles s'effectuent également en temps constant, précisément parce qu'il s'agit de tableaux. La pratique de l'informatique nous rappelle que d'autres structures de données analogues (c'est-à-dire dans lesquelles les éléments sont rangés consécutivement) ne possèdent pas cette propriété : en particulier, dans une *liste* ou dans un *fichier*, le temps d'accès à un élément dépend de la position de cet élément.

Le *déroulement* de l'algorithme est décrit par des *structures de contrôle* :

1. Le *séquencement* : notre algorithme est composé de trois étapes principales. La première est l'instruction écrite à la ligne 1 (c'est une affectation) ; la deuxième instruction, plus complexe, occupe les lignes 2 à 6 ; la troisième est l'affectation de la ligne 7. L'algorithme tout entier est composé du séquencement de ces trois instructions. C'est bien sûr une contrainte chronologique. De même, si l'on considère isolément le contenu des lignes 3 à 6, on voit qu'il s'agit d'une seule instruction composée du séquencement de deux instructions : la première est en ligne 3, la deuxième occupe les lignes 4 à 6, et la séquence qu'elles forment a été mise entre parenthèses pour

en faire une seule instruction. Cette mise entre parenthèses équivaut donc, dans nos notations, au classique **`debut ... fin`**, ou, si l'on préfère, **`begin ... end`**.

2. L'*itération* ou *répétition* : on peut voir que les lignes 2 à 6 représentent une seule instruction (elle-même un séquencement, d'ailleurs), laquelle dépend d'une sorte de paramètre **`i`**. Notons temporairement **`Proc1(i)`** cette instruction (que l'on peut donc assimiler à l'appel d'une procédure). Les lignes 3 à 6 équivalent à l'instruction suivante :

```
pour i de 0 a k faire
   Proc1(i) ;
```

Cette instruction, à son tour, équivaut au séquencement

```
Proc1(0) ; Proc1(1) ; ... ; Proc1(k) ;
```

Mais celui-ci ne peut être écrit de cette manière, puisque l'on ne sait pas concrètement combien d'instructions il comporterait ! C'est l'utilité de la *boucle « pour »* utilisée.

3. Le *branchement conditionnel* : l'instruction **`Proc1(i)`** se compose du séquencement de deux instructions dont la deuxième occupe les lignes 4 à 6. Cette dernière comprend un test (comparaison **`x < b`**) et deux instructions (elles-mêmes composées !) dont une seule sera effectuée.

Remarquons enfin que le type de répétition rencontré ici, également appelé *boucle inconditionnelle* ou encore *répétition définie*, n'est pas le plus général possible. En effet, si l'on ne sait pas, au moment où l'on écrit l'algorithme, combien de fois $Proc1(i)$ devra être exécutée (cela dépendra de la valeur de k), ce nombre de fois est totalement déterminé dès « l'entrée dans la boucle ». Dans certains algorithmes, ce nombre n'est pas connu et dépend des résultats des calculs : c'est le cas de l'algorithme qui va suivre.

1.1.3 Codage d'un entier dans une base

Pratiquement, comment obtient-on le codage de l'entier n dans la base b ? La question est ambiguë, car cela dépend de la forme sous laquelle est donnée l'entier n. Par exemple, si l'on demande l'écriture en base $b = 10$ du nombre $n = 1496$, la réponse est contenue dans la question. Si l'on part de l'écriture romaine $MCDXCVI$, on voit mieux où est le problème. Supposons donc que l'entier n est donné par son écriture dans une *autre* base et qu'on veut le *convertir*. Par exemple, n pourrait être le résultat codé en base 2 d'un calcul d'ordinateur, que l'on désirerait transcrire en base 10 pour usage humain.

Le procédé de conversion repose sur l'observation suivante : le quotient et le reste de la division euclidienne de $n = (c_k \dots c_0)_b$ par b sont respectivement $q = (c_k \dots c_1)_b$ et $r = c_0$. En effet, $q, r \in \mathbb{N}$ et $0 \leqslant r < b$ par définition ; de plus :

$$qb + r = b \sum_{i \geqslant 0} c_{i+1} b^i + c_0 = \sum_{i \geqslant 0} c_i b^i = n.$$

Il faut naturellement admettre que l'on sait effectuer des divisions euclidiennes sur des entiers naturels donnés dans la forme de départ choisie. L'algorithme suivant suppose l'entier n codé sous une forme quelconque, et il suppose aussi que l'on dispose d'une procédure

(ou fonction) **diveucl** telle que l'appel **diveucl(n,b)** rende le couple **(q,r)** résultat de la division euclidienne $n = qb + r$.

```
1. tant que n > 0 faire
2.     ((q,r) := diveucl(n,b) ;
3.      ecrire(r) ;
4.      n := q) ;;
```

Sans détailler à nouveau l'anatomie de cet algorithme, voici quelques observations :

1. La principale structure de contrôle est une répétition de la forme **tant que [condition] faire [instruction]**. La condition porte ici sur l'entier **n**. S'il est strictement positif, l'instruction est exécutée, puis l'on recommence. Ce type de répétition est donc appelé *boucle conditionnelle* ou encore *répétition indéfinie*. On peut noter **Proc2(n)** l'instruction qui occupe les lignes 2 à 4 : son effet est d'écrire un chiffre *et de modifier* n. La boucle est équivalente au séquencement :

   ```
   si n > 0 alors Proc2(n) ; si n > 0 alors Proc2(n) ; ...
   ```

   Bien sûr, il faut espérer que, l'entier **n** étant modifié par **Proc2(n)**, la condition **n > 0** finisse par être fausse, pour que l'algorithme s'arrête !
2. Les données **n** et **b** et les variables intermédiaires **q** et **r** sont présentes sous des formes non spécifiées, sur lesquelles la procédure **diveucl** peut opérer. De même, on fait ici abstraction de la manière dont est réalisée la division euclidienne. Ce n'est pas pour autant une véritable opération élémentaire, et l'on ne suppose pas qu'elle s'effectue en un temps constant : elle peut mettre en jeu un algorithme complexe.
3. Le résultat est la suite des chiffres de l'écriture de n en base b, mais ceux-ci ne sont pas stockés : on les écrit, et le problème de la structure de données qui permette de les conserver ne se pose donc pas.

Exercice 4.

L'algorithme ci-dessus écrit les chiffres de poids faibles en premier. Comment le modifier pour écrire les chiffres de poids fort d'abord ?

Solution. Il faut d'abord obtenir l'encadrement $b^k \leqslant n < b^{k+1}$; le chiffre de poids fort est alors le quotient q de la division euclidienne de $n = qb^k + r$. Puis on recommence avec le reste r. Le lecteur vérifiera que $1 \leqslant q < b$ et que $0 \leqslant r < b^k$, ce qui entraîne que $c_k = q \neq 0$. L'algorithme suivant ne produit que les chiffres non nuls, mais il donne leurs poids :

```
tant que n > 0 faire
   (k := 0 ; m := 1 ; M := b ;
    tant que M <= n faire
       (k := k + 1 ; m := M ; M := M * b) ;
    (q,r) := diveucl(n,m) ;
    ecrire("poids:") ; ecrire(k) ;
    ecrire("chiffre:") ; ecrire(q) ;
    n := r) ;;
```

La *boucle interne* **tant que M <= n faire ..** a pour but de calculer le poids k, et l'encadrement $m \leqslant n < M$ par des puissances successives $m = b^k$ et $M = b^{k+1}$ de la base.

1.2 Analyse de l'algorithme d'addition

Une fois un algorithme écrit, et avant même de s'en servir, il peut être utile de le juger. Une propriété fondamentale est évidemment la *correction* : l'algorithme a-t-il l'effet désiré ? Mais il est encore plus fondamental que l'algorithme *se termine*, sans quoi on ne pourrait rien juger du tout ! Enfin, étant donné un algorithme qui se termine et qui est correct, il possède des propriétés *opérationnelles* comme son temps d'exécution, les ressources qu'il consomme (espace mémoire, processeurs), etc. L'étude de toutes ces propriétés constitue *l'analyse* de l'algorithme, que nous allons illustrer sur l'algorithme d'addition et sur l'algorithme de conversion.

1.2.1 Terminaison

L'algorithme d'addition se compose de séquencements, d'un branchement conditionnel, et d'une répétition **pour i de 0 a k faire Proc1(i)**. Son exécution met en jeu un nombre fini d'opérations (que nous estimerons plus loin) et se déroule donc en temps fini.

L'algorithme de conversion se compose de séquencements, d'appels à la fonction **diveucl** et d'une boucle conditionnelle **tant que n > 0 faire Proc2(n)**. Nous admettrons que l'appel de **diveucl** (qui sous-entend l'exécution d'un autre algorithme) rend un résultat en temps fini. Il reste à vérifier que l'instruction **Proc2(n)** (lignes 2,3,4) ne s'exécute qu'un nombre fini de fois. Mathématiquement, cela se démontre ainsi : la variable **n** prend successivement les valeurs entières $n_0, n_1, \ldots$. L'entier n_0 est celui qui a été donné à l'algorithme. De plus, n_{k+1} est défini si, et seulement si, $n_k > 0$ et c'est le quotient de la division euclidienne de n_k par b. On a donc une suite *strictement décroissante* $n_0 > n_1 > \cdots > n_k$ d'entiers naturels, et cette suite ne peut pas être infinie.

Algorithmiquement, on raisonne plutôt ainsi. La variable **n** a une valeur *avant* l'exécution de **Proc2(n)** et une valeur *après* : notons les respectivement n et n'. Par hypothèse, $n > 0$ (condition d'exécution de **Proc2(n)**) et $n' = n \operatorname{div} b$ (définition de **Proc2(n)**), donc $0 \leqslant n' < n$. Ainsi l'entier naturel **Proc2(n)** décroît strictement à chaque itération de la boucle, ce qui ne peut se produire qu'un nombre fini de fois.

Exercice 5.

On peut se rendre compte des pièges que recèlent les boucles conditionnelles en examinant la variante suivante de notre algorithme de conversion : déceler l'anomalie

```
1. tant que n > 0 faire
2.     ((q,r) := diveucl(n,b) ;
3.      ecrire(r)) ;;
```

Solution. Ici, **n** n'est jamais modifié et l'algorithme écrit indéfiniment le même chiffre.

1.2.2 Correction

L'algorithme d'addition est correct parce qu'il reflète de manière triviale un théorème mathématique. L'algorithme de conversion est justifié mathématiquement par récurrence. Outre la notation $n_0, n_1, \ldots$ pour les valeurs successives de la variable **n**, introduisons les valeurs successives $r_0, r_1, \ldots$ de la variable **r**. Ce sont donc les chiffres écrits par l'algorithme. L'entier n_0 est celui que l'on veut convertir, puis l'on a, tant que $n_i > 0$, les divisions euclidiennes : $n_i = bn_{i+1} + r_i$. On a alors :

$$\forall k\ ,\ n_0 = \sum_{i=0}^{k-1} r_i b^i + b^k n_k, \quad \text{et} \quad r_0, \ldots, r_{k-1} \in [\![0, b-1]\!].$$

Le lecteur prouvera sans peine cette affirmation par récurrence. Comme il existe un plus petit entier k tel que $n_{k+1} = 0$, qu'il marque l'arrêt de l'algorithme et que les chiffres écrits sont alors $r_0, \ldots, r_k$, on a bien établi la correction.

Lorsqu'on raisonne algorithmiquement, le principe de récurrence est remplacé par la notion d'*invariant de boucle*. Cette notion (essentielle !) sera précisée lorsque nous parlerons de *correction* (voir page 406). Pour l'illustrer dans notre cas, rappelons que n_0 désigne l'entier à convertir, et que c'est la valeur de la variable **n** au départ. Nous allons prouver qu'après i itérations de la boucle, les chiffres écrits sont les i chiffres de poids faible de n_0 et la variable **n** a pour valeur l'entier formé des chiffres suivants. Autrement dit, si $n_0 = (c_k \ldots c_0)_b$, alors (après i itérations) les chiffres écrits sont $c_0, \ldots, c_i$ et **n** a pour valeur $(c_k \ldots c_{i+1})_b$. C'est évidemment vrai au départ. Supposons que cela soit vrai avant la $i^{ème}$ itération, et notons n et n' les valeurs de la variable **n** avant et après cette itération. Ainsi, $n = (c_k \ldots c_i)_b$ et les chiffres déjà écrits sont $c_0, \ldots, c_{i-1}$ (c'est l'hypothèse). Par définition de l'instruction **Proc2(n)**, on a une division euclidienne $n = bn' + r$, et r est le prochain chiffre écrit. La division euclidienne est ici :

$$(c_k \ldots c_i)_b = b(c_k \ldots c_{i+1})_b + c_i,$$

d'où $n' = (c_k \ldots c_{i+1})_b$ et $r = c_i$. L'invariant de boucle ayant été *prouvé*, reste à voir ce qu'il donne quand l'algorithme se termine. La valeur de **n** est alors 0, car c'est un entier naturel, et s'il était non nul, l'itération continuerait. Selon l'invariant de boucle, tous les chiffres de n_0 ont été écrits.

1.2.3 Coût

L'algorithme d'addition comporte deux instructions exécutées une seule fois (lignes 1 et 7) et l'instruction **Proc1(i)** exécutée $k+1$ fois. Nous admettrons que l'exécution de **Proc1(i)** prend toujours le même temps A. Cela pourrait être affiné si l'on tenait compte de la différence entre la ligne 5 et la ligne 6, mais il est plus raisonnable de penser à une « routine d'addition avec retenue » à temps constant. En conclusion, le coût *en temps* de l'algorithme est de la forme $A(k+1) + B$, où $k+1$ est le nombre de chiffres des nombres additionnés. On parle de *coût linéaire*. L'idée est que, si l'on double la taille des données (par exemple, si l'on additionne des nombres à 100 chiffres au lieu de 50), le temps d'exécution est à peu près doublé.

Il faut remarquer que l'on considère ici comme taille des données leur nombre de chiffres, ce qui ne va pas de soi. Si l'on considérait comme taille d'un entier l'entier lui-même, le

coût de l'addition serait plutôt *logarithmique* (puisque le nombre de chiffres de n est du même ordre de grandeur que $\log n$). Mais il est courant de considérer comme taille d'une donnée la taille de sa représentation informatique, ce qui nous amène à parler du coût *en espace* de l'algorithme d'addition. Il faut quelques variables et trois tableaux de $k+1$ chiffres, l'espace consommé est donc également linéaire en la taille des données. Dans les algorithmes arithmétiques courants, on s'intéresse rarement au coût en espace, qui ne pose habituellement pas de problème majeur.

L'algorithme de conversion fait appel à la fonction **`diveucl`** ce qui pose problème pour l'estimation de son coût. Nous ferons l'hypothèse simplificatrice que l'algorithme qui réalise cette division euclidienne s'exécute en temps et en espace constant. Ce n'est pas une hypothèse réaliste si les nombres manipulés sont grands, mais elle n'est pas aberrante dans le cas contraire. Dans ce cas, il est immédiat que notre algorithme utilise un espace constant (quelques variables ...) et que le temps d'exécution est essentiellement proportionnel au nombre d'itérations, c'est à dire au nombre de chiffres de n. Il peut sembler paradoxal de prendre comme base de calcul du coût ce nombre de chiffres, qui dépend justement du résultat de l'algorithme de conversion ! En fait, ce nombre de chiffres est en gros proportionnel à $\log n$ et son ordre de grandeur est donc indépendant du choix de la base b.

Nous pouvons affiner l'estimation du coût en temps de l'algorithme de conversion de n. Pour cela, nous supposerons que l'appel de **`diveucl(n,b)`** a un coût proportionnel à la taille de n. Une étude plus précise des algorithmes arithmétiques montrerait que c'est une hypothèse réaliste. Nous écrirons donc : $D(n) = A|n|_b$, où $D(n)$ désigne le coût de **`diveucl(n,b)`** et où $|n|_b$ désigne la taille de n, ici son nombre de chiffres en base b. Avec les notations précédentes, on voit donc que $C(n_0) = \sum_{i=0}^{k} (D(n_i) + B)$, où $C(n)$ désigne le coût de la conversion de n et B une constante (coût des lignes 3 et 4). Comme le nombre de chiffres de n_i est $i+1$, on trouve finalement :

$$C(n) = \sum_{i=0}^{k} (A(i+1) + B) = \frac{A}{2}|n|_b^2 + \left(\frac{A}{2} + B\right)|n|_b.$$

2 Vocabulaire

2.1 Langage algorithmique simplifié

Le langage de description des algorithmes est parfois appelé « pidgin algol » ou « pseudo-Pascal ». Nous n'en donnerons pas une description formelle, les nombreux exemples de ce module devraient suffire à le comprendre.

2.1.1 Les objets

Les *types* de base dont nous aurons l'usage sont les suivants :

1. Les entiers ; remarquons d'ailleurs que, du point de vue de certains algorithmes, les entiers sont eux-mêmes des objets composés. Par exemple, dans les exemples précédents, un entier était assimilé à un tableau de chiffres.

2. les booléens : leurs valeurs possibles sont **`vrai`** et **`faux`**, ils servent à représenter les résultats des tests.

On pourrait également prendre en compte d'autres types numériques, comme les réels, les complexes, ou encore les éléments d'un corps fini (*arithmétique modulaire*). Par ailleurs, les algorithmes de tri manipulent fréquemment des caractères et des chaînes de caractères.

Les objets que l'on peut manipuler sont des constantes et des variables de l'un de ces types, ou bien des objets plus structurés, comme des tableaux ou des listes (la nuance entre ces deux structures est qu'un tableau est en général de taille fixe alors qu'une liste peut être allongée ou raccourcie). Ainsi, un polynôme peut être assimilé à une liste de nombres, un vecteur à un tableau de nombres et une matrice à un tableau de tableaux de vecteurs. Dans ce module, nous ne manipulerons que des tableaux.

2.1.2 Les actions

Le langage algorithmique sert à décrire des actions sur ces objets. Un algorithme se présente comme une suite d'instructions et celles-ci comportent des opérations élémentaires. Voici les plus fondamentales :

1. Les calculs numériques (pour nous, dans $\mathbb{N}$ et dans $\mathbb{Z}$).
2. Les calculs logiques sur les booléens, mettant en jeu le **`et`**, le **`ou`**, le **`non`**, etc. (les règles en ont été décrites au module I.1, section 7).
3. Les affectations, qui ont la forme **`x := a`** ou bien **`T[i] := a`**, où **`x`** est une variable, **`T[i]`** une case d'un tableau et **`a`** une valeur du type convenable.
4. Les tests de conditions (égalité, inégalités, ...), qui ont pour résultats des booléens et qui peuvent s'insérer dans des calculs logiques ou bien dans des instructions conditionnelles (c'est-à-dire des branchements conditionnels ou bien des boucles conditionnelles).

Si l'on veut considérer un algorithme comme le processus de calcul d'une *fonction*, on le fera précéder d'une *déclaration* du nom de cette fonction indiquant ses arguments (ou *données* de l'algorithme) et suivre d'une instruction **`rendre [resultat]`**. Voici un exemple trivial :

```
Produit(p,q)
x := p * q ;
rendre(x) ;
```

On fait parfois appel à des *procédures*, qui ne rendent pas de résultat mais agissent par « effet de bord », par exemple en modifiant leurs arguments :

```
Echange(a,b)
temp := a ;
a := b ;
b := temp ;
```

C'est le cas des algorithmes de tri, qui réordonnent un tableau « sur place », c'est-à-dire en modifiant les valeurs contenues dans ce tableau.

2.1.3 Les structures de contrôle

Enfin, pour décrire la manière dont se déroule l'exécution des instructions de l'algorithme, on dispose de *structures de contrôle*. La principale remarque est que l'*on exclut tout parallélisme* : il y a une seule action à la fois. Nous avons :

1. Le *séquencement* : si **A**, **B**, **C** ... sont des instructions, alors **A ; B ; C ; ...** est l'instruction qui équivaut aux exécutions *successives* de **A**, **B**, **C** ...
2. Les instructions conditionnelles :

   si [condition] alors [instruction1] sinon [instruction2] ;

   Le sens est clair : on teste **[condition]**. Si le résultat est **vrai**, on exécute **[instruction1]** ; si le résultat est **faux**, on exécute **[instruction2]**. Il y a d'ailleurs une variante plus simple :

   si [condition] alors [instruction] ;

   Dans ce cas, si **[condition]** a la valeur **faux**, on ne fait rien. Il peut y avoir ambiguïté en cas d'imbrication : l'instruction
   si A alors si B alors C sinon D
   équivaut à l'une des deux suivantes :

```
si A
  alors si B
           alors C
           sinon rien
  sinon D ;
si A
  alors si B
           alors C
           sinon D
  sinon rien ;
```

   On voit l'intérêt de la version « complète » **si... alors ... sinon ...**, quitte à compléter par l'instruction vide **rien**.
3. Les *répétitions indéfinies* (ou *boucles conditionnelles*) :

   tant que [condition] faire [instruction] ;

   Encore une fois, le sens est clair : on teste **[condition]**. Si le résultat est **vrai**, on exécute **[instruction]**, *puis l'on recommence* ; si le résultat est **faux**, on sort de la boucle de répétition et l'on passe éventuellement à la suite. En voici un exemple. L'algorithme suivant calcule le produit des deux entiers naturels **p** et **q** :

```
Produit1(p,q)
x := 0 ; y := q ;
tant que y > 0 faire
  (x := x + p ; y := y - 1);
rendre(x) ;
```

Une variante (moins puissante) de la boucle **tant que** est la boucle **repete** :

```
repete [instruction] jusque [condition] ;
```

Dans ce cas, on exécute au moins une fois **[instruction]**, puis l'on teste **[condition]**. Si le résultat est **vrai**, on sort de la boucle de répétition et l'on passe éventuellement à la suite ; si le résultat est **faux**, on recommence. Ainsi, *si l'on suppose q strictement positif*, l'algorithme précédent équivaut au suivant :

```
Produit2(p,q)
x := 0 ; y := q ;
repete (x := x + p ; y := y - 1) jusque y = 0;
rendre(x) ;
```

4. Les *répétitions définies* (ou *boucles inconditionnelles*) :

```
pour i de 1 a n faire [instruction] ;
```

Comme on l'a déjà indiqué, la boucle **pour** se comprend mieux si l'on suppose que **[instruction]** est de la forme **Proc(i)**. Cette répétition est alors équivalente au séquencement :

```
Proc(1) ; Proc(2) ; ... ; Proc(n) ;
```

Naturellement, on pourrait tout aussi bien prendre 0 comme première valeur du compteur i. La boucle **pour** est équivalente à un cas particulier de la boucle **tant que** :

```
i := 1 ;
tant que i <= n faire
   (Proc(i) ;
    i := i+1) ;
```

L'algorithme de multiplication ci-dessus équivaut donc au suivant :

```
Produit3(p,q)
pour i de 1 a q faire
   x := x + p ;
rendre(x) ;
```

Exercice 6.

Écrire un algorithme qui calcule p^n en fonction de p et de n.

Solution. On utilise une variable **x** initialisée à la valeur $1 = p^0$ et qui prendra successivement les valeurs $p^0, \ldots, p^n$: on la multiplie donc n fois par p.

```
Puissance0(p,n)
x := 1 ;
pour i de 1 a n faire
   x := x * p ;
rendre(x) ;
```

2.2 Des mathématiques aux algorithmes

Alors qu'en mathématiques on définit une *suite* de valeurs $u_0, u_1, \ldots$, par exemple par la valeur initiale $u_0 := a$ et par la relation de récurrence $u_{n+1} := f(u_n)$, l'écriture correspondante d'un algorithme a l'allure suivante :

```
1. x := a ;
2. pour i de 1 a n faire
3.     x := f(x) ;
```

On dit que l'on a *initialisé* la variable **x** avec la valeur **a**. Une exécution de l'instruction **x := f(x)** est appelée *mise à jour* de la variable **x**. Après i *itérations*, (et donc i mises à jour), la valeur contenue dans **x** est u_i. Par exemple, l'algorithme **Puissance0(p,n)** traduit la suite définie par la valeur initiale $u_0 := 1$ et la relation de récurrence $u_{n+1} := pu_n$. Le cas d'une récurrence mutuelle (ou simultanée) se ramène trivialement à celui-ci, en remplaçant la variable **x** par plusieurs variables ou bien par un tableau. Par exemple, les suites définies par les valeurs initiales $u_0 := p, v_0 := q$ et la relation de récurrence mutuelle $\begin{cases} u_{n+1} := au_n + bv_n, \\ v_{n+1} := cu_n + dv_n \end{cases}$ sont calculées par l'algorithme :

```
x := p ; y := q ;
pour i de 1 a n faire
   (x,y) := (a*x + b*y,c*x + d*y) ;
```

Si l'on est perturbé par l'*affectation simultanée* **(x,y) := (a*x + b*y,c*x + d*y)**, on peut la remplacer par la séquence :

```
x0 := x ; y0 := y ; x := a*x0 + b*y0 ;y := c*x0 + d*y0
```

On a fait appel aux *variables tampons* **x0** et **y0**. De même, une récurrence à k pas $u_{n+k} := f(u_n, \ldots, u_{n+k-1})$ se ramène au schéma précédent à l'aide de k variables.

Exercice 7.

Traiter l'exemple de la suite de Fibonacci, définie par $F_0 := 0$, $F_1 := 1$ et, pour tout $n \in \mathbb{N}$, $F_{n+2} := F_n + F_{n+1}$.

Solution. Puisque chaque valeur de la suite dépend des deux valeurs précédentes, il convient d'utiliser deux variables **x** et **y** qui, à chaque étape de l'itération, contiennent les valeurs F_{i-1} et F_i. Elles seront donc initialisées avec les valeurs 0 et 1. L'itération repose sur la fonction de deux variables $f : (x, y) \mapsto (y, x + y)$, car $(F_i, F_{i+1}) = f(F_{i-1}, F_i)$.

```
Fibonacci(n)
si n = 0 alors rendre 0 sinon
si n = 1 alors rendre 1 sinon
   (x := 0 ; y := 1 ;
    pour i de 1 a n-1 faire
       (x,y) := (y,x+y) ;
    rendre(y)) ;
```

Voici un autre exemple, tiré du module II.6 : l'évaluation d'un polynôme

$$P(X) := a_0 + \cdots + a_d X^d$$

en x. On initialise une variable **S** à la valeur a_0, puis on l'incrémente successivement par les d termes $a_i x^i$. Nous représenterons les coefficients de **P** dans un tableau **a[0..d]**. Dans la variante la plus brutale, chaque terme est calculé à partir de rien :

```
Evaluation0(P,x)
S := a[0] ;
pour i de 1 a d faire
   (terme := a[i] ;
    pour j de 1 a i faire
       terme := terme * x ;
    S := S + terme) ;
rendre(S) ;
```

Dans une version plus économe de l'algorithme, on mémorise chaque puissance x^i pour calculer plus facilement la suivante ; la variable **xi** sert à contenir les x^i :

```
Evaluation1(P,x)
S := a[0] ; xi := 1 ;
pour i de 1 a d faire
  (t := t * x ;
   S := S + a[i] * t) ;
rendre(S) ;
```

Enfin, le schéma de Horner s'exprime comme suit :

```
Evaluation2(P,x)
S := a[d] ;
pour i de 1 a d faire
   S := S * x + a[d-i] ;
rendre(S) ;
```

2.2.1 Analyse d'algorithmes : la terminaison

La principale qualité d'un algorithme est de terminer à coup sûr, c'est-à-dire pour toute valeur des données. Le problème se pose pour des algorithmes où figure une répétition indéfinie, c'est-à-dire de l'une des formes suivantes :

```
tant que [condition] faire [instruction]
repete [instruction] jusque [condition]
```

Dans le premier cas (boucle **tant que**), il faut s'assurer qu'après un certain nombre d'exécutions de **[instruction]**, la valeur de **[condition]** (*condition de continuation*) devient **faux**. Dans le second cas (boucle **repete**), il faut au contraire s'assurer qu'après un certain nombre d'exécutions de **[instruction]**, la valeur de **[condition]** (*condition d'arrêt*) devient **vrai**.

Par exemple, lors de l'exécution de **`Produit1(p,q)`**, la variable **`y`** décroît de 1 à chaque étape. On ne peut donc avoir indéfiniment $y > 0$. Quelles que soient les valeurs de p et q, la terminaison est donc garantie. De même, lors de l'exécution de **`Produit1(p,q)`**, la variable décroît strictement à chaque étape. Si la valeur initiale q de **`y`** est un entier naturel non nul, cette variable ne prend que des valeurs entières et elle doit finir par s'annuler. Dans tous les autres cas (valeur initiale non entière ou non strictement positive), la condition de terminaison **`y = 0`** ne sera jamais réalisée et l'algorithme « bouclera ».
Toutes les méthodes de preuve de terminaison reposent en derrière analyse sur le fait que *toute suite strictement décroissante d'entiers naturels est finie*. On cherche donc à faire apparaître une quantité qui ne prend que des valeurs entières naturelles et qui décroît strictement à chaque étape. Cette quantité est parfois appelée *compteur*. Ce peut être la valeur contenue dans l'une des variables de l'algorithme, mais ce n'est pas nécessaire.

Lorsque l'on veut démontrer la terminaison de programmes *récursifs* généraux (voir le cours de L2), il peut arriver que l'on ait besoin de raisonner sur des ensembles ordonnés beaucoup plus compliqués que $\mathbb{N}$. C'est notamment le cas des « systèmes de réécriture ». Cependant, l'affirmation qui précède est valable pour les programmes *itératifs* du type considéré ici.

2.2.2 Analyse d'algorithmes : la correction

La deuxième qualité fondamentale d'un algorithme est sa correction : fait-il ce que l'on attend ? Outre l'arsenal des preuves mathématiques, les démonstrations de correction utilisent surtout des « invariants de boucle », qui sont une variante de la démonstration par récurrence. Ces invariants de boucle servent à maîtriser la principale difficulté non mathématique, qui vient de la distinction entre les *variables* et les *valeurs* prises par ces variables : au cours du déroulement de l'algorithme, une variable donnée pourra être affectée et prendre successivement différentes valeurs. On peut penser à la variable comme à un contenant et à sa valeur (à un stade donné de l'exécution) comme à un contenu.
Ainsi, lors de l'exécution de **`Produit1(p,q)`**, les variables **`x`** et **`y`** changent de valeur à chaque itération. Pour visualiser cela, nous noterons x, y les valeurs respectivement contenues dans **`x`**, **`y`** à un moment donné (la typographie est différente). Cependant, à chaque stade de l'exécution, on a la relation :

$$x + py = pq. \tag{1}$$

C'est une telle affirmation que l'on appelle *invariant de boucle*. Comment la démontre-t-on ? Rappelons que, dans (1), p et q désignent des valeurs fixées (les arguments de la fonction), alors que x et y désignent les contenus respectifs des variables **`x`** et **`y`**, qui évoluent. On démontre d'abord que la relation (1) est vérifiée à l'initialisation. Comme les variables **`x`** et **`y`** ont été initialisées avec les valeurs 0 et q, c'est évident. On démontre ensuite que, si la relation (1) est vérifiée *juste avant* une itération, alors elle est encore vérifiée *juste après*. Pour cela, nous allons utiliser une notation qui permet de distinguer la « valeur avant » de la « valeur après » d'une variable. Ainsi, nous noterons x, y les « valeurs avant » de **`x`**,**`y`** et x', y' leurs « valeurs après ». L'instruction de mise à jour **`(x := x + p ; y := y - 1)`** a pour conséquence les relations $x' = x + p$ et $y' = y - 1$. On a donc $x' + py' = (x + p) + p(y - 1) = x + py$. Ainsi, si $x + py = pq$ (l'invariant de boucle est valide « avant »), alors $x' + py' = pq$ (l'invariant de boucle est valide « après »), ce qu'il fallait démontrer.

En quoi la preuve d'un invariant de boucle nous aide-t-elle à établir la correction d'un algorithme ? Lorsque l'algorithme **`Produit1(p,q)`** se termine, la condition de répétition est devenue fausse donc, à la fin de l'algorithme, y vaut 0. L'invariant de boucle nous dit alors que x vaut pq. La variable **`x`** contient donc bien le produit des arguments. Cet exemple est évidemment très simple, mais nous détaillerons à la section 2.3 une application probante de cette méthode.

2.2.3 Analyse d'algorithmes : la complexité

La troisième propriété importante d'un algorithme est son coût (les algorithmiciens l'appellent également *complexité*). Nous nous intéresserons essentiellement au *coût en temps*, bien que le *coût en espace* ait parfois de l'importance. Dans les algorithmes mathématiques, on considère généralement comme coût le nombre d'opérations élémentaires effectuées. Dans presque tous les algorithmes qui comportent une répétition, le nombre d'opérations effectuées à chaque itération est « essentiellement constant » : cela signifie qu'il est encadré par deux nombres $a, b > 0$ fixés. Si la répétition a comporté n itérations, son coût vérifie donc : $an \leqslant C(n) \leqslant bn$. Autrement dit, on a à la fois $C(n) = O(n)$ et $n = O\big(C(n)\big)$, ce que l'on résume par la notation $C(n) = \Theta(n)$. On dit parfois que $C(n)$ *est de l'ordre de grandeur de* n. Pour évaluer le coût, on est donc en général ramené à évaluer le nombre d'itérations. Dans le cas d'une boucle **`pour`**, c'est trivial, mais dans le cas d'une répétition indéfinie, cela peut nécessiter des raisonnements sophistiqués. Pour certains algorithmes, le coût est essentiellement le même quelles que soient les données. Pour d'autres algorithmes, il dépend de la « taille » de ces données. Dans ce cas, on distinguera parfois le *coût moyen* du *coût maximum*, ou *coût le pire*. On cherchera alors à comprendre comment les fonctions $C_{moy}(N)$ et $C_{max}(N)$ dépendent de la taille N des données.

Une fois que l'on a évalué la fonction $C(N)$ (générale ou moyenne ou au pire), on la *classe*. Bien qu'il existe une théorie mathématique de la complexité des algorithmes, nous l'envisagerons plutôt comme une activité d'ingénieur, où l'on porte des jugements qualitatifs sur la valeur des algorithmes utilisés, en se fondant sur des estimations qui peuvent être approximatives. La principale *classe de complexité* des algorithmes est celle des fonctions de coût *linéaires*, c'est-à-dire telles que $C(N) = \Theta(N)$ ou simplement $C(N) = O(N)$. Il est d'usage de ne pas marquer de différence entre l'affirmation « tel algorithme ne comporte pas plus de aN opérations » et l'affirmation plus forte : « tel algorithme comporte de l'ordre de aN opérations ». Pour une telle fonction, on voit qu'en gros, lorsque la taille des données double, le coût double lui aussi.

Nous rencontrerons également la classe des coûts *polynomiaux* : $C(N) = \Theta(N^k)$ ou $C(N) = O(N^k)$. Par exemple, le produit de deux polynômes de degré d s'effectue en $\Theta(d^2)$ opérations. Lorsque l'on double le degré, le nombre d'opérations est à peu près multiplié par 4. De même, la multiplication de deux matrices carrées d'ordre N comporte $\Theta(N^3)$ opérations (voir les exercices). Lorsque l'on double la taille des matrices, le coût de la multiplication est donc multiplié par 8 (à peu près). Nous rencontrerons enfin la classe des coûts *logarithmiques* : $C(N) = \Theta(\log N)$ ou $C(N) = O(\log N)$; c'est par exemple celle de l'exponentiation « dichotomique ». Pour un tel algorithme, lorsque l'on double la taille des données, le coût est augmenté d'une constante.

2.3 Un exemple détaillé : l'algorithme d'Euclide

Nous nous référons ici au théorème 41 de la page 89 du module II.1, avec toutefois de légères modifications dans les notations. Soient $a \geqslant b > 0$ des entiers naturels. On définit une suite $x_0, x_1, \ldots$ de la manière suivante : $x_0 := a$, $x_1 := b$ puis, tant que $x_i \neq 0$, $x_{i+1} := x_{i-1} \mod x_i$ (rappelons que, pour $x \in \mathbb{N}$ et $y \in \mathbb{N}^*$, $x \mod y$ désigne le reste de la division euclidienne de x par y). Comme on l'a vu, pour $i \geqslant 1$, les x_i forment une suite strictement décroissante d'entiers naturels ; c'est donc une suite finie : $x_0 \geqslant x_1 > \cdots > x_k > x_{k+1} = 0$. Alors le dernier terme non nul de la suite est le pgcd de a et de b : $x_k = a \wedge b$. Nous allons retrouver ces propriétés par les méthodes de la section précédente.

Puisque cette suite est définie par une récurrence à deux pas, son écriture algorithmique met en jeu deux variables **x** et **y** qui, après i étapes, contiennent x_i et x_{i+1} :

```
PGCD(a,b)
x := a ; y := b ;
tant que y > 0 faire
   (x,y) := (y,x mod y) ;
rendre(x) ;
```

Si l'on veut éviter l'affectation simultanée, on peut la remplacer par la séquence : `(r := x mod y ; x := y ; y := r)`.

Le « compteur » qui garantit la terminaison de cet algorithme est le contenu y de la variable **y**. En effet, de la propriété $\forall x \in \mathbb{N}, y \in \mathbb{N}^*$, $0 \leqslant x \mod y < y$, on déduit que les valeurs prises par la variable **y** forment une suite strictement décroissante d'entiers naturels ; et une telle suite est nécessairement finie.

Nous prouverons l'invariant de boucle suivant : à tout stade de la répétition, x et y sont des entiers naturels et $x \wedge y = a \wedge b$. C'est en effet trivial à l'initialisation puisqu'alors $x = a$ et $y = b$. Notons x, y, x', y' les valeurs de **x** et **y** avant et après une itération : on a donc $x' = y$ et $y' = x \mod y$. De l'égalité (dont la preuve est laissée en exercice) $y \wedge (x \mod y) = x \wedge y$, on tire $x' \wedge y' = x \wedge y = a \wedge b$ et la conclusion s'ensuit.

Lorsque la boucle `tant que` s'arrête, on a $y \leqslant 0$ (par négation de la condition de continuation `y > 0`) ; comme y est un entier naturel, il est donc nul. Comme $x \wedge 0 = x$, l'invariant de boucle $x \wedge y = a \wedge b$ nous dit alors que $x = a \wedge b$, ce qu'il fallait démontrer.

2.3.1 Étude du coût

Nous supposerons que les opérations arithmétiques se font à coût constant. Avec les notations du début, le coût total d'exécution de l'algorithme est donc de la forme $Ak + B$, avec $A, B > 0$. L'étude du coût revient donc à l'étude de l'entier k comme fonction des entiers $a \geqslant b > 0$. Notons $K(a, b)$ cette fonction. Elle vérifie les relations suivantes : si b divise a, alors $K(a, b) = 1$; dans le cas contraire, $K(a, b) = K(b, a \mod b) + 1$.

Exercice 8.

Rappelons que la suite de Fibonacci est définie par $F_0 := 0$, $F_1 := 1$ et, pour $n \geqslant 1$, $F_{n+1} := F_n + F_{n-1}$. Calculer $K(F_{n+1}, F_n)$.

Solution. On vérifie aisément que cette suite est strictement croissante pour $n \geqslant 2$. On en déduit que $F_{n+1} \mod F_n = F_{n-1}$ pour $n \geqslant 3$. Prenant $a = F_{n+1}$ et $b = F_n$, on voit donc que $K(F_{n+1}, F_n) = n - 1$ pour $n \geqslant 2$. Nous allons voir qu'il s'agit du « cas le pire » de l'algorithme d'Euclide.

Théorème 1 (de Lamé). Supposons $k = K(a,b)$. Alors $a \geqslant F_{k+2}$ et $b \geqslant F_{k+1}$.

Démonstration. Si $k = 1$, il faut prouver que $a, b \geqslant 1$, ce qui est trivial. Nous supposerons donc $k \geqslant 2$. Reprenons la suite $x_0 \geqslant x_1 > \cdots > x_k > x_{k+1} = 0$. De ces inégalités et des divisions euclidiennes $x_{i-1} = q_i x_i + x_{i+1}$ (pour $i \in [\![1,k]\!]$), on déduit d'abord les inégalités $q_i \geqslant 1$, puis les inégalités $x_{i-1} \geqslant x_i + x_{i+1}$ (pour $i \in [\![1,k]\!]$). Comme $x_k \geqslant 1 = F_2$ et $x_{k-1} \geqslant 2 = F_3$, on montre par récurrence sur $i \in [\![0,k]\!]$ que $x_{k-i} \geqslant F_{i+2}$. En prenant $i = k-1$ et $i = k$, on obtient la conclusion voulue. ■

Corollaire 2.
Le coût $K(a,b)$ est majoré par le plus grand entier k tel que $F_{k+1} \leqslant b$ et $F_{k+2} \leqslant a$

Notons maintenant $\rho := \dfrac{1+\sqrt{5}}{2}$ et $\rho' := \dfrac{1-\sqrt{5}}{2}$.
La théorie des récurrences linéaires (exercice II.5.6 de la page 278) nous donne l'égalité : $F_k = \dfrac{\rho^k - \rho'^k}{\rho - \rho'}$, d'où l'on tire : $k \sim \dfrac{\log F_k}{\log \rho}$.
Le théorème de Lamé nous dit que le coût le pire envisageable est de l'ordre de grandeur de $\min(\log a, \log b)$. Pour l'étude (beaucoup plus difficile) du coût moyen, le lecteur consultera « The Art of Computer Programming » de Donald E. Knuth.

2.3.2 Algorithme d'Euclide étendu

On a vu au module II.6 qu'une variante de l'algorithme d'Euclide fournit des coefficients de Bézout, c'est-à-dire $u, v \in \mathbb{Z}$ tels que $a \wedge b = ua + vb$. En effet, chacun des x_i est de la forme $u_i a + v_i b$, où les couples (u_i, v_i) sont déterminés comme suit : $(u_0, v_0) := (1,0)$; $(u_1, v_1) := (0,1)$; et, pour $i \in [\![1,k]\!]$, notant $q_i := x_{i-1} \operatorname{div} x_i$, $(u_{i+1}, v_{i+1}) := (u_{i-1}, v_{i-1}) - q_i(u_i, v_i)$. Il suffit donc de poser $(u,v) := (u_k, v_k)$.

Pour retrouver les coefficients de Bézout par les méthodes de l'algorithmique, on fait appel à quatre nouvelles variables **s**, **t**, **u**, **v** (destinées à contenir respectivement les valeurs de $u_{i-1}, v_{i-1}, u_i, v_i$) et l'on veut que soient vérifiés les invariants de boucle suivants : $x = au + bv$ et $y = as + bt$. Il est donc clair que ces quatre variables doivent être initialisées aux valeurs $0, 1, 1, 0$. Pour écrire les mises à jour, nous partons des relations entre valeurs avant et après : $x' = y$ et $y' = x \bmod y = x - qy$, où $q = x \operatorname{div} y$. Il faut de plus que les relations $x = au + bv$ et $y = as + bt$ entraînent les relations $x' = au' + bv'$ et $y' = as' + bt'$, ce qui marchera parfaitement si l'on prend $(u', v') = (s, t)$ et $(s', t') = (u, v) - q(s, t)$. Voici l'algorithme étendu :

```
PGCD-BEZOUT(a,b)
x := a ; y := b ; u := 1 ; v := 0 ; s := 0 ; t := 1 ;
tant que y > 0 faire
   ((q,r) := diveucl(x,y) ;
    x := y ; y := r ;
    (u,s) := (s,u - q*s) ; (v,t) := (t , v - q*t)) ;
rendre (x,u,v) ;
```

Naturellement, l'appel à **diveucl(x,y)** doit retourner le quotient et le reste de la division euclidienne correspondante ; les affectations simultanées peuvent être remplacées par les séquences

```
temp := u ; u := s ; s := temp - q*s ;
temp := v ; v := t ; t := temp - q*t ;
```

L'analyse de l'algorithme étendu est en tous points similaire à celle de l'algorithme simple, à cela près que l'on doit prendre en compte les deux nouveaux invariants de boucle.

3 Quelques exemples fondamentaux

Les exemples qui suivent sont tirés de l'arithmétique, de la combinatoire et de l'algèbre. Il y a une omission notable, l'algèbre linéaire. L'algorithme le plus important dans ce domaine est l'algorithme du pivot et nous l'avons abondamment étudié au module II.4. De plus, sa programmation est nettement plus compliquée et représente un exercice un peu indigeste pour une initiation à l'algorithmique.

3.1 L'exponentiation dichotomique

Soit n un entier naturel et soit x un nombre entier, réel ou complexe.

Le calcul de x^n par la méthode la plus évidente comporte les $n-1$ multiplications $x^{k+1} = x^k * x$, $k = 1, \ldots, n-1$. Nous supposerons, pour simplifier, que le coût de ces multiplications est constant (voir les exercices II.8.1 et II.8.3 de la page 422 pour une hypothèse différente). Ainsi, le coût du calcul de x^n est-il essentiellement proportionnel à n.

Cependant, une méthode d'exponentiation plus rapide est connue depuis fort longtemps ; elle a d'ailleurs pris les noms d'exponentiation *chinoise* ou *indienne* ou *babylonienne*. Nous l'appellerons *dichotomique* car elle repose sur une des idées les plus fondamentales de l'algorithmique, *diviser pour régner* (en anglais : *divide and conquer*). Il s'agit de ramener le problème à deux sous-problèmes plus simples, et ainsi de suite, en espérant que la somme des coûts de résolution de ces sous-problèmes est inférieure au coût de résolution du problème de départ. C'est notamment le cas lorsqu'il y a *partage* entre les sous-problèmes.

Par exemple, on peut calculer x^{16} comme carré de x^8, celui-ci comme carré de x^4, etc. et l'on aura effectué 4 multiplications au lieu de 15. La méthode dichotomique marche aussi en présence d'exposants impairs, bien qu'avec un rendement légèrement inférieur. Ainsi, les égalités $x^{15} = (x^7)^2 * x$, $x^7 = (x^3)^2 * x$, $x^3 = (x)^2 * x$ permettent-elles de calculer x^{15} en 6 multiplications au lieu de 14. On peut même faire mieux : on calcule successivement $x^3 = x * x * x$, $x^6 = (x^3)^2$, $x^{12} = (x^6)^2$ et l'on obtient $x^{15} = x^{12} * x^3$ en 5 multiplications ! Cette dernière méthode est optimale, mais elle n'est pas aisément généralisable alors que la méthode dichotomique l'est.

```
Puissance1(a,n)
r := 1 ; x := a ; p := n ;
tant que p > 0 faire
   (si p impair r := r * x ;
    p := p div 2 ;
    x := x * x) ;
rendre(r);
```

La terminaison de **`Puissance1(a,n)`** est garantie par le compteur p (contenu de la variable **`p`**) : en effet, si $p \in \mathbb{N}^*$, alors $0 \leqslant p \operatorname{div} 2 < p$. La correction découlera de l'invariant de boucle suivant : $rx^p = a^n$. Il est trivialement vérifié à l'initialisation. La mise à jour de **`r`**, **`x`** et **`p`** donne les relations suivantes entre valeurs avant et valeurs après : $p = 2p' + \varepsilon$ pour un certain $\varepsilon \in \{0, 1\}$; $r' = rx^\varepsilon$; et $x' = x^2$. On a alors $r'x'^{p'} = rx^\varepsilon x^{2p'} = rx^p$, ce qu'il fallait démontrer. À la fin de l'itération, la condition de non continuation $p = 0$ entraîne que $r = a^n$, comme désiré. Avant de calculer le coût, notons que la dernière multiplication **`x := x * x`** est inutile. La version suivante de l'algorithme n'effectue que les opérations nécessaires.

```
Puissance2(a,n)
Si n = 0
  alors rendre(1)
  sinon si n = 1
    alors rendre x
    sinon
     (r := 1 ; x := a ; p := n ;
      tant que p > 1 faire
          (si p impair r := r * x ;
           p := p div 2 ;
           x := x * x) ;
      r := r * x;
      rendre(r)) ;
```

Nous ne comptabiliserons pas les opérations **`p := p div 2`** : si p est codé en binaire, elles correspondent à un décalage des bits et s'effectuent en un temps négligeable. À chaque itération de la boucle **`tant que`** sont effectuées 1 ou 2 multiplications, selon la parité de p. Le nombre d'itérations est égal à $\lfloor \log_2 n \rfloor$. Le coût de **`Puissance2(a,n)`** est donc $\Theta(\log n)$. L'exercice suivant permet de préciser ce résultat.

Exercice 9.

Combien de fois est effectuée l'opération **`r := r * x`** ?

Solution. Écrivons $n = (b_k \dots b_0)_2$ (écriture en base 2). Les valeurs successives de p sont donc les $(b_k \dots b_i)_2$, $0 \leqslant i \leqslant k$; il est impair chaque fois que $b_i = 1$. Le nombre cherché est donc le nombre de 1 dans l'écriture de n en base 2. Cela explique pourquoi l'algorithme est moins rentable avec l'exposant $15 = (1111)_2$ qu'avec l'exposant $16 = (10000)_2$.

En réalité, ce calcul comptabilise une multiplication par 1 (valeur initiale de r) qui pourrait être économisée, quitte à modifier légèrement l'algorithme. Nous conclurons que le coût $C(n)$ de **`Puissance2(a,n)`** (compté en nombre de multiplications) est égal à $\lfloor \log_2 n \rfloor + C'(n) - 1$, où $C'(n)$ désigne le nombre de chiffres 1 dans l'écriture de n en base 2. Cela sera retrouvé dans l'exercice II.8.4 de la page 422.

Comme on l'a vu sur l'exemple $n = 15$, il est parfois possible de calculer a^n avec moins de multiplications. Supposons par exemple que $n = pq$. On calcule d'abord $b = a^p$ par exponentiation dichotomique puis $c = b^q$ encore par exponentiation dichotomique. Est-ce plus

rentable que le calcul direct de $c = a^n$ par exponentiation dichotomique ? Autrement dit, a-t-on $C(pq) > C(p)+C(q)$? C'est bien le cas si $(p,q) = (3,5)$, car $C(3) = 2$, $C(5) = 3$ et $C(15) = 6$. En revanche, si $(p,q) = (3,7)$, cette astuce n'apporte rien car $C(21) = 6$ et $C(3)+C(7) = 2+4$. Si $(p,q) = (3,11)$ on y perd : $C(33) = 6 < C(3)+C(11) = 2+5$.

Dans tous les cas, *on ne peut pas améliorer l'ordre de grandeur* $\Theta(\log n)$. Considérons en effet une succession de k multiplications effectuées à partir de l'unique donnée de départ x ; on a pu ainsi calculer un certain nombre de puissances de x. Notons m_i le plus grand exposant ainsi apparu après i multiplications. On a donc $m_0 = 1$ (au départ, on ne dispose que de x) et la relation de récurrence $m_{i+1} \leqslant 2m_i$ (car en multipliant x^p et x^q tels que $p, q \leqslant m_i$, on obtient x^r, où $r = p + q \leqslant 2m_i$). On en déduit que $m_k \leqslant 2^k$. Pour que l'on puisse calculer x^n en k multiplications, il est donc nécessaire que $n \leqslant m_k \leqslant 2^k$, *i.e.* que $k \geqslant \log_2 n$.

3.2 Tris et permutations

La donnée de départ est un tableau **`T[1..n]`** d'objets que l'on veut ranger par ordre croissant. Cela sous-entend que ces objets sont d'un type pour lequel on dispose d'une relation d'ordre total. L'exemple le plus important dans la pratique est celui des chaînes de caractères, que l'on veut trier par ordre lexicographique : index terminologique d'un livre, liste des étudiants d'une filière, etc. Naturellement, seule l'existence de cet ordre total est pertinente pour décrire et comprendre les algorithmes, et nous supposerons que **`T[1..n]`** est un tableau d'entiers naturels et que la relation d'ordre utilisée est celle de $\mathbb{N}$.

Pour étudier les algorithmes de tri, *nous ferons l'hypothèse que le tableau à trier ne contient pas de duplicata,* autrement dit, les entiers $T[1], \ldots, T[n]$ sont deux à deux distincts. Cette hypothèse admet une justification « physique » : les valeurs qui figurent dans le tableau proviennent d'un domaine très vaste et n'en sont souvent qu'un échantillonnage restreint. Il est donc peu probable que deux valeurs soient égales. Par exemple, il est peu courant que deux étudiants d'un même amphi portent le même nom. De plus, on a observé empiriquement que, s'il y a peu de duplicata, le comportement des algorithmes est sensiblement le même que s'il n'y en a pas du tout. Il y a de plus une justification méthodologique : nous allons en effet constater que cette hypothèse permet d'analyser rigoureusement les algorithmes, ce qui serait beaucoup plus difficile sans elle.

Puisque les entiers $T[1], \ldots, T[n]$ ont été supposés deux à deux distincts et puisque seules leurs relations d'ordre comptent lorsque l'on veut les trier, on peut supposer que leurs valeurs sont les éléments de $[\![1, n]\!]$ mais en désordre. Ainsi est associée au tableau T une permutation de $[\![1, n]\!]$, celle qui à l'indice $i \in [\![1, n]\!]$ associe la valeur $T[i] \in [\![1, n]\!]$. Il y a $n!$ telles permutations et il y a donc $n!$ ordres possibles dans le tableau. Notre algorithme devra effectuer des comparaisons pour reconnaître l'ordre présent et des affectations pour rétablir l'ordre croissant. Ces affectations prendront la forme d'*échanges* **`T[i] <-> T[j]`**, cette dernière opération étant équivalente à la séquence **`(temp := T[i] ; T[i] := T[j] ; T[j] := temp)`**. Le coût de l'algorithme se comptera donc en comparaisons ou en échanges.

3.2.1 Le coût moyen optimal d'un algorithme de tri

Pour acquérir une intuition du problème, tentons de trier trois entiers a, b, c (supposés distincts) :

```
PetitTri1(a,b,c)
si a < b
  alors si b < c
          alors rendre(a,b,c)
          sinon si a < c
                alors rendre(a,c,b)
                sinon rendre(c,a,b)
  sinon si b < c
          alors si a < c
                  alors rendre(b,a,c)
                  sinon rendre(b,c,a)
          sinon rendre (c,b,a) ;
```

Ici, nous avons évité les échanges en créant directement dans chaque cas le triplet ordonné ; nous avons donc une *fonction*, qui renvoie le résultat des calculs. Dans le cas d'un tableau, cela ne sera évidemment pas possible, et des échanges seront nécessaires. Voici donc ici la version avec échanges (c'est donc une *procédure*, qui modifie les données) :

```
PetitTri2(a,b,c)
si a < b
  alors si b < c
          alors rien
          sinon si a < c
                alors b <-> c
                sinon (b <-> c ; a <-> b)
  sinon si b < c
          alors si a < c
                  alors a <-> b
                  sinon (a <-> b ; b <-> c)
          sinon a <-> c ;
```

Combien de comparaisons ont été effectuées ? Selon les cas deux ou trois. En considérant les six différents ordres comme équiprobables, on voit que le nombre de comparaisons est 2 avec probabilité $1/3$ et 3 avec probabilité $2/3$, soit une espérance de $8/3 \approx 2{,}67$. On peut estimer le nombre de comparaisons nécessaire en général de la manière suivante. Si k comparaisons sont effectuées par l'algorithme dans le pire des cas, cela signifie que l'on peut distinguer entre 2^k situations différentes. Comme il faut distinguer entre les $n!$ ordres possibles, on doit avoir $k \geqslant \log_2 n!$. Invoquant l'exercice II.8.6 de la page 422, on en déduit que k doit être au moins de l'ordre de grandeur de $n \log n$. Ce « raisonnement » est évidemment très approximatif, mais il est possible de le rendre rigoureux et l'on démontre que, pour tout algorithme de tri, le nombre moyen de comparaisons effectué pour classer un tableau de taille n est au moins de l'ordre de $\Theta(n \log n)$.

3.2.2 Tri par sélection

Le principe du tri par sélection est le suivant : on cherche le plus petit élément du tableau, soit $T[j]$ (c'est la sélection) ; on échange $T[j]$ avec $T[1]$. Puis l'on recommence avec la suite du tableau, etc.

```
Triselection(T)
pour i de 1 a n-1 faire
   (j := i ;
    pour k de i + 1 a n faire
       si T[k] < T[j] alors j := k ;
    T[j] <-> T[i]) ;
```

La terminaison est garantie, puisqu'il s'agit de boucles **pour**. L'invariant de boucle de la boucle « externe » (d'indice i) est le suivant : au démarrage de la $i^{ème}$ itération, les $i-1$ plus petits éléments du tableau sont en place, c'est à dire aux positions d'indices $1, \ldots, i-1$. L'invariant de boucle de la boucle « interne » (d'indice k) est le suivant : à tout moment, $T[j]$ est le plus petit parmi $T[i], \ldots, T[k]$. Le nombre de comparaisons effectué par l'algorithme est $\sum_{i=1}^{n-1} \sum_{k=i+1}^{n} 1 = \dfrac{n(n-1)}{2}$; le nombre d'échanges **T[j] <-> T[i]** est $n-1$. Ce nombre peut d'ailleurs être amélioré en modifiant ainsi la boucle interne :

```
j := i ; bouge := faux
pour k de i + 1 a n faire
   si T[k] < T[j] alors (j := k ; bouge := vrai) ;
si bouge alors T[j] <-> T[i] ;
```

Il n'est pas clair qu'une telle amélioration soit très sensible.

3.2.3 Tribulle

L'algorithme *tribulle* (en anglais *bubblesort*) n'est pas particulièrement performant, mais il est simple à écrire et à comprendre et son analyse est intéressante.

```
Tribulle(T)
pour i de 1 a n-1 faire
   pour j de 1 a n-i faire
      si T[j] > T[j+1] alors T[j] <-> T[j+1] ;
```

Pour se faire une idée intuitive du phénomène, observons l'effet de la « première passe » sur un exemple ; il s'agit de la première itération de la boucle externe : on fixe $i := 1$ et l'on observe une exécution complète de la boucle interne (j varie de 1 à $n-1$). Nous prenons $n := 7$ et le tableau $3, 1, 5, 2, 7, 6, 4$. À chaque étape, nous indiquons en gras l'entier qui se déplace vers la droite :

$$\mathbf{3}152764 \quad 1\mathbf{3}52764 \quad 13\mathbf{5}2764 \quad 132\mathbf{5}764 \quad 1325\mathbf{7}64 \quad 13256\mathbf{7}4 \quad 132564\mathbf{7}$$

Si l'on pense que la droite c'est le haut et que 7 est le plus léger, on voit bien l'analogie avec la montée des bulles (7 est la plus grosse bulle). Quoiqu'il en soit, la terminaison de l'algorithme est évidente puisqu'il ne comporte que des répétitions définies. Pour établir sa correction, on introduit des invariants de boucles. Pour la boucle externe, on constate que, juste après l'exécution de la $i^{ème}$ itération, les i cases de droite

du tableau contiennent les i plus grands éléments rangés par ordre croissant ; autrement dit, les éléments $T[n-i+1], \ldots, T[n]$ sont en position définitive. Pour la boucle interne, on constate que, juste après l'itération d'indice j, $T[j+1]$ contient le plus grand des éléments $T[1], \ldots, T[j+1]$. À ce stade du module, notre lecteur n'aura aucun mal à établir ces deux affirmations ! Le nombre de comparaisons lors d'une exécution est $\sum_{i=1}^{n-1} \sum_{j=1}^{n-i} 1 = \dfrac{n(n-1)}{2}$; il peut d'ailleurs être amélioré. Pour voir comment, regardons les passes suivantes dans l'exemple précédent. Voici la 2ème passe :

$$\mathbf{1}325647 \quad 1\mathbf{3}25647 \quad 12\mathbf{3}5647 \quad 123\mathbf{5}647 \quad 1235\mathbf{6}47 \quad 12354\mathbf{7}$$

Voici la 3ème passe :

$$\mathbf{1}235467 \quad 1\mathbf{2}35467 \quad 12\mathbf{3}5467 \quad 123\mathbf{5}467 \quad 1234\mathbf{5}67$$

Voici la 4ème passe :

$$\mathbf{1}234567 \quad 1\mathbf{2}34567 \quad 12\mathbf{3}4567 \quad 123\mathbf{4}567$$

Ici, rien n'a bougé, car le tableau était déjà trié à la fin de la passe précédente (mais on ne le savait pas). Toutes les passes suivantes ne serviront donc à rien. Si, lors d'une passe, on n'a détecté aucune inversion $T[j] > T[j+1]$ c'est que le tableau était déjà trié et il n'est pas nécessaire de poursuivre :

```
Tribulle2(T)
i := 1 ;
repete
   (inversion := faux ;
    pour j de 1 a n-i faire
       si T[j] > T[j+1]
         alors (T[j] <-> T[j+1] ; inversion := vrai) ;
    i := i + 1)
jusque ((i = n) ou (non inversion)) ;
```

Le nombre de comparaisons de l'algorithme ainsi amélioré dépend du nombre de passes, dont l'analyse est abordée dans l'exercice II.8.9 de la page 422. L'analyse du nombre d'échanges (qui est évidemment le même pour les deux variantes) occupera le reste de cette section.

Nous noterons sous forme de tableau à deux lignes une permutation $\sigma \in \mathfrak{S}_n$; la notation :

$$\sigma = \begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma_1 & \sigma_2 & \ldots & \sigma_n \end{pmatrix}$$

signifie donc que σ est la bijection $i \mapsto \sigma_i$ de l'ensemble $[\![1, n]\!]$ sur lui-même. Ainsi, σ_i et $\sigma(i)$ désignent le même élément et le choix de l'une ou l'autre notation dépend du contecte. Rappelons (module II.2, proposition 27 de la page 130) que l'on a défini le nombre d'inversions de la permutation σ comme l'entier :

$$N(\sigma) := \operatorname{card}\{(i,j) \mid 1 \leqslant i < j \leqslant n \text{ et } \sigma_i > \sigma_j\}.$$

Par exemple, notant, pour $1 \leqslant i < j \leqslant n$, $\tau_{i,j}$ la transposition de i et de j, on a $N(\tau_{i,j}) = 2(i-j) - 1$. La démonstration de cette égalité, ainsi que celle du lemme suivant, sont laissées en exercice au lecteur.

Lemme 3. Pour tout entier $k \in [\![1, n-1]\!]$, et pour toute permutation $\sigma \in \mathfrak{S}_n$, on a : $N(\sigma \circ \tau_{k,k+1}) = \begin{cases} N(\sigma) - 1 \text{ si } \sigma_k > \sigma_{k+1} \\ N(\sigma) + 1 \text{ si } \sigma_k < \sigma_{k+1} \end{cases}$

Théorème 4. Le nombre d'échanges lors d'une exécution de l'algorithme **`Tribulle`** est égal au nombre d'inversions de la permutation associée.

Démonstration. Soit T' le tableau obtenu à partir de T par échange de $T[k]$ et de $T[k+1]$, et soient $\sigma, \sigma' \in \mathfrak{S}_n$ les permutations associées. On a alors : $\sigma' = \sigma \circ \tau_{k,k+1}$. Or, l'algorithme **`Tribulle`** effectue un tel échange chaque fois qu'il rencontre une inversion de deux éléments contigus. D'après le lemme, chaque échange diminue donc le nombre d'inversions de 1. À la fin de l'algorithme, le tableau est trié et le nombre d'inversions est nul. La conclusion est alors immédiate. ∎

3.2.4 Étude du nombre d'inversions d'une permutation

Soit σ une permutation. Pour tout $i \in [\![1, n]\!]$, le nombre $N_i(\sigma)$ d'indices j tels que $i < j \leqslant n$ et $\sigma_i > \sigma_j$ est compris entre 0 et $n - i + 1$ (nombre total des indices j tels que $i < j \leqslant n$), d'où l'encadrement :

$$0 \leqslant N(\sigma) \leqslant \sum_{i=1}^{n} (n - i + 1) = \frac{n(n-1)}{2}.$$

Le minorant est atteint si, et seulement si, pour tout i, $i < j \Rightarrow \sigma_i < \sigma_j$, c'est-à-dire si σ est la permutation identité. Le majorant est atteint si, et seulement si, pour tout i, $i < j \Rightarrow \sigma_i > \sigma_j$, c'est-à-dire si σ est la permutation $\begin{pmatrix} 1 & 2 & \dots & n \\ n & n-1 & \dots & 1 \end{pmatrix}$.

Proposition 5. (i) Soient $\sigma \in \mathfrak{S}_n$ et $k \in [\![0, n]\!]$.
L'application σ' définie sur $[\![1, n+1]\!]$ par :

$$\sigma'(i) := \begin{cases} \sigma(i) \text{ si } i < n - k + 1 \\ n + 1 \text{ si } i = n - k + 1 \\ \sigma(i - 1) \text{ si } i > n - k + 1 \end{cases}$$

est une permutation de $[\![1, n+1]\!]$. Notons-la $\langle \sigma, k \rangle$. On a $N\big(\langle \sigma, k \rangle\big) = N(\sigma) + k$.
(ii) L'application $(\sigma, k) \mapsto \langle \sigma, k \rangle$ de $\mathfrak{S}_n \times \{0, \dots, n\}$ dans $\mathfrak{S}_{n+1}$ est une bijection.

Démonstration. Les deux premières affirmations deviennent évidentes si l'on comprend que σ' s'obtient à partir de σ en insérant la valeur $n + 1$ parmi $\sigma_1, \dots, \sigma_n$ de manière à laisser k valeurs à sa droite :

$$\sigma = \begin{pmatrix} 1 & \dots & n-k & n-k+1 & \dots & n \\ \sigma_1 & \dots & \sigma_{n-k} & \sigma_{n-k+1} & \dots & \sigma_n \end{pmatrix}$$

et :

$$\sigma' = \begin{pmatrix} 1 & \dots & n-k & n-k+1 & n-k+2 & \dots & n & n+1 \\ \sigma_1 & \dots & \sigma_{n-k} & \mathbf{n+1} & \sigma_{n-k+1} & \dots & \sigma_{n-1} & \sigma_n \end{pmatrix}.$$

On voit alors que, par rapport à σ, σ' présente k inversions de plus : les couples $(n-k+1, j)$ avec $n-k+1 < j \leqslant n+1$.
L'assertion (ii) est facile et laissée au lecteur. ■

Corollaire 6. Le nombre moyen d'inversions d'une permutation de $\mathfrak{S}_n$ est :

$$\overline{N}_n = \frac{n(n-1)}{4}.$$

Démonstration. Pour $n = 1$, la seule permutation est l'identité et son nombre d'inversions est 0, d'où $\overline{N}_1 = 0$. De la bijection obtenue au point (ii) de la proposition et de la formule donnée en (i), on déduit :

$$\overline{N}_{n+1} = \frac{1}{(n+1)!} \sum_{\sigma \in \mathfrak{S}_n} \sum_{k=0}^{n} \big(N(\sigma) + k\big).$$

Un petit calcul donne alors $\overline{N}_{n+1} = \overline{N}_n + \dfrac{n}{2}$, d'où la conclusion par récurrence. ■

Corollaire 7. (i) Lors d'une exécution de **`Tribulle`**, le nombre minimum d'échanges est 0, et il correspond au cas d'un tableau déjà trié.
(ii) Le nombre maximum d'échanges est $\dfrac{n(n-1)}{2}$, et il correspond au cas d'un tableau complètement inversé (décroissant).
(i) Le nombre moyen d'échanges est $\dfrac{n(n-1)}{4}$.

3.3 Polynômes

Les opérations sur les polynômes sont d'abord des opérations sur leurs coefficients et il est naturel de considérer comme *taille* $|P|$ d'un polynôme P le nombre de ses coefficients, c'est-à-dire $\deg P + 1$. Un polynôme de taille n est donc de la forme $P(X) := a_0 + \cdots + a_{n-1}X^{n-1}$, et il sera représenté par un tableau de nombres **`a[0..n-1]`** de taille n. Il peut arriver que le coefficient dominant s'annule (donc $\deg P < n-1$) mais nous dirons tout de même que P est de taille n dans ce cas.

Nous supposerons que les opérations sur ces nombres se font en temps constant. Le coût d'un algorithme sur les polynômes sera donc assimilé au nombre total de telles opérations ; nous ne distinguerons donc pas entre additions, multiplications, etc. Par exemple, le coût de la multiplication de P par un scalaire est $|P| = \deg P + 1$ et le coût de l'addition de P et de Q est $\min(|P|, |Q|) = 1 + \min(\deg P, \deg Q)$. La multiplication de P par un monôme X^e se ramène à un simple décalage des coefficients et s'effectue donc en temps $\Theta(|P|)$.

3.3.1 Multiplication simple

Pour évaluer le coût de la multiplication de deux polynômes P et Q, notons $P(X) := a_0 + \cdots + a_d X^d$, $Q(X) := b_0 + \cdots + b_e X^e$ et $PQ(X) = c_0 + \cdots + c_f X^f$; nous avons donc $f = d + e$ et, pour $k \in [\![0, f]\!]$: $c_k = \sum_{i+j=k} a_i b_j$. Au lieu de parcourir le tableau $C[0..f]$ qui code PQ, il est plus simple de parcourir les tableaux $A[0..d]$ et $B[0..e]$ qui codent respectivement P et Q :

```
MultPols(A,B,C)
pour k de 0 a f faire C[k] := 0 ;
pour i de 0 a d faire
   pour j de 0 a e faire
      C[i+j] := C[i+j] + A[i] * B[j] ;
```

Il y a donc $(d+1)(e+1)$ additions et $(d+1)(e+1)$ multiplications. On aurait pu économiser $(f+1)$ additions en initialisant $C[k]$ à une autre valeur que 0 (l'une des valeurs `A[i] * B[j]`), mais l'algorithme aurait été moins simple pour une faible amélioration.

3.3.2 Division euclidienne

L'algorithme de la division euclidienne décrit au module II.6 s'écrit ainsi :

```
Diveucl(A,B)
R := A ; Q := 0 ;
tant que deg R >= deg B faire
   (c X^e := td(R) / td(B) ;
    R := R - c X^e B ;
    Q := Q + c X^e) ;
rendre(Q,R) ;
```

Le compteur qui garantit la terminaison est $\deg R$. En effet, si $\deg R > \deg B$, la valeur après $R' = R - \frac{\operatorname{td} R}{\operatorname{td} B} B$ est telle que $\deg R' < \deg R$. L'invariant de boucle qui garantit la correction de l'algorithme est la relation $A = QB + R$. Sa validité à l'initialisation est triviale. Les relations $R' = R - \frac{\operatorname{td} R}{\operatorname{td} B} B$ et $Q' = Q + \frac{\operatorname{td} R}{\operatorname{td} B}$ montrent que $Q'B + R' = QB + R$, d'où la conservation de l'invariant à chaque itération. En sortie de répétition, la même relation jointe à la condition de non continuation $\deg R < \deg B$ entraîne que l'on a bien une division euclidienne.

À chaque itération, on effectue les opérations suivantes :

1. La division de monômes $\frac{\operatorname{td} R}{\operatorname{td} B}$: il y a donc une soustraction d'entiers (les exposants) et une division de nombres (les coefficients), le coût est constant ;
2. l'addition du monôme $\frac{\operatorname{td} R}{\operatorname{td} B}$ à Q : un seul coefficient de Q est modifié, c'est une addition de nombres et son coût est constant ;
3. l'addition de $\frac{\operatorname{td} R}{\operatorname{td} B} B$ à R, qui revient à $\deg B + 1$ modifications de coefficients, des additions de nombres.

Le coût total d'une itération est donc $\Theta(\deg B)$. Le nombre d'itérations est, dans le pire des cas, $\deg A - \deg B + 1$ (on admet ici que $\deg A \geqslant \deg B$). Le coût total de la division euclidienne des polynômes dans le pire des cas est donc :

$$\Theta\big((\deg A - \deg B)\deg B\big) = \Theta\big((|A| - |B|)|B|\big).$$

Remarque. En fait, le pire des cas est le cas général. On peut en effet démontrer que la probabilité que le degré de R chute de plus de 1 lors d'une itération est nulle (on admet ici que A et B sont des polynômes à coefficients dans $\mathbb{Q}$, $\mathbb{R}$ ou $\mathbb{C}$, plus généralement dans un corps infini). Pour des polynômes A et B pris « au hasard » (avec $\deg A \geqslant \deg B$), le nombre d'itérations est donc exactement $\deg A - \deg B + 1$.

3.3.3 Multiplication dichotomique

Une nouvelle application de l'idée de dichotomie (« diviser pour régner »), due à Karatsuba, va nous permettre d'améliorer substantiellement les performances de la multiplication des polynômes. Pour simplifier l'exposition, nous nous restreindrons au cas de deux polynômes de même taille, c'est-à-dire de même degré. Comme pour beaucoup d'algorithmes dichotomiques, la programmation de celui-ci en style « itératif » est très compliquée : c'est une programmation récursive qui s'imposerait. Cela relève du cours de L2.

Notons $C(n)$ le coût de la multiplication de deux polynômes de taille n. Si l'on emploie la méthode la plus évidente, nous avons vu que ce coût est *quadratique*, autrement dit, polynomial de degré 2 : $C(n) \sim an^2$. Cela entraîne que *lorsque la taille n double, le coût quadruple* : pour n grand, $\lim\limits_{n\to+\infty} \dfrac{C(2n)}{C(n)} = 4$. Si nous pouvons diminuer ce nombre 4, la performance en sera améliorée d'autant. Dans l'analyse qui suit, nous noterons encore $C(n)$ le coût de la multiplication de deux polynômes de taille n.

Prenons donc deux polynômes P et Q de taille $2n$ (c'est-à-dire de degré $2n-1$), que nous « coupons en deux » : nous les écrivons $P = P_0 + X^n P_1$ et $Q = Q_0 + X^n Q_1$, où P_0, P_1, Q_0, Q_1 sont de taille n. Nous utiliserons la formule magique (dont la vérification est laissée en exercice au lecteur) :

$$PQ = P_0Q_0 + X^n\big((P_0 + P_1)(Q_0 + Q_1) - P_0Q_0 - P_1Q_1\big) + X^{2n}P_1Q_1.$$

Cette formule ramène la multiplication de deux polynômes de taille $2n$ aux opérations suivantes :

1. Des additions de polynômes de tailles inférieures ou égales à $2n$ (notons en effet que la multiplication par X^n du terme central revient à un simple décalage dans le tableau des coefficients) ;
2. trois produits de polynômes de taille n : P_0Q_0, P_1Q_1 et $(P_0 + P_1)(Q_0 + Q_1)$.

On en déduit l'existence d'une constante $b > 0$ telle que :

$$C(2n) = 3C(n) + bn.$$

Ainsi, même si l'on effectue le produit des polynômes de taille n par la méthode « simple », on trouvera $C(2n) \sim 3an^2$ au lieu de $4an^2$, ce qui est déjà une amélioration. Cependant,

l'idée de la méthode dichotomique va plus loin. Par exemple, pour multiplier astucieusement deux polynômes de taille 8, on se ramène à trois produits de taille 4 (et des additions) ; ces trois produits se ramènent à 9 produits de taille 2, donc à 27 multiplications de nombres, au lieu de 64 ! Bien sûr, il y a également des additions et nous allons faire un compte plus précis.

Nous supposerons d'abord que nous n'opérons sur des polynômes dont la taille est une puissance de 2. Ainsi, on peut effectivement couper les tailles en 2 jusqu'à se ramener à des multiplications de nombres. Notons donc, pour $k \geqslant 0$, $u_k := C(2^k)$. Par exemple, u_0 est le coût de la multiplication de deux nombres. De plus, la relation entre $C(2n)$ et $C(n)$ nous donne :

$$\forall k \geqslant 0\ ,\ u_{k+1} = 3u_k + b2^k.$$

Ce type de récurrence se résout en écrivant $\dfrac{u_{k+1}}{3^{k+1}} - \dfrac{u_k}{3^k} = b\dfrac{2^k}{3^{k+1}}$, puis en sommant (télescopage). On obtient :

$$\forall k \geqslant 0\ ,\ C(2^k) = 3^k(u_0 + b) - b2^k.$$

Ainsi, lorsque n est une puissance de 2, $C(n) \sim Kn^\alpha$, où $K := u_0+b > 0$ et $\alpha := \log_2 3$. Lorsque n est un entier quelconque, on complète les polynômes à multiplier par des coefficients nuls de manière à obtenir des polynômes de taille 2^k, avec $k \geqslant \log_2 n > k-1$, c'est-à-dire $k = \lceil \log_2 n \rceil$ (on désigne par $\lceil x \rceil$ l'entier immédiatement supérieur au réel x). On obtient ainsi l'estimation $C(n) \sim K2^{\alpha\lceil \log_2 n \rceil}$.

Théorème 8. Le coût de la multiplication dichotomique de deux polynômes de taille n est $\Theta(n^\alpha)$, où $\alpha = \log_2 3$.

Démonstration. De l'encadrement $\log_2 n - 1 \leqslant \lceil \log_2 n \rceil \leqslant \log_2 n$, on tire :

$$\frac{n^\alpha}{3} \leqslant 2^{\alpha\lceil \log_2 n \rceil} \leqslant n^\alpha \qquad \text{(car } 2^\alpha = 3\text{).}$$

On peut apprécier le gain par rapport au coût quadratique (exposant 2) en remarquant que $\alpha = \log_2 3 \approx 1{,}585 < 2$.

Exercice type corrigé

II.8.0 1) Soit `T[1..n]` un tableau d'entiers de $[\![1, n]\!]$ deux à deux distincts.
Expliquer l'algorithme suivant :

```
i := 1;
tant que (i < n et T[i] < T[i+1]) faire i := i + 1 ;
rendre(i=n) ;
```

2) On prend pour critère de coût le nombre de comparaisons de la forme `T[i] < T[i+1]`. Décrire le cas le meilleur et le cas le pire.
3) On introduit la permutation $\sigma : i \mapsto \sigma_i = T[i]$ et l'on définit sa *première montée* $\Lambda(\sigma)$ comme le plus grand entier k tel que $\sigma_1 < \cdots < \sigma_k$. (Si $\sigma_1 > \sigma_2$, $\Lambda(\sigma) := 1$.)
Exprimer le coût de l'algorithme en fonction de $\Lambda(\sigma)$.
4) Pour $k \in [\![1, n]\!]$, calculer le nombre de $\sigma \in \mathfrak{S}_n$ telles que $\Lambda(\sigma) \geqslant k$.
5) En déduire le coût moyen de l'algorithme.

Solution. 1) La répétition s'arrête lorsque $i = n$ ou bien lorsque $T[i] > T[i+1]$. Dans le premier cas, le tableau est trié et la condition **`i=n`** vaut **`vrai`** ; dans le second cas, le tableau n'est pas trié et la condition **`i=n`** vaut **`faux`**. Cet algorithme rend donc la valeur **`vrai`** si le tableau est trié et la valeur **`faux`** dans le cas contraire.
2) Si $n = 1$, il n'y a aucune comparaison. Si $n = 2$, il y a 1 comparaison dans tous les cas. Supposons donc $n \geqslant 3$. Le nombre minimum de comparaisons est 1. Il correspond au cas où $T[1] > T[2]$ (le tableau n'est pas trié et l'on s'en rend compte dès le début). Le nombre maximum de comparaisons est $n-1$. Il correspond au cas où $T[1] < \cdots < T[n-1]$; il se produit si le tableau est trié mais également si l'une des valeurs $1, \ldots, n-1$ a été « déplacée » à la fin du tableau (ce cas correspond donc à n configurations).
3) Le coût est égal à $\Lambda(\sigma)$ si $\Lambda(\sigma) < n$, et à $n-1$ si $\Lambda(\sigma) = n$ (c'est-à-dire si le tableau est trié) ; c'est donc, dans tous les cas, $\min\big(\Lambda(\sigma), n-1\big)$.
4) On a $\Lambda(\sigma) \geqslant k$ si, et seulement si, $\sigma_1 < \cdots < \sigma_k$. Pour tout $\sigma \in \mathfrak{S}_n$, il y a $k!$ ordres possibles des k premiers éléments et, pour chacun de ces ordres, le même nombre A de $\sigma \in \mathfrak{S}_n$ qui lui correspondent : donc $k!A = n!$. Le nombre de $\sigma \in \mathfrak{S}_n$ qui correspondent à l'ordre croissant est donc $A = \dfrac{n!}{k!}\cdot$
5) Pour $k \in [\![1, n-1]\!]$, le nombre de permutations telles que $\Lambda(\sigma) = k$ est donc égal à $\dfrac{n!}{k!} - \dfrac{n!}{(k+1)!}\cdot$ Le coût moyen recherché est donc :

$$\frac{1}{n!}\left(\sum_{\sigma \neq Id} \Lambda(\sigma) + n - 1\right) = \sum_{k=1}^{n-1}\left(\frac{k}{k!} - \frac{k}{(k+1)!}\right) + \frac{n-1}{n!} = \sum_{k=1}^{n-1}\frac{1}{k!},$$

formule que l'on vérifiera aisément par exemple pour $n = 3$ (on trouve $3/2$). Remarquons que cette expression tend vers $\mathrm{e} - 1 \approx 1{,}718$ lorsque $n \to \infty$.

On suppose ici que, si **`i < n`** vaut **`faux`**, alors l'expression booléenne

```
(i < n et T[i] < T[i+1])
```

est directement évaluée à la valeur **`faux`** sans qu'il soit nécessaire d'évaluer **`T[i] < T[i+1]`**. Si ce n'était pas le cas, l'algorithme tenterait d'accéder à la valeur de **`T[n+1]`**, qui n'a pas de sens ; dans la plupart des langages de programmations, cela provoquerait une erreur à l'exécution du programme. On appelle *évaluation paresseuse* cette manière de calculer des booléens. Une conséquence en est que la condition **`1 = 0 et 1 = 1/0`** est évaluée à **`faux`**, alors que la condition **`1 = 1/0 et 1 = 0`** ne peut être évaluée (erreur « division par zéro »).

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercice sur la section 1

II.8.1 Écrire un algorithme pour la multiplication $m \times n$ en base b dont le coût est $\Theta(|m||n|)$ (« modèle des coûts bilinéaires »).

Exercice sur la section 2

II.8.2 Démontrer les propriétés suivantes invoquées lors de l'étude de l'algorithme d'Euclide : $y \wedge (x \mod y) = x \wedge y$. La suite de Fibonacci est strictement croissante pour $n \geqslant 2$.

Pour $i \in [\![0,k]\!]$, $x_{k-i} \geqslant F_{i+2}$. Pour $k \in \mathbb{N}$, $F_k = \dfrac{\rho^k - \rho'^k}{\rho - \rho'}$. En déduire que $k \sim \dfrac{\log F_k}{\log \rho}$.

Exercices sur la section 3

II.8.3 $*$ Évaluer le coût de l'exponentiation normale et de l'exponentiation dichotomique avec le modèle des coûts bilinéaires.

II.8.4 Soit $C(n)$ le nombre de multiplications lors de l'exécution de **`Puissance2(x,n)`**. Que vaut $C(1)$? Exprimer $C(2n)$ et $C(2n+1)$ en fonction de $C(n)$. Retrouver ainsi le calcul de $C(n)$ en fonction de n.

II.8.5
Quel est le nombre moyen d'échanges lors d'une exécution de l'algorithme **`PetitTri2`** ?

II.8.6 En comparant $\sum_{k=1}^{n} \ln k$ à $\int_1^n \ln t \, dt$, montrer que $\ln n! \sim n \ln n$.

II.8.7 (i) Démontrer la formule $N(\tau_{i,j}) = 2(i-j) - 1$ et celles du lemme 3 de la page 416.
(ii) Vérifier le petit calcul qui donne $\overline{N}_{n+1} = \overline{N}_n + \dfrac{n}{2}$ (démonstration du corollaire 6 de la page 417).

II.8.8 $*$ On définit le tableau des inversions d'une permutation $\sigma \in \mathfrak{S}_n$ comme le n-uplet $\big(b_1(\sigma), \ldots, b_n(\sigma)\big)$, où :

$$\forall i \in \{1, \ldots, n\}\ ,\ b_i(\sigma) = \operatorname{card}\{j \in [\![1,n]\!] \mid j < \sigma^{-1}(i) \text{ et } \sigma(j) > i\}.$$

(i) Vérifier que $b_i(\sigma) \in [\![0, n-i]\!]$ et que $\sum_{i=1}^{n} b_i(\sigma) = N(\sigma)$.
(ii) Décrire un algorithme qui permet de reconstituer σ à partir de son tableau d'inversions et en déduire une bijection de $\mathfrak{S}_n$ sur le produit cartésien $\{0\} \times \{0,1\} \times \cdots \times \{0, \ldots, n-1\}$.

II.8.9 Décrire l'effet d'une passe de l'algorithme **`Tribulle2`** sur le tableau des inversions. En déduire que le nombre de passes de l'algorithme est $1 + \max\big(b_1(\sigma), \ldots, b_n(\sigma)\big)$.

II.8.10 Écrire l'algorithme d'addition de deux polynômes de même degré et estimer son coût.

II.8.11 $*$ Dans le produit PQ de deux polynômes, calculer le nombre N_k des termes $a_i b_j$ qui interviennent dans c_k.
Distinguer les cas $k \leqslant \min(d,e)$, $\min(d,e) \leqslant k \leqslant \max(d,e)$, $\max(d,e) \leqslant k \leqslant f$.
Calculer à nouveau le nombre total de multiplications (c'est $\sum_{k=0}^{f} N_k$).

II.8.12 Majorer le coût de l'algorithme d'Euclide pour les polynômes en tenant compte de l'estimation du coût de la division euclidienne.

II.8.13 (i) Évaluer le nombre d'additions et de soustractions nécessaires pour additionner et pour multiplier deux matrices carrées.
(ii) Une méthode dichotomique de multiplication due à Strassen permet de ramener la multiplications de deux matrices carrées d'ordre n aux opérations suivantes : 18 additions de matrices carrées d'ordre n ; 7 multiplications de matrices carrées d'ordre n. Montrer que le coût de cette multiplication vérifie la relation de récurrence : $C(2n) = 7C(n) + an^2$ (où $a > 0$). En déduire que $C(n) = \Theta(n^\beta)$, où $\beta = \log_2 7 \approx 2{,}81 < 3$.

Partie III

Géométrie

Les mathématiques sont nées du point et du nombre ; aujourd'hui le second (via ses extensions) est souverain, mais ce fut longtemps le contraire. La tradition grecque régnait encore sans partage en 1794 quand Legendre proposait une édition d'Euclide, proche de l'original, qui restera une Bible pour un siècle (l'enseignement secondaire en était encore influencé en 1950). Au nom d'Euclide, il faut ajouter ceux d'Archimède et d'Apollonius, qui n'eurent guère de postérité créative pendant quinze siècles. Ce si long règne s'explique par deux points essentiels : une technique axiomatique clarifiant les règles du jeu d'une part, le fait que le sujet repose sur des bases intuitives fortes liées à l'expérience spatiale d'autre part. Toutefois les axiomes euclidiens étaient insuffisants et la géométrie finira par souffrir de la comparaison avec l'algèbre et l'analyse, aux fondements plus sûrs.
Les découvertes simultanées de la géométrie analytique par Descartes en 1637 dans le *Discours de la Méthode* et Fermat, puis son exploitation intensive par Euler – découvrant par le *calcul* la droite qui porte son nom ! –, ouvrirent de nouveaux horizons, notamment par la création d'objets d'étude à l'infini. La géométrie devint ainsi un outil pour la physique et la mécanique céleste, vers la maîtrise de l'Univers.
Un autre atout de la géométrie moderne est le calcul vectoriel, issu des quaternions découverts par Hamilton en 1843. Dès lors, la percée de l'algèbre linéaire pendant les cinquante ans qui suivirent changea profondément la donne. Aujourd'hui les axiomatiques de Hilbert, corrigeant Euclide, sont « marginalisées » par l'inclusion de la géométrie dans cette nouvelle et puissante discipline. L'enseignement en fut, un temps, fortement bouleversé.
Pourtant, loin d'être morte, la géométrie est toujours secrètement présente. Une vision géométrique sous-tend la théorie des nombres et une partie de l'analyse. Elle féconde les travaux les plus actuels, comme par exemple la recherche de symétries cachées des mathématiques ou de la physique théorique. La notion de point reprend sa souveraineté, même si les représentations intuitives sous-jacentes diffèrent largement des figures euclidiennes de base.

Dans ce cours de première année nous avons croisé des approches géométriques traditionnelles et analytiques d'objets simples (droites, plans, coniques...), renouvelées toutes deux par l'algèbre linéaire. L'étudiant devra d'abord apprendre à les manipuler par le calcul, en utilisant des bases, puis à s'abstraire du choix de celles-ci, ou à exploiter au mieux ce choix, ce qui est l'essence des méthodes vraiment géométriques.

Géométrie dans les espaces affines

III.1

Comme signalé dans le module II.3, une des sources de l'algèbre linéaire est la géométrie d'Euclide. Dans ce module, nous allons illustrer cela sur quelques exemples.
En algèbre linéaire, nous avons manipulé des vecteurs. Nous les avons ajoutés. Nous les avons multipliés par des scalaires. Nous les avons transformés à l'aide d'applications linéaires. Nous les avons repérés à l'aide de leurs coordonnées dans des bases.
Dans ce module, les acteurs seront des points, par exemple, les points du plan. Il n'est pas sensé d'ajouter des points, pas plus qu'il n'est sensé d'ajouter les positions de deux points A et B sur une feuille de papier. En revanche, on peut parler de la distance qui sépare A et B. On peut dire que A se trouve 10 cm au dessus et 3 cm à droite de B. Cela va se traduire par le fait que dans un espace affine, on peut passer d'un point à un autre en ajoutant un vecteur.
La notion d'espace affine apparaît dans le module II.7 dans le cadre d'un sous-ensemble d'un espace vectoriel pour étudier les solutions d'un système linéaire. Ici, nous donnerons une définition plus générale d'espace affine. Nous introduirons la notion d'application affine (analogue de celle d'application linéaire pour les espaces vectoriels) et la notion de repère (analogue de la notion de base pour les espaces vectoriels).
Muni de ces notions, nous étudierons des sous-ensembles particuliers. Nous commencerons par les sous-espaces affines, en particulier, les droites, les plans et les hyperplans. Nous verrons que nous pouvons définir ces ensembles à l'aide d'équations cartésiennes ou d'équations paramétriques.
Puis, nous nous initierons à l'étude des courbes planes définies par des équations. Nous nous intéresserons à l'étude des coniques, sous-ensembles du plan définis par des équations de degré 2.

1 Espaces affines

Un espace affine est un ensemble de points muni d'une structure proche de celle d'espace vectoriel.

1.1 Structure d'espace affine

Définition 1. Un K*-espace affine* est la donnée d'un espace de points X non vide, d'un K-espace vectoriel E et d'une application :

$$\begin{array}{ccc} X \times X & \longrightarrow & E \\ (A, B) & \longmapsto & \overrightarrow{AB} \end{array}$$

qui vérifie les deux propriétés suivantes :

1. pour tous points A, B, C dans X, $\overrightarrow{AB} + \overrightarrow{BC} = \overrightarrow{AC}$ (relation de Chasles)
2. pour tout point $A \in X$ et tout vecteur $\vec{u} \in E$, il existe un unique point $B \in X$ tel que $\overrightarrow{AB} = \vec{u}$.

On note $B := A + \vec{u}$ l'unique point $B \in X$ tel que $\overrightarrow{AB} = \vec{u}$.

Nous dirons que X est un espace affine et que E est la direction de l'espace affine X. Dans toute la suite, X désigne un espace affine et E le K-espace vectoriel qui dirige X. Les propriétés suivantes découlent directement de la définition d'espace affine.

Proposition 1.

1. $\overrightarrow{AB} = \overrightarrow{0}$ si, et seulement si, $A = B$;
2. $\overrightarrow{AB} = -\overrightarrow{BA}$;
3. $\overrightarrow{AB} = \overrightarrow{DC}$ si, et seulement si, $\overrightarrow{AD} = \overrightarrow{BC}$ (identité du parallélogramme).

Démonstration.

1. D'après la relation de Chasles, pour tout A et tout B dans X, on a $\overrightarrow{AA} + \overrightarrow{AB} = \overrightarrow{AB}$. Donc $\overrightarrow{AA} = \overrightarrow{0}$. Réciproquement, si $\overrightarrow{AB} = \overrightarrow{0} = \overrightarrow{AA}$, l'unicité dans le point 2 garantit que $B := A + \overrightarrow{AB} = A + \overrightarrow{AA} = A$.

2. D'après la relation de Chasles et ce qui précède, $\overrightarrow{AB} + \overrightarrow{BA} = \overrightarrow{AA} = \overrightarrow{0}$.

   D'où $\overrightarrow{AB} = -\overrightarrow{BA}$.

3. Si $\overrightarrow{AB} = \overrightarrow{DC}$, alors $\overrightarrow{AD} = \overrightarrow{AB} + \overrightarrow{BD} = \overrightarrow{DC} + \overrightarrow{BD} = \overrightarrow{BC}$.
   On a la réciproque en échangeant le rôle des points. ∎

FIGURE III.1.1 – Identité du parallélogramme

Exemples.

1. L'ensemble X des points du plan de coordonnées $(x, y) \in \mathbb{R}^2$ est un espace affine dirigé par l'espace vectoriel $E := \mathbb{R}^2$. L'application $X \times X \to E$ est l'application :
$$\big((x_1, y_1), (x_2, y_2)\big) \mapsto (x_2 - x_1, y_2 - y_1).$$
2. L'ensemble X des points de l'espace de coordonnées $(x, y, z) \in \mathbb{R}^3$ est un espace affine dirigé par l'espace vectoriel $E := \mathbb{R}^3$. L'application $X \times X \to E$ est l'application :
$$\big((x_1, y_1, z_1), (x_2, y_2, z_2)\big) \mapsto (x_2 - x_1, y_2 - y_1, z_2 - z_1).$$
3. Plus généralement, si E est un K-espace vectoriel, alors, l'application :
$$\begin{array}{ccc} E \times E & \longrightarrow & E \\ (x, y) & \longmapsto & y - x \end{array}$$
muni E d'une structure de K-espace affine. Par exemple : K^n. On dit alors que E est muni de sa structure affine canonique.

Exercice 1.
Montrer que l'ensemble X des fonctions $f : \mathbb{R} \to \mathbb{R}$ telles que $f(0) = 1$ est un espace affine dirigé par l'espace vectoriel E des fonctions $f : \mathbb{R} \to \mathbb{R}$ qui s'annulent en 0.
Solution. L'application $X \times X \to E$ est l'application $(f, g) \mapsto g - f$.

Définition 2. La *dimension* (finie ou infinie) d'un espace affine X dirigé par l'espace vectoriel E est la dimension de E.

Exemple. Ainsi, $\mathbb{R}^2$ est un espace affine de dimension 2, $\mathbb{R}^3$ un espace affine de dimension 3, et l'espace affine des fonctions $f : \mathbb{R} \to \mathbb{R}$ telles que $f(0) = 1$ est de dimension infinie.

Définition 3. On appelle *droite affine* tout espace affine de dimension 1 et *plan affine* tout espace affine de dimension 2.

1.2 Barycentres

Les barycentres sont aux espaces affines ce que les combinaisons linéaires sont aux espaces vectoriels.

Théorème 2. Soit $(A_i)_{i\in I}$ une famille de points du K-espace affine X indexée par l'ensemble I. Soit $(\lambda_i)_{i\in I}$ une famille de scalaires (éléments de K) à support fini telle que $\sigma := \sum_{i\in I} \lambda_i \neq 0$. Alors il existe un unique point $G \in X$ tel que :

$$\sum_{i\in I} \lambda_i \overrightarrow{GA_i} = 0.$$

Pour tout point $O \in X$, on a :

$$G = O + \frac{1}{\sigma} \sum_{i\in I} \lambda_i \overrightarrow{OA_i}.$$

Démonstration. Soit $O \in X$ un point quelconque. Alors, pour tout $M \in X$, on peut écrire :

$$\sum_{i\in I} \lambda_i \overrightarrow{MA_i} = \sum_{i\in I} \lambda_i \left(\overrightarrow{OA_i} - \overrightarrow{OM}\right) = \left(\sum_{i\in I} \lambda_i \overrightarrow{OA_i}\right) - \sigma\overrightarrow{OM}.$$

Il existe un unique point $G \in X$ tel que :

$$\left(\sum_{i\in I} \lambda_i \overrightarrow{OA_i}\right) - \sigma\overrightarrow{OG} = 0.$$

C'est le point :

$$G := O + \overrightarrow{OG} = O + \frac{1}{\sigma} \sum_{i\in I} \lambda_i \overrightarrow{OA_i}.$$

■

Définition 4. Sous les hypothèses du théorème 2, on appelle *barycentre* de la famille $(A_i, \lambda_i)_{i\in I}$ l'unique point G tel que $\sum_{i\in I} \lambda_i \overrightarrow{GA_i} = 0$.

Exemple. Considérons deux points A et B du plan affine $\mathbb{R}^2$. Le barycentre de la famille $\{(A, 1), (B, 1)\}$ est le milieu du segment $[A, B]$. Plus généralement, si $\lambda_1 + \lambda_2 \neq 0$ et si G est le barycentre de la famille $\{(A, \lambda_1), (B, \lambda_2)\}$ alors, les points A, B et G sont alignés. En effet, $\lambda_1\overrightarrow{GA} + \lambda_2\overrightarrow{GB} = 0$ et donc, les vecteurs $\overrightarrow{GA}$ et $\overrightarrow{GB}$ sont colinéaires.

Nous allons maintenant montrer une propriété d'associativité du barycentre.

Théorème 3. Soit $(A_i)_{i\in I}$ une famille de points de l'espace affine X indexée par l'ensemble I. Soit $(\lambda_i)_{i\in I}$ une famille de scalaires à support fini telle que $\sum_{i\in I}\lambda_i \neq 0$. Soit enfin $(I_j)_{j\in J}$ une partition de I telle que pour tout $j \in J$, $\sigma_j := \sum_{i\in I_j}\lambda_i \neq 0$. Alors, le barycentre de la famille $(A_i,\lambda_i)_{i\in I}$ est également le barycentre de la famille $(G_j,\sigma_j)_{j\in J}$ où G_j est le barycentre de la famille $(A_i,\lambda_i)_{i\in I_j}$.

Démonstration. Il suffit d'écrire :

$$\sum_{i\in I}\lambda_i\overrightarrow{GA_i} = \sum_{j\in J}\left(\sum_{i\in I_j}\lambda_i\overrightarrow{GA_i}\right) = \sum_{j\in J}\sigma_j\overrightarrow{GG_j}.$$

Exemple. Considérons un triangle ABC dans le plan affine $\mathbb{R}^2$. Le barycentre G de la famille $\{(A,1),(B,1),(C,1)\}$ est le *centre de gravité* du triangle. Il se trouve à l'intersection des médianes. En effet, G est le barycentre de la famille $\{(A,1),(A',2)\}$ avec A' le barycentre de la famille $\{(B,1),(C,1)\}$, c'est-à-dire le milieu du segment $[BC]$. Le point G appartient donc à la médiane (AA'). On montre de même qu'il appartient aux deux autres médianes.

FIGURE III.1.2 – Les médianes d'un triangle sont concourantes

1.3 Sous-espaces affines

Définition 5.
Un sous-ensemble non vide $Y \subset X$ est un *sous-espace affine* de X dirigé par F si :

1. F est un sous-espace vectoriel de E ;
2. pour tous points A et B dans Y, le vecteur $\overrightarrow{AB}$ appartient à F ;
3. pour tout point A dans Y et tout vecteur $\vec{u}$ dans F, le point $A+\vec{u}$ appartient à Y.

Exemple. Dans le plan affine $\mathbb{R}^2$, les points et les droites sont des sous-espaces affines.

Exercice 2.

Soit Y un sous-espace affine de X dirigé par F. Montrer que pour tout $A \in Y$, on a l'égalité $Y = A + F$.

Solution. D'après le point 2 de la définition 5, si $B \in Y$, $B = A + \overrightarrow{AB} \in A + F$. Donc $Y \subset A + F$. D'après le point 3 de la définition 5, on a $A + F \subset Y$

Proposition 4. Soient $(A_i)_{i\in I}$ une famille de points de l'espace affine X et Y l'ensemble des barycentres de familles $(A_i, \lambda_i)_{i\in I}$.
Alors Y est un sous-espace affine de X.

Démonstration. Soit $A := A_{i_0}$ un point quelconque de la famille $(A_i)_{i\in I}$. Soit $F \subset E$ le sous-espace vectoriel engendré par la famille $\left(\overrightarrow{AA_i}\right)_{i\in I}$. On va montrer que Y est un sous-espace affine dirigé par F.
Supposons que G est le barycentre de la famille $(A_i, \lambda_i)_{i\in I}$ et que G' est le barycentre de la famille $(A_i, \lambda'_i)_{i\in I}$. Alors :

$$\overrightarrow{AG} = \frac{1}{\sigma}\sum_{i\in I}\lambda_i\overrightarrow{AA_i} \in F \quad \text{et} \quad \overrightarrow{AG'} = \frac{1}{\sigma'}\sum_{i\in I}\lambda'_i\overrightarrow{AA_i} \in F.$$

Par conséquent, $\overrightarrow{GG'} = \overrightarrow{AG'} - \overrightarrow{AG} \in F$.
Supposons que G est le barycentre de la famille $(A_i, \lambda_i)_{i\in I}$ et que $\vec{u} \in F$. Alors, il existe une famille de scalaires $(\lambda'_i)_{i\in I}$, à support fini, telle que :

$$\vec{u} = \sum_{i\in I}\lambda'_i\overrightarrow{AA_i}.$$

On peut choisir n'importe quelle valeur pour λ'_{i_0}. On peut donc supposer que $\displaystyle\sum_{i\in I}\lambda'_i = 0$.
On a alors :

$$\overrightarrow{AG} + \vec{u} = \frac{1}{\sigma}\sum_{i\in I}(\lambda_i + \sigma\lambda'_i)\overrightarrow{AA_i} \quad \text{avec} \quad \sum_{i\in I}(\lambda_i + \sigma\lambda'_i) = \sum_{i\in I}\lambda_i = \sigma \neq 0.$$

Le point $G + \vec{u}$ est donc barycentre de la famille $(A_i, \lambda_i + \sigma\lambda'_i)_{i\in I}$ et appartient à Y. ■

Proposition 5 (Stabilité par barycentre). Soit Y un sous-espace affine de X et $(A_i)_{i\in I}$ une famille de points de Y. Si G est le barycentre d'une famille $(A_i, \lambda_i)_{i\in I}$, alors $G \in Y$.

Démonstration. Soit F le sous-espace vectoriel de E qui dirige Y. Soit $A := A_{i_0}$ un point quelconque de la famille A_i. Puisque Y est un sous-espace affine et puisque $A_i \in Y$, on a $\overrightarrow{AA_i} \in F$ pour tout $i \in I$. De plus, $G = A + \overrightarrow{AG}$ avec $\overrightarrow{AG} = \displaystyle\frac{1}{\sigma}\sum_{i\in I}\lambda_i\overrightarrow{AA_i} \in F$.
Donc, $G \in A + F = Y$. ■

Définition 6. Soient Y_1 et Y_2 deux sous-espaces affines de X dirigés par F_1 et F_2. Alors, Y_1 est *parallèle à* Y_2 si $F_1 \subset F_2$; Y_1 et Y_2 sont *parallèles* si $F_1 = F_2$.

Étant donnés un point A du plan affine et une droite D, il passe par A une unique droite parallèle à D. Le résultat suivant est plus général. On lui donne parfois le nom de cinquième postulat d'Euclide (bien que ce postulat ne soit pas énoncé en ces termes par Euclide).

Théorème 6. Dans un espace affine X dirigé par l'espace vectoriel E, étant donné un point quelconque A, et un sous-espace vectoriel $F \subset E$, il existe un unique sous-espace affine passant par A dirigé par F.

Démonstration. Ce sous-espace affine est $A + F$. ■

Proposition 7. Soient Y_1 et Y_2 deux sous-espaces affines de X. Si Y_1 est parallèle à Y_2, alors soit $Y_1 \cap Y_2 = \varnothing$, soit $Y_1 \subset Y_2$.

Démonstration. Soient F_1 et F_2 les espaces vectoriels qui dirigent Y_1 et Y_2. Comme Y_1 est parallèle à Y_2, on a $F_1 \subset F_2$. Supposons que $Y_1 \cap Y_2 \neq \varnothing$ et choisissons $A \in Y_1 \cap Y_2$. Alors $Y_1 = A + F_1 \subset A + F_2 = Y_2$. ■

Corollaire 8. Deux sous-espaces affines parallèles et distincts sont disjoints.

Théorème 9 (Intersection d'espaces affines).

1. Si Y_1 et Y_2 sont deux sous-espaces affines de X, dirigés respectivement par F_1 et F_2, alors :
   - soit $Y_1 \cap Y_2 = \varnothing$,
   - soit $Y_1 \cap Y_2$ est un sous-espace affine de X dirigé par $F_1 \cap F_2$.
2. Si de plus $F_1 + F_2 = E$, alors $Y_1 \cap Y_2 \neq \varnothing$. Si $F_1 \oplus F_2 = E$ (les espaces sont supplémentaires) alors $Y_1 \cap Y_2$ contient exactement 1 point.

Démonstration.

1. Supposons $Y_1 \cap Y_2$ non vide. Nous devons montrer que c'est un sous-espace affine de X dirigé par $F_1 \cap F_2$. Rappelons que l'intersection de deux sous-espaces vectoriels est un sous-espace vectoriel. Par conséquent, $F_1 \cap F_2$ est un sous-espace vectoriel de E.

   Si $A \in Y_1 \cap Y_2$ et $B \in Y_1 \cap Y_2$, alors $\overrightarrow{AB} \in F_1$ car A et B sont tous deux dans Y_1 qui est un sous-espace affine de X dirigé par F_1. De même, $\overrightarrow{AB} \in F_2$ car A et B sont tous deux dans Y_2. Par conséquent, $\overrightarrow{AB} \in F_1 \cap F_2$.

   Si $A \in Y_1 \cap Y_2$ et si $\vec{u} \in F_1 \cap F_2$, alors, $A + \vec{u} \in Y_1$ car $A \in Y_1$ et $\vec{u} \in F_1$. De même, $A + \vec{u} \in Y_2$ car $A \in Y_2$ et $\vec{u} \in F_2$. Donc, $A + \vec{u} \in Y_1 \cap Y_2$.

2. Choisissons deux points $A_1 \in Y_1$ et $A_2 \in Y_2$. Si $F_1 + F_2 = E$, alors $\overrightarrow{A_1A_2} = \vec{u}_1 + \vec{u}_2$ avec $\vec{u}_1 \in F_1$ et $\vec{u}_2 \in F_2$. Posons :

$$M := A_1 + \vec{u}_1 = A_2 - \vec{u}_2.$$

Alors $M \in A_1 + F_1 = Y_1$ et $M \in A_2 + F_2 = Y_2$. Donc $M \in Y_1 \cap Y_2$.
Inversement, si $M \in Y_1 \cap Y_2$, alors :

$$\overrightarrow{A_1A_2} = \underbrace{\overrightarrow{A_1M}}_{\in F_1} + \underbrace{\overrightarrow{MA_2}}_{\in F_2}.$$

Si $F_1 \oplus F_2 = E$, alors la décomposition est unique. Donc, si M et M' appartiennent à $Y_1 \cap Y_2$, on a $\overrightarrow{A_1M} = \overrightarrow{A_1M'}$ et $M = M'$. ■

1.4 Applications affines

Définition 7. Soient X et Y deux espaces affines dirigés respectivement par E et F. Une application $f : X \to Y$ est une *application affine* s'il existe une application linéaire $L : E \to F$ telle que pour tous points A et B dans X, on a :

$$\overrightarrow{f(A)f(B)} = L\big(\overrightarrow{AB}\big).$$

L'application linéaire L s'appelle la *partie linéaire* de f.

Exemples.

1. Si X et Y sont deux espaces affines, toute application constante $f : X \to Y$ est affine. Sa partie linéaire est l'application nulle.
2. Étant donnés un espace affine X dirigé par E et un vecteur $\vec{u} \in E$, la translation :

$$\begin{array}{rccc} f: & X & \longrightarrow & X \\ & M & \longmapsto & M + \vec{u} \end{array}$$

est une application affine. En effet, pour tout A et tout B dans X, on a :

$$\overrightarrow{f(A)f(B)} = \overrightarrow{f(A)A} + \overrightarrow{AB} + \overrightarrow{Bf(B)} = -\vec{u} + \overrightarrow{AB} + \vec{u} = \overrightarrow{AB}.$$

La partie linéaire de f est l'identité.

Notons que si $f : X \to Y$ est une application affine, si $L : E \to F$ est la partie linéaire de f et si $A \in X$, alors pour tout $M \in X$, on a :

$$f(M) = f(A) + L\big(\overrightarrow{AM}\big).$$

Réciproquement, s'il existe un point $A \in X$ et une application linéaire $L : E \to F$ telle que pour tout $M \in X$, on a $f(M) = f(A) + L\big(\overrightarrow{AM}\big)$ alors pour tout $M \in X$ et tout $N \in X$, on a :

$$\overrightarrow{f(M)f(N)} = \overrightarrow{f(A)f(N)} - \overrightarrow{f(A)f(M)} = L\big(\overrightarrow{AN}\big) - L\big(\overrightarrow{AM}\big) = L\big(\overrightarrow{MN}\big)$$

et f est une application affine.

Exemples.

1. Si K est un corps muni de sa structure affine canonique, et si $f : K \to K$ est une application affine, alors $f(x) = f(0) + L(x - 0)$ avec $L : K \to K$ une application

linéaire. Les applications linéaires $L : K \to K$ sont de la forme $x \mapsto a \cdot x$ avec $a \in K$. Par conséquent, les applications affines $f : K \to K$ sont les applications de la forme :

$$f : \begin{array}{ccc} K & \longrightarrow & K \\ x & \longmapsto & ax + b \end{array} \qquad \text{avec} \quad (a, b) \in K^2.$$

2. Supposons K^m et K^n munis de leurs structures affines canoniques, et considérons $f : K^m \to K^n$ une application affine de partie linéaire $L : K^m \to K^n$. Alors pour tout $M := (x_1, \ldots, x_m) \in K^m$, on a $f(M) = f(O) + L\big(\overrightarrow{OM}\big)$, avec $O := (0, \ldots, 0)$. Cela s'écrit sous forme matricielle :

$$f(M) = A \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

avec $(b_1, \ldots, b_n) := f(O) \in K^n$ et $A \in M_{m,n}(K)$ une matrice à m lignes et n colonnes. Autrement dit, les applications affines $f : K^m \to K^n$ sont les applications de la forme :

$$f : \begin{array}{ccc} K^m & \longrightarrow & K^n \\ x & \longmapsto & Ax + b \end{array} \qquad \text{avec} \quad A \in M_{m,n}(K) \quad \text{et} \quad b \in K^n.$$

Exercice 3.

On considère que le corps K est muni de sa structure affine canonique. Pour $(a, b) \in K^2$ on note $f_{a,b}$ l'application affine :

$$f_{a,b} : \begin{array}{ccc} K & \longrightarrow & K \\ x & \longmapsto & ax + b. \end{array}$$

Si $a = 1$, $f_{a,b}$ est une translation. Si $a \neq 1$ ou si $(a, b) = (1, 0)$, $f_{a,b}$ est une homothétie de rapport a. Montrer que l'ensemble des applications affines inversibles de K dans K est un groupe (pour la loi de composition). On l'appelle le groupe des homothéties-translations.

Solution. Nous ne donnerons que les grandes lignes de la preuve.

1. La composée de deux applications affines de K dans K est une application affine de K dans K.

2. L'application affine $f_{a,b}$ est inversible si, et seulement si, $f_{a,b}$ est une homothétie ou une translation. Son inverse est $f_{1/a,-b/a}$.

3. La composée de deux translations est une translation.

4. La composée de deux homothéties $f_{a,b}$ et $f_{c,d}$ est une homothétie si $ac \neq 1$ et une translation si $ac = 1$.

Théorème 10 (Image réciproque d'un sous-espace affine par une application affine). Soit $f : X \to Y$ une application affine, $L : E \to F$ sa partie linéaire et Y' un sous-espace affine de Y dirigé par $F' \subset F$. Posons :

$$X' := f^{-1}(Y') := \{M \in X \mid f(M) \in Y'\}.$$

Alors, soit $X' = \varnothing$, soit X' est un sous-espace affine de X dirigé par :

$$E' := L^{-1}(F') := \{\vec{u} \in E \mid L(\vec{u}) \in F'\}.$$

Démonstration. Supposons que X' ne soit pas vide et soit A un point de X'. Alors, $f(A) \in Y'$ et $Y' = f(A) + F'$. Pour tout $B \in X$:

$$\overrightarrow{f(A)f(B)} = L\big(\overrightarrow{AB}\big).$$

Donc :

$$X' = \{B \in X \mid f(B) \in f(A) + F'\} = \{B \in X \mid L\big(\overrightarrow{AB}\big) \in F'\} = A + E'.$$

Notons que E' est un sous-espace vectoriel de E. En effet, $\overrightarrow{0} \in E'$ car $L(\overrightarrow{0}) = \overrightarrow{0} \in F'$. De plus pour tout $\vec{u} \in E'$, tout $\vec{v} \in E'$ et tout $\lambda \in K$:

$$L(\lambda\vec{u} + \vec{v}) = \lambda L(\vec{u}) + L(\vec{v}) \in F' \quad \text{et donc} \quad \lambda\vec{u} + \vec{v} \in E'.$$

Par conséquent, $X' = A + E'$ est un sous-espace affine de X dirigé par E'. ■

1.4.1 Homothéties

Définition 8. Soit A un point de X et k un réel non nul. On appelle *homothétie* de centre A et de rapport k l'application affine :

$$\begin{array}{rccc} f : & X & \longrightarrow & X \\ & M & \longmapsto & A + k\overrightarrow{AM}. \end{array}$$

Une homothétie de rapport 1 est l'identité. Tous les points de X sont alors des points fixes, c'est-à-dire tels que $f(M) = M$.

Proposition 11. Si $f : X \to X$ est une homothétie de rapport $k \neq 1$, il y a un unique point fixe : le centre de l'homothétie.

Démonstration. On a $f(M) = M$ si, et seulement si, $\overrightarrow{AM} = k\overrightarrow{AM}$ donc si, et seulement si, $(1-k)\overrightarrow{AM} = \overrightarrow{0}$ donc si, et seulement si, $M = A$. ■

1.4.2 Formes affines

Définition 9. On appelle *forme affine* sur le K-espace affine X toute application affine de X dans K, muni de sa structure affine canonique.

La partie linéaire $L : E \to K$ d'une forme affine $f : X \to K$ est une forme linéaire sur E. Une telle forme linéaire est soit nulle soit surjective. Il s'en suit qu'une forme affine est soit constante, soit surjective.

1.4.3 Projections, symétries, affinités

Proposition 12. Soit $Y \subset X$ un sous-espace affine dirigé par $F \subset E$ et soit $G \subset E$ un supplémentaire de F dans E.

Pour tout point $M \in X$, il existe un unique point $M' \in Y$ tel que $\overrightarrow{MM'} \in G$.

Démonstration. D'après le théorème 9 de la page 431, le sous-espace affine $M + G$ coupe Y en un unique point M'. ∎

Définition 10. Le point M' est le *projeté* de M sur Y parallèlement à G. L'application qui à M associe M' s'appelle la *projection* sur Y parallèlement à G.

FIGURE III.1.3 – Projection sur le plan Y parallèlement à la droite vectorielle G.

Proposition 13. La projection sur Y parallèlement à G est une application affine dont la partie linéaire est la projection (vectorielle) sur F parallèlement à G.

Démonstration. Soit A un point de Y et soit $f : X \to Y$ la projection sur Y parallèlement à G. Par définition, on a $A = f(A)$ et :

$$f(M) = M' = A + \overrightarrow{AM'} \quad \text{avec} \quad \overrightarrow{AM} = \underbrace{\overrightarrow{AM'}}_{\in F} + \underbrace{\overrightarrow{M'M}}_{\in G}.$$

En particulier, $\overrightarrow{AM'} = L\big(\overrightarrow{AM}\big)$ avec $L : E \to E$ la projection vectorielle sur F parallèlement à G, et on a bien $f(M) = f(A) + L\big(\overrightarrow{AM}\big)$. ∎

Définition 11. Soit M' le *projeté* de M sur Y parallèlement à G. L'application qui à M associe $M + 2\overrightarrow{MM'} = M' - \overrightarrow{M'M}$ s'appelle la *symétrie* par rapport à Y parallèlement à G.

FIGURE III.1.4 – La symétrie par rapport au plan Y parallèlement à la droite vectorielle G.

Proposition 14. La symétrie par rapport à Y parallèlement à G est une application affine dont la partie linéaire est la symétrie (vectorielle) par rapport à F parallèlement à G.

Démonstration. Soit $L : E \to E$ la projection vectorielle sur F parallèlement à G et soit $L' : E \to E$ la symétrie vectorielle par rapport à F parallèlement à G.
On a $L' = 2L - \mathrm{Id}$. Soit $f : X \to X$ la projection affine sur Y parallèlement à G et soit $f' : X \to X$ la symétrie par rapport à Y parallèlement à G. Alors, pour tout A et tout B dans X, on a :

$$\begin{aligned}
\overrightarrow{f'(A)f'(B)} &= \overrightarrow{f'(A)A} + \overrightarrow{AB} + \overrightarrow{Bf'(B)} \\
&= 2\overrightarrow{f(A)A} + 2\overrightarrow{AB} - \overrightarrow{AB} + 2\overrightarrow{Bf(B)} \\
&= 2\overrightarrow{f(A)f(B)} - \overrightarrow{AB} \\
&= 2L\big(\overrightarrow{AB}\big) - \overrightarrow{AB} = L'\big(\overrightarrow{AB}\big).
\end{aligned}$$

■

Définition 12. Soit M' le *projeté* de M sur Y parallèlement à G. Si $k \in \mathbb{R}$, l'application qui à M associe $M' + k\overrightarrow{M'M}$ s'appelle l'*affinité* de rapport k par rapport à Y parallèlement à G.

Une projection est une affinité de rapport 0. Une symétrie est une affinité de rapport -1. Nous verrons à la fin de ce module qu'une ellipse est l'image d'un cercle par une affinité.

2 Représentation des sous-espaces affines

2.1 Hyperplans

La notion d'hyperplan d'un espace vectoriel a été introduite dans le module II.3 : c'est le noyau d'une forme linéaire non nulle.

Définition 13. On appelle *hyperplan affine* de X tout sous-espace affine dirigé par un hyperplan vectoriel de E.

Exemples.

1. Si X est une droite affine, les hyperplans affines sont les singletons. Ce sont les points de la droite.
2. Si X est un plan affine, les hyperplans sont les droites affines.
3. Dans l'espace affine $\mathbb{R}^3$, les hyperplans affines sont les plans de $\mathbb{R}^3$.

Exercice 4.

Étant donné un ensemble I, on munit le $\mathbb{R}$-espace vectoriel $\mathcal{F}(I,\mathbb{R})$ des fonctions numériques $f : I \to \mathbb{R}$ de sa structure affine canonique. Étant donné un point $a \in I$, soit $\mathcal{H}$ l'ensemble des fonctions $f : I \to \mathbb{R}$ telles que $f(a) = 1$.

1. Montrer que $\mathcal{H}$ est un sous-espace affine de $\mathcal{F}(I,\mathbb{R})$ et déterminer le sous-espace vectoriel $H \subset \mathcal{F}(I,\mathbb{R})$ qui dirige $\mathcal{H}$.
2. Montrer que l'application :
$$\begin{array}{rccc} \delta : & \mathcal{F}(I,\mathbb{R}) & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & f(a) \end{array}$$
est une forme linéaire.
3. Montrer que H est un hyperplan vectoriel de $\mathcal{F}(I,\mathbb{R})$ et en déduire que $\mathcal{H}$ est un hyperplan affine de $\mathcal{F}(I,\mathbb{R})$.

Solution.

1. Soit $H \subset \mathcal{F}(I,\mathbb{R})$ l'ensemble des fonctions $f : I \to \mathbb{R}$ qui s'annulent en a. La fonction identiquement nulle sur I appartient à H. Toute combinaison linéaire de fonctions qui s'annulent en a s'annule encore en a. Par conséquent, H est un sous-espace vectoriel de $\mathcal{F}(I,\mathbb{R})$.

   Montrons que $\mathcal{H}$ est un sous-espace affine de $\mathcal{F}(I,\mathbb{R})$ dirigé par H :
   - Si f_1 et f_2 sont deux fonctions de $\mathcal{H}$, le vecteur $f_2 - f_1$ s'annule en a (car $f_1(a) = f_2(a) = 1$). Par conséquent, $f_2 - f_1 \in H$.
   - Si $f_1 \in \mathcal{H}$ et si f_2 est un vecteur de H, alors $f_1(a) = 1$ et $f_2(a) = 0$. Par conséquent, $(f_1 + f_2)(a) = 1$ et $f_1 + f_2 \in \mathcal{H}$.
2. L'application δ est à valeur dans le corps $\mathbb{R}$. C'est une application linéaire puisque :
$$\delta(f_1 + \lambda f_2) = (f_1 + \lambda f_2)(a) = f_1(a) + \lambda f_2(a) = \delta(f_1) + \lambda\delta(f_2).$$
C'est donc une forme linéaire.
3. Remarquons que la forme linéaire δ est non identiquement nulle. Par conséquent, son noyau est un hyperplan de $\mathcal{F}(I,\mathbb{R})$. Or, $\mathrm{Ker}(\delta)$ est précisément l'ensemble des fonctions $f : I \to \mathbb{R}$ telles que $f(a) = 0$, c'est-à-dire l'espace H. On en déduit que H est un hyperplan vectoriel de $\mathcal{F}(I,\mathbb{R})$, et donc que $\mathcal{H}$ est un hyperplan affine de $\mathcal{F}(I,\mathbb{R})$.

Proposition 15. Soit $f : X \to K$ une forme affine non constante.
Alors, $\{M \in X \mid f(M) = 0\}$ est un hyperplan affine de X dirigé par le noyau de la partie linéaire de f.

Démonstration. Le singleton $\{0\}$ est un sous-espace affine de K muni de sa structure canonique d'espace affine. Il est dirigé par l'espace vectoriel trivial $\{0\}$. D'après le théorème 10 de la page 434, $Y := f^{-1}\{0\}$ est un sous-espace affine de X dirigé par le noyau de la partie linéaire L de f. Cette partie linéaire est une forme linéaire non nulle (car f est non constante). Par conséquent le noyau de L est un hyperplan vectoriel de E et Y est un hyperplan affine de X. ∎

La réciproque est vraie : tout hyperplan affine de X est donné par une équation.

Proposition 16. Soit Y un hyperplan affine d'un K-espace affine X. Alors, il existe une forme affine non constante $f : X \to K$ telle que $Y = \{M \in X \mid f(M) = 0\}$. Si f_1 et f_2 sont deux telles formes, alors $f_1 = \lambda f_2$ avec $\lambda \in K \setminus \{0\}$.

Remarque. On dit que $f(M) = 0$ est une équation de l'hyperplan Y.

Démonstration. L'hyperplan affine Y est dirigé par un hyperplan vectoriel H de E. On sait qu'il existe une forme linéaire non nulle $L : E \to K$ telle que $\mathrm{Ker}(L) = H$. Soit A un point de Y et définissons $f : X \to K$ par :

$$f(M) := 0 + L\big(\overrightarrow{AM}\big).$$

Alors, f est une application affine, donc une forme affine.

De plus, $f(M) = 0$ si, et seulement si, $L\big(\overrightarrow{AM}\big) = 0$ donc si, et seulement si, $\overrightarrow{AM} \in H$ donc si, et seulement si, $M = A + \overrightarrow{AM} \in Y$.

Supposons maintenant que f_1 et f_2 sont deux formes affines qui s'annulent exactement sur Y. Soit $B \in X \setminus Y$. Alors, $f_1(B) \neq 0$ et $f_2(B) \neq 0$ et pour tout $M \in X$:

$$\vec{u}_1 := \overrightarrow{AM} - \frac{f_1(M)}{f_1(B)}\overrightarrow{AB} \in H.$$

En effet, si L_1 est la partie linéaire de f_1, on a :

$$\begin{aligned} L_1\left(\overrightarrow{AM} - \frac{f_1(M)}{f_1(B)}\overrightarrow{AB}\right) &= L_1\big(\overrightarrow{AM}\big) - \frac{f_1(M)}{f_1(B)} L_1\big(\overrightarrow{AB}\big) \\ &= \big(f_1(M) - f_1(A)\big) - \frac{f_1(M)}{f_1(B)}\big(f_1(B) - f_1(A)\big) = 0 \end{aligned}$$

car $f_1(A) = 0$. De même $\vec{u}_2 := \overrightarrow{AM} - \dfrac{f_2(M)}{f_2(B)}\overrightarrow{AB} \in H$. En soustrayant les deux vecteurs, on obtient :

$$\vec{u}_2 - \vec{u}_1 = \left(\frac{f_1(M)}{f_1(B)} - \frac{f_2(M)}{f_2(B)}\right)\overrightarrow{AB} \in H.$$

Comme $\overrightarrow{AB} \notin H$, on voit que pour tout $M \in X$:

$$f_1(M) = \lambda \cdot f_2(M) \quad \text{avec} \quad \lambda := \frac{f_1(B)}{f_2(B)} \in K \setminus \{0\}.$$ ∎

2.2 Repère

À partir de maintenant, nous allons restreindre notre étude au cas des espaces affines de dimension finie. Nous illustrerons nos propos dans le cas du plan affine $\mathbb{R}^2$ et de l'espace affine $\mathbb{R}^3$.
Dans un espace affine, on ne parle pas de base mais de repère.

Définition 14. On appelle *repère affine* (ou cartésien) de l'espace affine X tout couple $(O;\mathcal{B})$ où O est un point de X appelé *origine* et $\mathcal{B}$ est une base de l'espace vectoriel E qui dirige X.

Si l'espace affine X est de dimension $n > 0$, un repère de X est alors constitué d'un point de X, l'origine, et de n vecteurs de E qui forment une base de E.

Définition 15. On appelle *coordonnées d'un point* $M \in X$ dans le repère $(O;\mathcal{B})$ les coordonnées du vecteur $\overrightarrow{OM}$ dans la base $\mathcal{B}$.

Exemple. Dans le plan affine muni du repère $(O;\vec{\imath},\vec{\jmath})$, le point A de coordonnées $(2,3)$ est le point $O + 2\vec{\imath} + 3\vec{\jmath}$.

FIGURE III.1.5 – Le point A de coordonnées $(2,3)$ dans le plan muni du repère $(O;\vec{\imath},\vec{\jmath})$.

Si X est muni d'un repère cartésien $(O;\vec{u}_1,\ldots,\vec{u}_n)$, et si $M \in X$ est un point de coordonnées $(x_1,\ldots,x_n)$, alors :

$$M = O + x_1\vec{u}_1 + x_2\vec{u}_2 + \ldots + x_n\vec{u}_n.$$

Une forme affine $f : X \to K$ de partie linéaire $L : E \to K$ s'écrit alors en coordonnées :

$$f(x_1,\ldots,x_n) = a_1x_1 + \ldots + a_nx_n + b$$

avec :

$$a_1 := L(\vec{u}_1), \quad \ldots, \quad a_n := L(\vec{u}_n) \quad \text{et} \quad b := f(O).$$

Toute équation d'un hyperplan affine de X est donc de la forme :

$$a_1x_1 + \ldots + a_nx_n + b = 0 \quad \text{avec} \quad (a_1,\ldots,a_n) \neq (0,\ldots,0).$$

La condition $(a_1,\ldots,a_n) \neq (0,\ldots,0)$ garantit que la forme linéaire est non constante. D'après la proposition 16 de la page ci-contre, toutes les équations d'un même hyperplan affine sont proportionnelles. Autrement dit, si :

$$a_1x_1 + \ldots + a_nx_n + b = 0 \quad \text{et} \quad a'_1x_1 + \ldots + a'_nx_n + b' = 0$$

sont deux équations d'un même hyperplan, alors il existe $\lambda \in K \setminus \{0\}$ tel que :

$$a'_1 = \lambda a_1, \quad \ldots, \quad a'_n = \lambda a_n \quad \text{et} \quad b' = \lambda b.$$

2.3 Systèmes d'équations

Nous avons vu qu'un sous-espace de X défini par une équation affine est un hyperplan affine. Si X est de dimension n, cet hyperplan est de dimension $n-1$. Ce n'est donc généralement pas une droite (l'unique exception est le cas où la dimension de X est 2, comme dans le cas du plan affine $\mathbb{R}^2$). Par exemple, dans l'espace affine $\mathbb{R}^3$, l'ensemble défini par une équation affine est un plan.

Nous allons voir maintenant comment décrire des sous-espaces affines de X qui ne sont pas des hyperplans à l'aide de systèmes d'équations : nous allons décrire les sous-espaces affines de X comme intersections d'hyperplans. D'après le théorème 9 de la page 431, une intersection d'hyperplans affines de X est un sous-espace affine de X.

En dimension finie, il y a un lien très fort avec les systèmes linéaires (qui ont été introduits dans le modules II.4 et II.7). Comme un hyperplan peut être décrit par une équation, l'intersection de p hyperplans $Y_1 \cap \cdots \cap Y_p$ peut être décrite comme le lieu où p formes affines $f_1 : X \to K, \ldots, f_p : X \to K$ s'annulent simultanément. En d'autre termes, l'intersection de p hyperplans affines peut être décrite par :

$$Y_1 \cap \cdots \cap Y_p = \{M \in X \mid f(M) = 0\} \quad \text{avec} \quad f := (f_1, \ldots, f_p) : X \to K^p.$$

Si $L_1 : E \to K, \ldots, L_p : E \to K$ sont les parties linéaires des formes affines $f_1, \ldots, f_p$, alors la partie linéaire de f est l'application :

$$L := (L_1, \ldots, L_p) : E \to K^p.$$

Le sous-espace affine $Y_1 \cap \cdots \cap Y_p$ est alors dirigé par $\mathrm{Ker}(L)$. D'après le théorème du rang, la dimension de cette intersection est égale à :

$$\dim(Y_1 \cap \cdots \cap Y_p) = \dim(E) - \mathrm{rang}(L) \quad \text{avec} \quad \mathrm{rang}(L) \leqslant p.$$

Donc, dans un espace affine X de dimension n, pour décrire une droite à l'aide d'un système d'équations, il faut au moins $n-1$ équations. Pour décrire un plan à l'aide d'un système d'équations, il faut au moins $n-2$ équations.

Nous allons maintenant nous intéresser à l'affirmation réciproque, dans le cas où X est un espace affine de dimension finie.

Proposition 17. Si X est de dimension finie, tout sous-espace affine Y de X est une intersection finie d'hyperplans affine.
Si la dimension de X est n et la dimension de Y est $n-p$, alors Y peut être obtenu comme intersection de p hyperplans affines de X.

Démonstration. Soit Y un sous-espace affine de X de dimension $n-p$ dirigé par F. Soit $(\vec{u}_1, \ldots, \vec{u}_{n-p})$ une base de F. D'après le théorème de la base incomplète, il existe des vecteurs $\vec{v}_1, \ldots, \vec{v}_p$ dans E tels que :

$$(\vec{u}_1, \ldots, \vec{u}_{n-p}, \vec{v}_1, \ldots, \vec{v}_p)$$

est une base de E.

Pour $i \in [\![1, p]\!]$, on note L_i la forme linéaire $L_i : E \to K$ qui prend la valeur 1 sur le vecteur $\vec{v}_i$ et s'annule sur tous les autres vecteurs de la base. Puis on choisit un point $A \in Y$ et pour tout $i \in [\![1, p]\!]$, on définit la forme affine $f_i : Y \to K$ par :

$$f_i(M) := L_i\big(\overrightarrow{AM}\big).$$

Montrons que :

$$Y = \big\{M \in X \;\big|\; \big(\forall i \in [\![1,p]\!]\big)\; f_i(M) = 0\big\}.$$

Pour cela, observons d'abord que les formes linéaires L_i s'annulent toutes sur F. Si $M \in Y$, alors $\overrightarrow{AM} \in F$. Donc, pour tout $i \in [\![1,p]\!]$:

$$f_i(M) = L_i\big(\overrightarrow{AM}\big) = 0.$$

De plus, si $M \notin Y$, alors le vecteur $\overrightarrow{AM}$ n'est pas contenu dans F et a donc une coordonnée non nulle selon au moins l'un des vecteurs $\vec{v}_i$. Mais cette coordonnée est précisément égale à $f_i(M) = L_i\big(\overrightarrow{AM}\big) \neq 0$. ∎

Corollaire 18. Toute droite affine d'un espace affine X de dimension n peut être caractérisée par un système de $n-1$ équations affines. Tout plan affine d'un espace affine X de dimension n peut être caractérisé par un système de $n-2$ équations affines.

Remarque. Nous voyons donc que si dans le plan affine $\mathbb{R}^2$, il suffit d'une équation affine pour caractériser une droite, dans l'espace affine $\mathbb{R}^3$, il faut deux équations.

FIGURE III.1.6 – Dans $\mathbb{R}^3$, une droite affine D peut être décrite comme l'intersection de deux plans affines P_1 et P_2.

3 Géométrie affine dans $\mathbb{R}^2$ et dans $\mathbb{R}^3$

Dans cette partie, nous allons restreindre notre étude au cas d'espaces affines de dimension 2 et 3. Si X est un espace affine de dimension 2, on peut le munir d'un repère $(O;\vec{\imath},\vec{\jmath})$. L'application qui à un point M associe ses coordonnées $(x,y) \in \mathbb{R}^2$ est une application affine et bijective. Elle nous permet d'identifier l'espace affine X muni du repère $(O;\vec{\imath},\vec{\jmath})$ avec $\mathbb{R}^2$ muni du repère $\big((0,0);(1,0),(0,1)\big)$. Avec un abus de langage, nous identifierons donc le point $M \in X$ et le couple $(x,y) \in \mathbb{R}^2$ dés lors que :

$$M = O + x\vec{\imath} + y\vec{\jmath}.$$

De manière similaire, si X est un espace affine de dimension 3 muni d'un repère $(O;\vec{\imath},\vec{\jmath},\vec{k})$, nous identifierons le point $M \in X$ et le point $(x,y,z) \in \mathbb{R}^3$ dés lors que :

$$M = O + x\vec{\imath} + y\vec{\jmath} + z\vec{k}.$$

3.1 Droites de $\mathbb{R}^2$

3.1.1 Équations cartésiennes d'une droite

Les hyperplans affines de $\mathbb{R}^2$ sont les droites de $\mathbb{R}^2$. Une forme affine $f : \mathbb{R}^2 \to \mathbb{R}$ s'écrit $f : (x, y) \mapsto ax + by + c$ avec a, b et c réels. La forme est non constante si, et seulement si, $(a, b) \neq (0, 0)$. Une équation de droite est donc de la forme $ax + by + c = 0$ avec $(a, b) \neq (0, 0)$. Autrement dit, si D est une droite de $\mathbb{R}^2$, il existe des réels (a, b, c) avec $(a, b) \neq (0, 0)$ tels que :

$$D = \left\{(x, y) \in \mathbb{R}^2 \;\middle|\; ax + by + c = 0\right\}.$$

Nous allons maintenant expliquer comment trouver une telle équation.
Tout d'abord, si D est une droite passant par deux points distincts $A := (x_0, y_0)$ et $B := (x_1, y_1)$, on peut rechercher une équation de D en étudiant un système. En effet, si l'on cherche une équation de la forme $ax + by + c = 0$, alors, on a nécessairement :

$$\begin{cases} ax_0 + by_0 + c = 0 \\ ax_1 + by_1 + c = 0. \end{cases}$$

C'est un système homogène de deux équations avec trois inconnues (a, b, c). On peut écrire ce système sous forme matricielle :

$$\begin{pmatrix} x_0 & y_0 & 1 \\ x_1 & y_1 & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Comme les points A et B sont distincts, le rang du système est égal à 2. L'ensemble des solutions (a, b, c) est donc une droite vectorielle. Chaque solution correspond à une équation de la droite D. Toutes les autres équations lui sont proportionnelles.
Il est possible de court-circuiter la résolution du système.

Proposition 19. Une équation cartésienne de la droite D de $\mathbb{R}^2$ passant par les points distincts $A := (x_0, y_0)$ et $B := (x_1, y_1)$ est :

$$\det \begin{pmatrix} x & x_0 & x_1 \\ y & y_0 & y_1 \\ 1 & 1 & 1 \end{pmatrix} = 0.$$

Démonstration. Les point $M := (x, y)$ appartient à la droite D si, et seulement si, le système suivant de trois équations à trois inconnues (a, b, c) a une solution non triviale (en effet, une telle solution non triviale ne peut pas être de la forme $(0, 0, c)$ avec $c \neq 0$) :

$$\begin{cases} ax + by + c = 0 \\ ax_0 + by_0 + c = 0 \\ ax_1 + by_1 + c = 0. \end{cases}$$

C'est le cas si, et seulement si, le déterminant du système est nul. ∎

Exercice 5.
Donner une équation de la droite passant par les points $A := (1, 1)$ et $B := (3, -2)$.

Solution. Une équation de la droite est :

$$\det \begin{pmatrix} x & 1 & 3 \\ y & 1 & -2 \\ 1 & 1 & 1 \end{pmatrix} = 0,$$

c'est-à-dire $3x + 2y - 5 = 0$.

Proposition 20. Une équation cartésienne de la droite $D \subset \mathbb{R}^2$ passant par le point $A := (x_0, y_0)$ et dirigée par le vecteur $\vec{u} := (a, b)$ est :

$$\det \begin{pmatrix} x - x_0 & a \\ y - y_0 & b \end{pmatrix} = 0.$$

Démonstration. Le point $M := (x, y)$ appartient à la droite D si, et seulement si, les vecteurs $\overrightarrow{AM} = (x - x_0, y - y_0)$ et $\vec{u} = (a, b)$ sont colinéaires. ∎

Exercice 6.
Donner une équation de la droite passant par les points $A := (1, 1)$ et $B := (3, -2)$.

Solution. La droite est dirigée par le vecteur $\overrightarrow{AB} := (2, -3)$. Une équation est :

$$\det \begin{pmatrix} x - 1 & 2 \\ y - 1 & -3 \end{pmatrix} = 0,$$

c'est-à-dire $-3x - 2y + 5 = 0$.

Le résultat suivant est en quelque sorte une réciproque.

Proposition 21.
La droite $D \subset \mathbb{R}^2$ d'équation $ax + by + c = 0$ est dirigée par le vecteur $(-b, a)$.

Démonstration. L'espace vectoriel qui dirige D est le noyau de la forme linéaire $L : (x, y) \mapsto ax + by$ et $L(-b, a) = 0$. ∎

Il est également possible de caractériser les droites de $\mathbb{R}^2$ à l'aide d'*équations paramétriques*; si la droite D passe par le point $A := (x_0, y_0)$ et est dirigée par le vecteur $\vec{u} := (a, b)$, alors :

$$\begin{aligned} D &= \left\{ M \in \mathbb{R}^2 \;\middle|\; \overrightarrow{AM} = \lambda \vec{u} \text{ avec } \lambda \in \mathbb{R} \right\} \\ &= \left\{ (x_0 + \lambda a, y_0 + \lambda b) \;\middle|\; \lambda \in \mathbb{R} \right\}. \end{aligned}$$

On dit que :

$$\begin{cases} x = x_0 + \lambda a \\ y = y_0 + \lambda b \end{cases} \qquad \lambda \in \mathbb{R}$$

est une équation paramétrique de la droite D.

Réciproquement, si l'on connaît une équation paramétrique d'une droite D de $\mathbb{R}^2$, on détermine facilement un point de la droite et un vecteur directeur de la droite. La proposition 20 de la page précédente permet alors d'en trouver une équation cartésienne.

Exercice 7.
Donner une équation cartésienne de la droite d'équation paramétrique :

$$\begin{cases} x = 2 + 3\lambda \\ y = 1 - 4\lambda \end{cases} \qquad \lambda \in \mathbb{R}.$$

Solution. La droite passe par le point $(2, 1)$ et est dirigée par le vecteur $(3, -4)$. Elle a donc pour équation cartésienne :

$$\det \begin{pmatrix} x-2 & 3 \\ y-1 & -4 \end{pmatrix} = 0,$$

c'est-à-dire $-4x - 3y + 11 = 0$.

3.1.2 Intersection de deux droites

Proposition 22. Considérons deux droites D_1 et D_2 d'équations respectives $a_1x + b_1y + c_1 = 0$ et $a_2x + b_2y + c_2 = 0$.

1. Les droites D_1 et D_2 sont parallèles si, et seulement s'il existe $\lambda \in \mathbb{R}$ tel que $(a_2, b_2) = \lambda(a_1, b_1)$ donc si, et seulement si :

$$\det \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} = 0.$$

2. Si les droites D_1 et D_2 ne sont pas parallèles, elles se coupent en un unique point.

Démonstration. Les droites D_1 et D_2 sont parallèles si, et seulement si, les sous–espaces vectoriels qui les dirigent sont égaux. Ces sous-espaces vectoriels sont les noyaux respectifs des formes linéaires non nulles $L_1 : (x, y) \mapsto a_1x + b_1y$ et $L_2 : (x, y) \mapsto a_2x + b_2y$. Ces noyaux sont égaux si, et seulement si, $L_2 = \lambda L_1$ avec $\lambda \in \mathbb{R} \setminus \{0\}$.

Les points qui appartiennent à D_1 et D_2 sont les points (x, y) avec :

$$\begin{cases} a_1x + b_1y + c_1 = 0 \\ a_2x + b_2y + c_2 = 0. \end{cases}$$

Si les droites ne sont pas parallèles, alors le rang du système est égal à 2. Dans ce cas, d'après le théorème 64 de la page 377, le système admet une unique solution. ■

Exercice 8.
Soit D la droite d'équation $ax + by + c = 0$.

1. Montrer que le vecteur (a, b) ne dirige pas D.
2. Montrer que la droite D' passant par $(0, 0)$ et dirigée par (a, b) coupe D en un unique point A.
3. Déterminer les coordonnées de A.

Solution.

1. Les vecteurs qui dirigent la droite D sont dans le noyau de l'application linéaire $L : (x, y) \mapsto ax + by$, et $L(a, b) = a^2 + b^2 > 0$.
2. Les droites D' et D ne sont pas parallèles. Elles se coupent donc en un unique point.
3. $A = (ta, tb)$ avec $t = -c/(a^2 + b^2)$.

La proposition suivante est une conséquence de cet exercice et de la proposition 21 de la page 443.

Proposition 23. La droite $D \subset \mathbb{R}^2$ d'équation $ax+by+c = 0$, avec $(a, b) \neq (0, 0)$, passe par le point $\left(-\dfrac{ac}{a^2+b^2}, -\dfrac{bc}{a^2+b^2}\right)$ et est dirigé par le vecteur $(-b, a)$.

Exemple. La droite d'équation $2x + 3y - 4 = 0$ passe par le point $(8/13, 12/13)$ et est dirigée par le vecteur $(-3, 2)$.

3.2 Plans de $\mathbb{R}^3$

3.2.1 Équations cartésiennes d'un plan

Nous pouvons mener une étude similaire des hyperplans de $\mathbb{R}^3$; ce sont les plans de $\mathbb{R}^3$. Une équation d'un tel plan est de la forme $ax + by + cz + d = 0$. Autrement dit, si P est un plan de $\mathbb{R}^3$, il existe des réels (a, b, c, d) avec $(a, b, c) \neq (0, 0, 0)$ tels que :

$$P = \left\{(x, y, z) \in \mathbb{R}^3 \;\middle|\; ax + by + cz + d = 0\right\}.$$

Comme dans le cas des droites de $\mathbb{R}^2$, si l'on connaît trois points non alignés dans un plan $P \subset \mathbb{R}^3$, on peut retrouver une équation cartésienne de P. On peut pour cela résoudre un système de trois équations (une par point) à quatre inconnues (a, b, c, d). La méthode est quelque peu laborieuse. Comme pour les équations cartésiennes de droites dans $\mathbb{R}^2$, on dispose de la méthode plus rapide suivante.

Proposition 24. Une équation cartésienne du plan P contenant les trois points non alignés $A := (x_0, y_0, z_0)$, $B := (x_1, y_1, z_1)$ et $C := (x_2, y_2, z_2)$ est :

$$\det \begin{pmatrix} x & x_0 & x_1 & x_2 \\ y & y_0 & y_1 & y_2 \\ z & z_0 & z_1 & z_2 \\ 1 & 1 & 1 & 1 \end{pmatrix} = 0.$$

Démonstration. Le point $M := (x, y, z)$ appartient au plan P si, et seulement si, le système suivant de quatre équations à quatre inconnues (a, b, c, d) admet une solution non triviale (qui est alors nécessairement telle que $(a, b, c) \neq (0, 0, 0)$) :

$$\begin{cases} ax + by + cz + d = 0 \\ ax_0 + by_0 + cz_0 + d = 0 \\ ax_1 + by_1 + cz_1 + d = 0 \\ ax_2 + by_2 + cz_1 + d = 0. \end{cases}$$

C'est le cas si, et seulement si, le déterminant du système est nul. ∎

Exercice 9.
Donner une équation cartésienne du plan contenant les points $A := (1,2,3)$, $B := (0,1,0)$ et $C := (0,0,1)$.
Solution. Une équation cartésienne du plan est :

$$\det \begin{pmatrix} x & 1 & 0 & 0 \\ y & 2 & 1 & 0 \\ z & 3 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

c'est-à-dire $-4x + y + z - 1 = 0$.

Nous allons maintenant étudier le cas où le plan P est donné par un point $A := (x_0, y_0, z_0)$ et deux vecteurs non colinéaires $\vec{u}_1 := (a_1, b_1, c_1)$ et $\vec{u}_2 := (a_2, b_2, c_2)$ qui le dirigent.
Rappelons que si $\vec{u}$, $\vec{v}$ et $\vec{w}$ sont trois vecteurs de $\mathbb{R}^3$ alors, le produit mixte de $\vec{u}$, $\vec{v}$ et $\vec{w}$ est défini par :

$$[\vec{u}, \vec{v}, \vec{w}] := \det_{\mathcal{B}}(\vec{u}, \vec{v}, \vec{w})$$

avec $\mathcal{B} := (\vec{\imath}, \vec{\jmath}, \vec{k})$ la base canonique de $\mathbb{R}^3$. Le produit mixte est nul si, et seulement si, les trois vecteurs sont coplanaires. Par conséquent, si on connaît une base de l'espace vectoriel qui dirige P (par exemple $\vec{u} := \overrightarrow{AB}$ et $\vec{v} := \overrightarrow{AC}$), un point M appartient au plan P si, et seulement si, $\overrightarrow{AM}$, $\vec{u}$ et $\vec{v}$ sont coplanaires donc si, et seulement si, $\left[\vec{u}, \vec{v}, \overrightarrow{AM}\right] = 0$.
Le produit vectoriel de deux vecteurs $\vec{u} := (a_1, b_1, c_1)$ et $\vec{v} := (a_2, b_2, c_2)$ est défini par :

$$\vec{u} \wedge \vec{v} := \begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix} \wedge \begin{pmatrix} a_2 \\ b_2 \\ c_2 \end{pmatrix} := \begin{pmatrix} b_1c_2 - c_1b_2 \\ c_1a_2 - a_1c_2 \\ a_1b_2 - b_1a_2 \end{pmatrix}.$$

Si de plus $\vec{w} := (a_3, b_3, c_3)$, en développant le déterminant $\det_{\mathcal{B}}(\vec{u}, \vec{v}, \vec{w})$ par rapport à la troisième colonne, on obtient :

$$[\vec{u}, \vec{v}, \vec{w}] = aa_3 + bb_3 + cc_3 \quad \text{avec} \quad (a, b, c) = \vec{u} \wedge \vec{v}.$$

On en déduit le résultat suivant.

Proposition 25. Soit P un plan de $\mathbb{R}^3$ contenant le point (x_0, y_0, z_0) et dirigé par le plan vectoriel engendré par (a_1, b_1, c_1) et (a_2, b_2, c_2). Une équation cartésienne de P est :

$$\det \begin{pmatrix} a_1 & a_2 & x - x_0 \\ b_1 & b_2 & y - y_0 \\ c_1 & c_2 & z - z_0 \end{pmatrix} = 0,$$

c'est-à-dire $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$ avec :

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} := \begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix} \wedge \begin{pmatrix} a_2 \\ b_2 \\ c_2 \end{pmatrix} x = \begin{pmatrix} b_1c_2 - c_1b_2 \\ c_1a_2 - a_1c_2 \\ a_1b_2 - b_1a_2 \end{pmatrix}.$$

Exercice 10.

Donner une équation cartésienne du plan contenant les points $A := (1,2,3)$, $B := (0,1,0)$ et $C := (0,0,1)$.

Solution. Comme $\overrightarrow{AB} = (-1,-1,-3)$ et $\overrightarrow{AC} = (-1,-2,-2)$, on a :

$$\overrightarrow{AB} \wedge \overrightarrow{BC} = \begin{pmatrix} -1 \\ -1 \\ -3 \end{pmatrix} \wedge \begin{pmatrix} -1 \\ -2 \\ -2 \end{pmatrix} = \begin{pmatrix} -4 \\ 1 \\ 1 \end{pmatrix}.$$

Une équation du plan est donc :

$$-4(x-1) + (y-2) + (z-3) = 0 \qquad \text{c'est-à-dire} \qquad -4x + y + z - 1 = 0.$$

Il est également possible de caractériser les plans de $\mathbb{R}^3$ à l'aide d'*équations paramétriques* : si le plan P contient le point $A := (x_0, y_0, z_0)$ et si les vecteurs $\vec{u} := (a_1, b_1, c_1)$ et $\vec{v} := (a_2, b_2, c_2)$ constituent une base de l'espace vectoriel qui dirige P :

$$\begin{aligned} P &= \left\{ M \in \mathbb{R}^3 \;\middle|\; \overrightarrow{AM} = \lambda\vec{u} + \mu\vec{v} \text{ avec } (\lambda,\mu) \in \mathbb{R}^2 \right\} \\ &= \left\{ (x_0 + \lambda a_1 + \mu a_2, y_0 + \lambda b_1 + \mu b_2, z_0 + \lambda c_1 + \mu c_2) \;\middle|\; (\lambda,\mu) \in \mathbb{R}^2 \right\}. \end{aligned}$$

On dit que :

$$\begin{cases} x = x_0 + \lambda a_1 + \mu a_2 \\ y = y_0 + \lambda b_1 + \mu b_2 \\ z = z_0 + \lambda c_1 + \mu c_2 \end{cases} \qquad (\lambda,\mu) \in \mathbb{R}^2$$

est une équation paramétrique de P.

Nous avons essentiellement exprimé le fait que $(A, \vec{u}, \vec{v})$ est un repère du plan affine P. Réciproquement, si on connaît une équation paramétrique d'un plan P de $\mathbb{R}^3$, on détermine facilement un point du plan et deux vecteurs non colinéaires dirigeant le plan. La proposition 25 de la page précédente permet alors d'en trouver une équation cartésienne.

Exercice 11.

Donner une équation cartésienne du plan d'équation paramétrique :

$$\begin{cases} x = 2 + 3\lambda + \mu \\ y = 1 - 4\lambda + \mu \\ z = -1 + \lambda - \mu \end{cases} \qquad (\lambda,\mu) \in \mathbb{R}^2.$$

Solution. Le plan contient le point $A := (2,1,-1)$ et est dirigé par les vecteurs :

$$\vec{u} := (3,-4,1) \qquad \text{et} \qquad \vec{v} := (1,1,-1).$$

Il a donc pour équation cartésienne :

$$a(x-2) + b(y-1) + c(z+1) = 0,$$

avec $(a,b,c) := \vec{u} \wedge \vec{v} = (3,4,7)$, *i.e.* $3x + 4y + 7z - 3 = 0$.

3.2.2 Intersection de deux plans

D'une part, l'intersection de deux plans de $\mathbb{R}^3$ est soit vide, soit un sous-espace affine de $\mathbb{R}^3$ de dimension 1 ou 2, c'est-à-dire un plan si $P_1 = P_2$ et une droite sinon. D'autre part, d'après le théorème 17 de la page 440, une droite de $\mathbb{R}^3$ peut toujours être représentée comme intersection de deux plans de $\mathbb{R}^3$.

Proposition 26. Dans $\mathbb{R}^3$, considérons deux plans P_1 et P_2 d'équations respectives :

$$a_1x + b_1y + c_1z + d_1 = 0 \qquad \text{et} \qquad a_2x + b_2y + c_2z + d_2 = 0.$$

1. Les plans P_1 et P_2 sont parallèles si, et seulement s'il existe un réel non nul λ tel que $(a_2, b_2, c_2) = \lambda(a_1, b_1, c_1)$ donc si, et seulement si, la matrice :

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{pmatrix}$$

est de rang 1 donc si, et seulement si :

$$\det\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} = \det\begin{pmatrix} a_1 & c_1 \\ a_2 & c_2 \end{pmatrix} = \det\begin{pmatrix} b_1 & c_1 \\ b_2 & c_2 \end{pmatrix} = 0.$$

2. Si les plans P_1 et P_2 ne sont pas parallèles, ils se coupent le long d'une droite dirigée par le vecteur :

$$\begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix} \wedge \begin{pmatrix} a_2 \\ b_2 \\ c_2 \end{pmatrix} = \begin{pmatrix} b_1c_2 - c_1b_2 \\ c_1a_2 - a_1c_2 \\ a_1b_2 - b_1a_2 \end{pmatrix}.$$

Démonstration. Les plans affines P_1 et P_2 sont parallèles si, et seulement si, les plans vectoriels qui les dirigent sont égaux. Ces sous-espaces vectoriels sont les noyaux respectifs des formes linéaires $L_1 : (x, y) \mapsto a_1x + b_1y + c_1z$ et $L_2 : (x, y) \mapsto a_2x + b_2y + c_2z$. Ces noyaux sont égaux si, et seulement si, $L_2 = \lambda L_1$ avec $\lambda \in \mathbb{R} \setminus \{0\}$.
Les points qui appartiennent à P_1 et P_2 sont les points (x, y, z) avec :

$$\begin{cases} a_1x + b_1y + c_1z + d_1 = 0 \\ a_2x + b_2y + c_2z + d_2 = 0. \end{cases}$$

Si les plans ne sont pas parallèles, alors le rang du système est égal à 2. Dans ce cas, d'après le théorème 64 de la page 377, l'ensemble des solutions est un sous-espace affine de dimension 1, c'est-à-dire une droite affine. Par définition du produit vectoriel (*cf.* (42), page 368), si :

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix} \wedge \begin{pmatrix} a_2 \\ b_2 \\ c_2 \end{pmatrix},$$

alors pour tout $(x, y, z) \in \mathbb{R}^3$:

$$ax + by + cz = \det\begin{pmatrix} a_1 & a_2 & x \\ b_1 & b_2 & y \\ c_1 & c_2 & z \end{pmatrix}.$$

Notons que (a, b, c) n'est pas le vecteur nul car (a_1, b_1, c_1) et (a_2, b_2, c_2) ne sont pas colinéaires.
On voit donc que :

$$a_1 a + b_1 b + c_1 c = \det \begin{pmatrix} a_1 & a_2 & a_1 \\ b_1 & b_2 & b_1 \\ c_1 & c_2 & c_1 \end{pmatrix} = 0$$

et le vecteur (a, b, c) appartient au noyau de la forme linéaire :

$$L_1 : (x, y, z) \mapsto a_1 x + b_1 y + c_1 z.$$

Par conséquent, (a, b, c) appartient au plan vectoriel qui dirige P_1. On montre de même que (a, b, c) appartient au plan vectoriel qui dirige P_2. Donc (a, b, c) est un vecteur directeur de la droite $D := P_1 \cap P_2$. ■

3.3 Droites de $\mathbb{R}^3$

3.3.1 Équations de droites

Nous venons de voir que deux plans de $\mathbb{R}^3$ qui ne sont pas parallèles se coupent le long d'une droite. Le théorème 17 de la page 440 implique que toute droite de $\mathbb{R}^3$ peut être caractérisée comme intersection de 2 plans. Autrement dit, on peut caractériser une droite de $\mathbb{R}^3$ à l'aide d'un système de deux équations affines. Pour trouver un tel système, il est possible de partir d'une *équation paramétrique* de la droite : la droite D passant par le point $A := (x_0, y_0, z_0)$ et dirigée par le vecteur $\vec{u} := (a, b, c)$ vérifie :

$$D = \left\{ \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} + \lambda \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad \lambda \in \mathbb{R} \right\}.$$

La droite D admet l'équation paramétrique :

$$\begin{cases} x = x_0 + \lambda a \\ y = y_0 + \lambda b \\ z = z_0 + \lambda c \end{cases} \qquad \lambda \in \mathbb{R}.$$

Pour trouver un système de deux équations caractérisant la droite D, on peut considérer les combinaisons linéaires :

$$\begin{cases} bx - ay = bx_0 - ay_0 \\ cy - bz = cy_0 - bz_0 \\ az - cx = az_0 - cx_0. \end{cases}$$

Parmi ces trois équations, on peut en trouver deux qui définissent des plans non parallèles (sinon, la proposition 26 de la page précédente impliquerait que les trois coefficients a, b et c seraient nuls). Ces deux équations caractérisent la droite D.

Exemple. Considérons la droite D passant par le point $(1, 2, 3)$ et dirigée par le vecteur $(0, 1, 2)$. Une équation paramétrique de la droite est :

$$\begin{cases} x = 1 \\ y = 2 + \lambda \\ z = 3 + 2\lambda. \end{cases}$$

On a donc :

$$\begin{cases} 1 \cdot x - 0 \cdot y = 1 \\ 2 \cdot y - 1 \cdot z = 1 \\ 0 \cdot z - 2 \cdot x = 2. \end{cases}$$

La première et la deuxième équation définissent deux plans qui ne sont pas parallèles et la droite D est caractérisée par le système d'équations :

$$\begin{cases} x = 1 \\ 2y - z = 1. \end{cases}$$

3.3.2 Intersection d'un plan avec une droite

Rappelons que par définition, la droite D dirigée par E est parallèle au plan P dirigé par F si, et seulement si, $E \subset F$. Si $\vec{u}$ est un vecteur directeur de D et si $(\vec{v}, \vec{w})$ est une base de l'espace vectoriel qui dirige P, la droite D est donc parallèle à P si, et seulement si, $\vec{u}$ est combinaison linéaire de $\vec{v}$ et $\vec{w}$.

Proposition 27.
Dans $\mathbb{R}^3$, considérons une droite D et un plan P d'équations respectives :

$$D : \begin{cases} a_1x + b_1y + c_1z + d_1 = 0 \\ a_2x + b_2y + c_2z + d_2 = 0 \end{cases} \quad \text{et} \quad P : \ a_3x + b_3y + c_3z + d_3 = 0.$$

1. La droite D est parallèle au plan P si, et seulement s'il existe $(\lambda, \mu) \in \mathbb{R}^2$ tel que $(a_3, b_3, c_3) = \lambda(a_1, b_1, c_1) + \mu(a_2, b_2, c_2)$ donc si, et seulement si :

$$\det \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} = 0.$$

2. Si la droite D n'est pas parallèle au plan P, l'intersection $D \cap P$ contient exactement un point.

Démonstration. La droite D est parallèle au plan P si, et seulement si, le vecteur directeur :

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix} \wedge \begin{pmatrix} a_2 \\ b_2 \\ c_2 \end{pmatrix}$$

appartient au plan vectoriel qui dirige P donc si, et seulement si, $a_3a + b_3b + c_3c = 0$. Or, par définition du produit vectoriel :

$$a_3a + b_3b + c_3c = \det \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix}.$$

Donc, la droite D est parallèle au plan P si, et seulement si :

$$\det \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix} = 0.$$

Les points qui appartiennent à $D \cap P$ sont les points (x, y, z) avec :

$$\begin{cases} a_1x + b_1y + c_1z + d_1 = 0 \\ a_2x + b_2y + c_2z + d_2 = 0 \\ a_3x + b_3y + c_3z + d_3 = 0. \end{cases}$$

Si la droite D n'est pas parallèle au plan P, le rang du système est égal à 3 et d'après le théorème 64 de la page 377, le système admet une unique solution. ■

Exercice 12.

Soit P le plan d'équation $ax + by + cz + d = 0$.

1. Montrer que le vecteur (a, b, c) n'est pas contenu dans l'espace vectoriel qui dirige P.
2. Montrer que la droite D passant par $(0, 0, 0)$ et dirigée par (a, b, c) coupe P en un unique point A.
3. Déterminer les coordonnées de A.

Solution.

1. Les vecteurs qui dirigent le plan P sont dans le noyau de l'application linéaire $L : (x, y) \mapsto ax + by + cz$, et $L(a, b, c) = a^2 + b^2 + c^2 > 0$.
2. La droite D n'est pas parallèle au plan P. L'intersection $D \cap P$ contient donc un unique point.
3. $A = (ta, tb, tc)$ avec $t = -c/(a^2 + b^2 + c^2)$.

Cet exercice montre qu'un plan d'équation $ax + by + cz + d = 0$ contient le point :

$$A := \left(-\frac{ad}{a^2 + b^2 + c^2}, -\frac{bd}{a^2 + b^2 + c^2}, -\frac{cd}{a^2 + b^2 + c^2}\right).$$

Pour trouver une base de vecteurs qui dirige le plan P, il faut trouver deux vecteurs non colinéaires contenus dans le noyau de la partie linéaire :

$$L : (x, y, z) \mapsto ax + by + cz.$$

Il n'y a pas de moyen canonique de choisir deux tels vecteurs.

3.4 Géométrie euclidienne dans $\mathbb{R}^2$ et $\mathbb{R}^3$

Nous allons maintenant introduire les notions de produit scalaire, d'orthogonalité et de distance. Ces notions et leurs généralisations à des espaces vectoriels ou affines quelconques seront étudiées systématiquement dans le volume de L2. Ici, nous nous contenterons de les utiliser dans le cas de $\mathbb{R}^2$ et de $\mathbb{R}^3$. À partir du moment où la notion de produit scalaire intervient, on fait de la géométrie euclidienne.

Dans toute cette partie, E désigne l'espace vectoriel $\mathbb{R}^2$ ou $\mathbb{R}^3$, suivant les cas, et X désigne l'espace affine $\mathbb{R}^2$ ou $\mathbb{R}^3$. Quand nous parlerons de vecteurs, sans plus de précision, ce seront des vecteurs de l'espace vectoriel $\mathbb{R}^2$ ou $\mathbb{R}^3$. Quand nous parlerons de points sans plus de précision, ce seront des points de l'espace affine $\mathbb{R}^2$ ou $\mathbb{R}^3$.

Définition 16. La *norme d'un vecteur* $\vec{u} := (x, y) \in \mathbb{R}^2$ est : $\|\vec{u}\| := \sqrt{x^2 + y^2}$.
La *norme d'un vecteur* $\vec{u} := (x, y, z) \in \mathbb{R}^3$ est : $\|\vec{u}\| := \sqrt{x^2 + y^2 + z^2}$.

Proposition 28 (Propriétés de la norme). Pour tous vecteurs $\vec{u}$ et $\vec{v}$ et pour tout $\lambda \in \mathbb{R}$, on a les trois propriétés suivantes :

1. $\|\vec{u}\| \geqslant 0$ et $\|\vec{u}\| = 0$ si, et seulement si, $\vec{u} = \overrightarrow{0}$.
2. $\|\lambda\vec{u}\| = |\lambda| \cdot \|\vec{u}\|$.
3. $\|\vec{u} + \vec{v}\| \leqslant \|\vec{u}\| + \|\vec{v}\|$ (inégalité triangulaire).

Démonstration. Les deux premières propriétés sont immédiates. Nous laissons au lecteur le soin de les vérifier. Pour une démonstration de l'inégalité triangulaire (dans un cadre un peu plus général), nous renvoyons le lecteur à la proposition 4 de la page 781 du module IV.6. ■

Définition 17. La *distance entre deux points* A et B est :

$$d(A, B) := \|\overrightarrow{AB}\|.$$

La distance entre deux points A et B est toujours positive. Elle est nulle si, et seulement si, $A = B$. Si $A := (x_0, y_0, z_0)$ et $B := (x_1, y_1, z_1)$, alors :

$$d(A, B) = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}.$$

Exemple. La distance entre les points $A := (0, 1, 3)$ et $B := (1, 2, -1)$ dans $\mathbb{R}^3$ est :

$$d(A, B) = \sqrt{(1 - 0)^2 + (2 - 1)^2 + (-1 - 3)^2} = \sqrt{18}.$$

L'inégalité triangulaire implique que pour tout triplet de points (A, B, C), on a :

$$d(A, C) \leqslant d(A, B) + d(B, C).$$

Définition 18 (Distance d'un point à un ensemble).
La distance du point $A \in X$ au sous-ensemble $Y \subset X$ est :

$$d(A, Y) := \inf_{M \in Y} d(A, M).$$

Exemple. La distance du point $(0, 0, z)$ au plan horizontal d'équation $z = 0$ est $|z|$.

Définition 19. Le *produit scalaire de deux vecteurs* $\vec{u} := (x_1, y_1)$ et $\vec{v} := (x_2, y_2)$ de $\mathbb{R}^2$ est $\left(\vec{u} \mid \vec{v}\right) := x_1x_2 + y_1y_2$.
Le *produit scalaire de deux vecteurs* $\vec{u} := (x_1, y_1, z_1) \in \mathbb{R}^3$ et $\vec{v} := (x_2, y_2, z_2) \in \mathbb{R}^3$ est $\left(\vec{u} \mid \vec{v}\right) := x_1x_2 + y_1y_2 + z_1z_2$.

Proposition 29 (Propriétés du produit scalaire). Si $\vec{u}$ et $\vec{v}$ sont deux vecteurs, $\left(\vec{u} \mid \vec{v}\right) = \left(\vec{v} \mid \vec{u}\right)$. Les applications $\vec{w} \mapsto \left(\vec{u} \mid \vec{w}\right)$ et $\vec{w} \mapsto \left(\vec{w} \mid \vec{v}\right)$ sont des formes linéaires. Enfin, $\left(\vec{u} \mid \vec{u}\right) \geqslant 0$ avec égalité si, et seulement si, $\vec{u} = \overrightarrow{0}$.

Ces propriétés sont immédiates. Nous laissons au lecteur le soin de les vérifier.

Définition 20.
Deux vecteurs $\vec{u}$ et $\vec{v}$ sont *orthogonaux* si $\left(\vec{u} \mid \vec{v}\right) = 0$. On note $\vec{u} \perp \vec{v}$.

Exemple.
Dans $\mathbb{R}^2$ et dans $\mathbb{R}^3$, les vecteurs de la base canoniques sont deux à deux orthogonaux.

Définition 21. Une base est *orthonormée* si les vecteurs de la base sont deux à deux orthogonaux et de norme 1.

Exemples.

1. Dans $\mathbb{R}^2$ et dans $\mathbb{R}^3$, la base canonique est orthonormée.
2. Rappelons que si $\vec{u}$ et $\vec{v}$ sont orthogonaux et de norme 1, alors la base $(\vec{u}, \vec{v}, \vec{u} \wedge \vec{v})$ est orthonormée. Pour cette propriété, nous renvoyons le lecteur au module II.7, page 367.

Les deux résultats qui suivent découlent directement des définitions de la norme d'un vecteur et du produit scalaire de deux vecteurs. Nous laissons les démonstrations aux soins du lecteur.

Théorème 30. Si $\vec{u}$ et $\vec{v}$ sont deux vecteurs, on a :

1. $\left\|\vec{u}\right\|^2 = \left(\vec{u} \mid \vec{u}\right)$;
2. $\left\|\vec{u} + \vec{v}\right\|^2 = \left\|\vec{u}\right\|^2 + \left\|\vec{v}\right\|^2 + 2\left(\vec{u} \mid \vec{v}\right)$.

Comme corollaire, on a le fameux théorème de Pythagore.

Théorème 31 (Pythagore).
Les vecteurs $\vec{u}$ et $\vec{v}$ sont orthogonaux si, et seulement si :

$$\left\|\vec{u} + \vec{v}\right\|^2 = \left\|\vec{u}\right\|^2 + \left\|\vec{v}\right\|^2.$$

Mentionnons également l'inégalité de Cauchy-Schwarz dont la démonstration (dans un cadre un peu plus général) est donnée à la proposition 3 de la page 781 du module IV.6.

Théorème 32 (Inégalité de Cauchy-Schwarz).
Si $\vec{u}$ et $\vec{v}$ sont deux vecteurs, on a :

$$\left|\left(\vec{u} \mid \vec{v}\right)\right| \leqslant \left\|\vec{u}\right\| \cdot \left\|\vec{v}\right\|$$

avec égalité si, et seulement si, $\vec{u}$ et $\vec{v}$ sont colinéaires.

Par conséquent, lorsque $\vec{u} \neq 0$ et $\vec{v} \neq 0$, on peut écrire :

$$\frac{\big(\vec{u} \mid \vec{v}\big)}{\|\vec{u}\|\|\vec{v}\|} = \cos\theta \in [-1,1] \quad \text{avec} \quad \theta \in [0,\pi].$$

L'angle θ est, par définition, l'angle entre les vecteurs $\vec{u}$ et $\vec{v}$.

Proposition 33. Si F est un sous-espace vectoriel de E, l'ensemble :

$$F^{\perp} := \big\{\vec{v} \in E \mid \forall \vec{u} \in F,\ \big(\vec{u} \mid \vec{v}\big) = 0\big\}$$

est un sous-espace vectoriel de E.

Démonstration. Pour tout $\vec{u} \in F$, $\big(\vec{u} \mid \overrightarrow{0}\big) = 0$. Par conséquent, $\overrightarrow{0} \in F^{\perp}$. Si $\vec{v}$ et $\vec{w}$ appartiennent à $F^{\perp}$ et si $\lambda \in \mathbb{R}$, alors pour tout $\vec{u} \in F$, on a :

$$\big(\vec{u} \mid \lambda\vec{v} + \vec{w}\big) = \lambda\big(\vec{u} \mid \vec{v}\big) + \big(\vec{u} \mid \vec{w}\big) = 0.$$

Donc, $\lambda\vec{v} + \vec{w}$ appartient à $F^{\perp}$. ∎

Définition 22. Si F est un sous-espace vectoriel de E, l'ensemble $F^{\perp}$ est appelé l'*orthogonal* de F. Si $\vec{u} \in F^{\perp}$, $\vec{u}$ est *orthogonal* à F. Si $Y \subset X$ est un sous-espace affine dirigé par F et si $\vec{u} \in F^{\perp}$, $\vec{u}$ est *orthogonal* à Y. Si Y et Y' sont deux sous-espaces affine de X dirigés respectivement par F et $F^{\perp}$, Y et Y' sont *orthogonaux*.

Exemple. Rappelons que si $\vec{u}$, $\vec{v}$ et $\vec{w}$ sont trois vecteurs de $\mathbb{R}^3$ alors :

$$[\vec{u}, \vec{v}, \vec{w}] = \big(\vec{u} \wedge \vec{v} \mid \vec{w}\big),$$

où $\vec{u} \wedge \vec{v}$ est le produit vectoriel de $\vec{u}$ et $\vec{v}$. Si $\vec{u}$ et $\vec{v}$ ne sont pas colinéaires, $\vec{u} \wedge \vec{v}$ est donc orthogonal au plan engendré par $\vec{u}$ et $\vec{v}$ (le produit mixte est nul si $\vec{w}$ appartient à ce plan).

Proposition 34. Si $D \subset \mathbb{R}^2$ est une droite d'équation $ax + by + c = 0$, alors le vecteur (a,b) est orthogonal à D. Si $P \subset \mathbb{R}^3$ est un plan d'équation $ax + by + cz + d = 0$, alors le vecteur (a,b,c) est orthogonal à P.

Démonstration. Si $D \subset \mathbb{R}^2$ est une droite d'équation $ax + by + c = 0$, alors tout vecteur qui dirige D est dans le noyau de la partie linéaire :

$$L : (x,y) \mapsto ax + by = \big((a,b) \mid (x,y)\big),$$

et donc est orthogonal à (a,b).

Si $P \subset \mathbb{R}^3$ est un plan d'équation $ax + by + cz + d = 0$, alors tout vecteur du plan qui dirige P est dans le noyau de la partie linéaire :

$$L : (x,y,z) \mapsto ax + by + cz = \big((a,b,c) \mid (x,y,z)\big),$$

et donc est orthogonal à (a,b,c). ∎

Théorème 35. Les espaces F et $F^\perp$ sont supplémentaires.

Démonstration. Si $F = \{0\}$, alors $F^\perp = E$ car pour tout $\vec{v} \in E$, $\left(\overrightarrow{0} \mid \vec{v}\right) = 0$.
Si $F = E$ alors $F^\perp = \{0\}$ car pour $\vec{v} \neq \overrightarrow{0}$, on a $\left(\vec{v} \mid \vec{v}\right) = \|\vec{v}\|^2 \neq 0$.
Dans ces deux cas, F et $F^\perp$ sont supplémentaires.
Observons que si $\vec{v} \in F \cap F^\perp$, alors $\|\vec{v}\|^2 = \left(\vec{v} \mid \vec{v}\right) = 0$. Donc, $\vec{v} = \overrightarrow{0}$ et $F \cap F^\perp = \{0\}$.
Il suffit donc de montrer que $\dim F + \dim F^\perp \geqslant \dim E$.
Si F est un hyperplan de E (une droite vectorielle dans $\mathbb{R}^2$ et un plan vectoriel dans $\mathbb{R}^3$), alors F est le noyau d'une forme linéaire $L : E \to \mathbb{R}$ non identiquement nulle. Cette forme linéaire est de la forme $(x, y) \mapsto ax + by$ avec $(a, b) \neq \overrightarrow{0}$ si $E = \mathbb{R}^2$ et de la forme $(x, y, z) \mapsto ax + by + cz$ avec $(a, b, c) \neq \overrightarrow{0}$ si $E = \mathbb{R}^3$. Dans le premier cas, le vecteur (a, b) appartient à $F^\perp$ et dans le second cas, le vecteur (a, b, c) appartient à $F^\perp$. Par conséquent, $\dim F^\perp \geqslant 1$.
Il ne reste plus qu'à traiter le cas où F est une droite de $\mathbb{R}^3$. Alors, cette droite est l'intersection de deux plans d'équations $a_1x + b_1y + c_1z = 0$ et $a_2x + b_2y + c_2z = 0$. Les deux vecteurs (a_1, b_1, c_1) et (a_2, b_2, c_2) sont non colinéaires (sinon les deux plans seraient confondus) et appartiennent tous deux à $F^\perp$. Par conséquent $\dim F^\perp \geqslant 2$. ■

Définition 23. Soit Y un sous-espace affine de X dirigé par F.

- La *projection orthogonale* sur Y est la projection sur Y parallèlement à $F^\perp$.
- La *symétrie orthogonale* par rapport à Y est la symétrie par rapport à Y parallèlement à $F^\perp$.
- L'*affinité orthogonale* de rapport k par rapport à Y est l'affinité de rapport k par rapport à Y parallèlement à $F^\perp$.

Proposition 36. Si A est un point de X et si Y est un sous-espace affine de X, il existe un unique point $B \in Y$ tel que $d(A, B) = d(A, Y)$. Le point B est le projeté orthogonal de A sur Y.

Démonstration. Soit F le sous–espace vectoriel qui dirige Y. Soit B le projeté orthogonal de A sur Y. Il suffit de montrer que pour tout $M \in Y$ avec $M \neq B$, on a $d(A, M) > d(A, B)$. D'une part, le vecteur $\overrightarrow{AB}$ appartient à $F^\perp$. D'autre part, le vecteur $\overrightarrow{BM}$ appartient à F. On a donc $\overrightarrow{AB} \perp \overrightarrow{BM}$ et d'après le théorème de Pythagore :

$$\|\overrightarrow{AM}\|^2 = \|\overrightarrow{AB}\|^2 + \|\overrightarrow{BM}\|^2 > \|\overrightarrow{AB}\|^2. \quad ■$$

FIGURE III.1.7 – Projeté orthogonal

Exercice 13.
Déterminer $d(A, D)$ pour $A := (x_0, y_0) \in \mathbb{R}^2$ lorsque $D \subset \mathbb{R}^2$ est la droite d'équation $ax + by + c = 0$.
Solution. Le projeté orthogonal B de A sur D appartient à la droite orthogonale à D passant par A, il est donc de la forme $(x_0, y_0) + \lambda(a, b)$ avec $\lambda \in \mathbb{R}$. En écrivant que $B \in D$, on obtient :

$$a(x_0 + \lambda a) + b(y_0 + \lambda b) + c = 0.$$

Par conséquent :

$$\lambda = -\frac{ax_0 + by_0 + c}{a^2 + b^2} \quad \text{et} \quad d(A, D) = \|\overrightarrow{AB}\| = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}}.$$

Exercice 14.
Déterminer $d(A, P)$ pour $A := (x_0, y_0, z_0) \in \mathbb{R}^3$ lorsque $P \subset \mathbb{R}^3$ est le plan d'équation $ax + by + cz + d = 0$.
Solution. Le projeté orthogonal B de A sur P appartient à la droite orthogonale à P passant par A, il est donc de la forme $(x_0, y_0, z_0) + \lambda(a, b, c)$ avec $\lambda \in \mathbb{R}$. En écrivant que $B \in P$, on obtient :

$$a(x_0 + \lambda a) + b(y_0 + \lambda b) + c(z_0 + \lambda c) + d = 0.$$

Par conséquent :

$$\lambda = -\frac{ax_0 + by_0 + cz_0 + d}{a^2 + b^2 + c^2} \quad \text{et} \quad d(A, P) = \|\overrightarrow{AB}\| = \frac{|ax_0 + by_0 + cz_0 + d|}{\sqrt{a^2 + b^2 + c^2}}.$$

4 Les coniques

Nous avons vu deux manières de décrire une droite dans $\mathbb{R}^2$. La première consiste à en donner une équation cartésienne. La seconde consiste à en donner une équation paramétrique. Dans le module suivant, nous étudierons les courbes paramétrées de $\mathbb{R}^2$ et de $\mathbb{R}^3$. Maintenant, nous allons étudier les courbes du plan définies par une équation algébrique de degré 2, c'est-à-dire une équation de la forme :

$$ax^2 + bxy + cy^2 + dx + ey + f = 0 \qquad \text{avec} \qquad (a, b, c) \neq (0, 0, 0).$$

Dans le volume de L2, nous étudierons de manière plus générale les courbes du plan définies par des équations de la forme $f(x, y) = 0$ ou des surfaces de l'espaces définies par des équations de la forme $f(x, y, z) = 0$. L'outil central sera le théorème des fonctions implicites dont nous ne disposons pas pour le moment.
Nous allons voir ici qu'une courbe du plan donnée par une équation algébrique de degré 2 est une conique : en général, une ellipse, une parabole ou une hyperbole, mais ce peut être aussi l'ensemble vide, un point ou la réunion de deux droites (éventuellement confondues).

4.1 Cercles

Définition 24. Étant donnés un point $A \in \mathbb{R}^2$ et un réel $R > 0$, le *cercle* $C(A, R)$ de centre A et de rayon R est l'ensemble des points $M \in \mathbb{R}^2$ tels que $d(A, M) = R$.

Notons que $d(A, M) = R$ si, et seulement si, $\|\overrightarrow{AM}\|^2 = R^2$. Le résultat suivant est donc immédiat.

Proposition 37. Le cercle de centre $A := (x_0, y_0)$ de rayon $R > 0$ admet l'équation cartésienne :

$$(x - x_0)^2 + (y - y_0)^2 = R^2.$$

Il est possible de développer cette équation sous la forme :

$$x^2 + y^2 - 2x_0x - 2y_0y + x_0^2 + y_0^2 - R^2 = 0.$$

Remarquons que le coefficient de x^2 est égal à celui de y^2 et que le coefficient de xy est nul.

Proposition 38. Étant donnés des réels $a \neq 0$, b, c, et d, l'ensemble des points $M := (x, y) \in \mathbb{R}^2$ tels que :

$$ax^2 + ay^2 + bx + cy + d = 0$$

est :

- l'ensemble vide si $b^2 + c^2 - 4ad < 0$;
- le point $\big(-b/(2a), -c/(2a)\big)$ si $b^2 + c^2 - 4ad = 0$;
- le cercle $C(A, R)$ si $b^2 + c^2 - 4ad > 0$ avec :

$$A := \left(-\frac{b}{2a}, -\frac{c}{2a}\right) \quad \text{et} \quad R := \sqrt{\frac{b^2 + c^2 - 4ad}{4a^2}}.$$

Démonstration. En divisant par $a \neq 0$ et en posant :

$$x_0 := -\frac{b}{2a}, \quad y_0 := -\frac{c}{2a} \quad \text{et} \quad \alpha := x_0^2 + y_0^2 - \frac{d}{a} = \frac{b^2 + c^2 - 4ad}{4a^2},$$

on voit que $M := (x, y)$ est solution de l'équation $ax^2 + ay^2 + bx + cy + d = 0$ si, et seulement si :

$$x^2 + y^2 - 2x_0x - 2y_0y + x_0^2 + y_0^2 - \alpha = 0,$$

c'est-à-dire :

$$(x - x_0)^2 + (y - y_0)^2 = \alpha.$$

Si $\alpha < 0$, il n'y a pas de solution. Si $\alpha = 0$, la seule solution est le point $A = (x_0, y_0)$.

Si $\alpha > 0$, l'ensemble des solutions est l'ensemble des points M tels que $d(A, M)^2 = \alpha$, c'est-à-dire le cercle de centre A et de rayon $\sqrt{\alpha}$. ∎

4.2 Coniques

Définition 25. Étant donnés un point $F \in \mathbb{R}^2$, une droite $D \subset \mathbb{R}^2$ ne contenant pas F et un réel $e > 0$, l'ensemble $\mathcal{C}$ des points $M \in \mathbb{R}^2$ tels que :

$$d(M, F) = e \cdot d(M, D)$$

est :

- une *ellipse* si $0 < e < 1$,
- une *parabole* si $e = 1$,
- une *hyperbole* si $e > 1$.

La courbe $\mathcal{C}$ est la *conique* de *foyer* F, de *directrice* D associée au foyer F et d'*excentricité* e.

Notons que si le foyer F a pour coordonnées (x_0, y_0) et si la directrice D a pour équation cartésienne $a_0x + b_0y + c_0 = 0$, alors la conique $\mathcal{C}$ a pour équation :

$$(x - x_0)^2 + (y - y_0)^2 = e^2 \frac{(a_0x + b_0y + c_0)^2}{a_0^2 + b_0^2}.$$

C'est une équation algébrique de degré 2. En multipliant par $a_0^2 + b_0^2$, on peut l'écrire sous la forme :

$$ax^2 + 2bxy + cy^2 + dx + fy + g = 0$$

avec :

$$a := a_0^2 + b_0^2 - e^2a_0^2, \quad b := -a_0b_0e^2 \quad \text{et} \quad c := a_0^2 + b_0^2 - e^2b_0^2.$$

Observons que :

$$ac - b^2 = (a_0^2 + b_0^2)^2(1 - e^2).$$

On voit donc que :

- lorsque $\mathcal{C}$ est une ellipse, $ac - b^2 > 0$,
- lorsque $\mathcal{C}$ est une parabole, $ac - b^2 = 0$,
- lorsque $\mathcal{C}$ est une hyperbole, $ac - b^2 < 0$.

4.2.1 Parabole

Quitte à faire un changement de repère orthonormé, on peut supposer que l'origine du repère est le foyer F et que la directrice associée est une droite horizontale d'équation $y + \alpha = 0$. Dans ce cas, l'équation de la parabole est $x^2 + y^2 = (y + \alpha)^2$. On peut simplifier cette équation : $x^2 - 2\alpha y - \alpha^2 = 0$. On peut également l'écrire sous la forme :

$$y = \frac{1}{2\alpha}x^2 - \frac{\alpha}{2}.$$

La parabole passe par le point $S := (0, \alpha/2)$ qui se trouve à mi-distance entre le foyer F et la directrice D. Ce point s'appelle le sommet de la parabole. La droite D_1 d'équation $x = 0$ est un axe de symétrie de la parabole : le point (x, y) appartient à la parabole si, et seulement si, le point $(-x, y)$ appartient à la parabole. La droite D_1 passe par le foyer F et est orthogonale à la directrice D.

FIGURE III.1.8 – La parabole de foyer F et de directrice D.

4.2.2 Ellipse

Comme précédemment, supposons que l'origine du repère soit le foyer F et que la directrice associée D soit une droite horizontale d'équation $y + \alpha = 0$. L'équation de l'ellipse est alors :

$$x^2 + y^2 = e^2(y + \alpha)^2$$

que l'on peut écrire :

$$x^2 + (1 - e^2)\left(y - \frac{\alpha e^2}{1 - e^2}\right)^2 = \alpha^2 e^2 + \frac{\alpha^2 e^4}{1 - e^2}.$$

Si l'on fait le changement de coordonnées :

$$X := x \quad \text{et} \quad Y := y - \frac{\alpha e^2}{1 - e^2},$$

l'équation devient :

$$X^2 + (1 - e^2)Y^2 = \frac{\alpha^2 e^2}{1 - e^2}.$$

Cela revient à se placer dans le repère $(O', \vec{\imath}, \vec{\jmath})$ avec :

$$O' := \left(0, \frac{\alpha e^2}{1 - e^2}\right).$$

Puisque $0 < e < 1$, il est également possible de l'écrire sous la forme :

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1 \quad \text{avec} \quad a := \frac{\alpha e}{\sqrt{1 - e^2}} \quad \text{et} \quad b := \frac{\alpha e}{1 - e^2}.$$

Le point O' est centre de symétrie de l'ellipse : le point (X, Y) appartient à l'ellipse si, et seulement si, le point $(-X, -Y)$ appartient à l'ellipse. La droite D_1 d'équation $X = 0$

est un axe de symétrie de l'ellipse : $(X,Y) \in \mathcal{C}$ si, et seulement si, $(-X,Y) \in \mathcal{C}$. La droite D_2 d'équation $Y = 0$ est un axe de symétrie de l'ellipse : $(X,Y) \in \mathcal{C}$ si, et seulement si, $(X,-Y) \in \mathcal{C}$. La droite D_1 est orthogonale à la directrice D et passe par le foyer F (ainsi que par le centre de symétrie O'). La droite D_2 est parallèle à la directrice D et passe par le centre de symétrie O'. L'ellipse coupe la droite D_1 aux points $(X,Y) = (0,b)$ et $(X,Y) = (0,-b)$. L'ellipse coupe la droite D_2 aux points $(X,Y) = (a,0)$ et $(X,Y) = (-a,0)$. Comme $|a| > |b| > 0$, la droite D_1 s'appelle le grand axe de l'ellipse et la droite D_2 s'appelle le petit axe de l'ellipse. L'ellipse est l'image du cercle centré en O' de rayon $|a|$ par l'affinité orthogonale de rapport $|b/a|$ d'axe D_1. En effet, le point $M := O' + X\vec{\imath} + Y\vec{\jmath}$ appartient au cercle de centre O' et de rayon $|a|$ si, et seulement si, le point $M' := O' + X\vec{\imath} + |b/a|Y\vec{\jmath}$ appartient à l'ellipse $\mathcal{C}$:

$$X^2+Y^2 = a^2 \quad \Longleftrightarrow \quad \frac{X^2}{a^2}+\frac{\big(|b/a|Y\big)^2}{b^2} = 1.$$

De même, l'ellipse est l'image du cercle centré en O' de rayon $|b|$ par l'affinité orthogonale de rapport $|a/b|$ d'axe D_2.

Enfin, comme O' est centre de symétrie de l'ellipse, elle admet un autre foyer F' et une directrice D' associée à F' : F' est le symétrique de F par rapport à O' et D' est la droite symétrique de D par rapport à O'.

FIGURE III.1.9 – L'ellipse de foyer F, de directrice D et d'excentricité $1/2$.

4.2.3 Hyperbole

Comme dans le cas d'une ellipse, on suppose d'abord que l'origine du repère est le foyer F et que la directrice associée est une droite horizontale d'équation $y + \alpha = 0$. L'équation de l'hyperbole est alors :

$$x^2 + y^2 = e^2(y+\alpha)^2$$

que l'on peut écrire :

$$x^2 + (1-e^2)\left(y + \frac{\alpha e^2}{e^2-1}\right)^2 = \alpha^2 e^2 - \frac{\alpha^2 e^4}{e^2-1}\cdot$$

Si on fait le changement de coordonnées :

$$X := x \quad \text{et} \quad Y := y + \frac{\alpha e^2}{e^2-1},$$

ce qui revient à se placer dans le repère $(O', \vec{\imath}, \vec{\jmath})$ avec :

$$O' := \left(0, -\frac{\alpha e^2}{e^2-1}\right),$$

l'équation devient :

$$-X^2 + (e^2 - 1)Y^2 = \frac{\alpha^2 e^2}{e^2 - 1}\cdot$$

Puisque $e > 1$, il est également possible de l'écrire sous la forme :

$$-\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1 \quad \text{avec} \quad a := \frac{\alpha e}{\sqrt{e^2 - 1}} \quad \text{et} \quad b := \frac{\alpha e}{e^2 - 1}\cdot$$

Par conséquent, le point $M := O' + X\vec{\imath} + Y\vec{\jmath}$ appartient à l'hyperbole si, et seulement si :

$$X \in \mathbb{R} \text{ et } Y = b\sqrt{1 + \frac{X^2}{a^2}} \quad \text{ou} \quad X \in \mathbb{R} \text{ et } Y = -b\sqrt{1 + \frac{X^2}{a^2}}\cdot$$

L'hyperbole est donc la réunion de deux graphes symétriques par rapport à l'axe des X.
Le point O' est centre de symétrie de l'hyperbole : le point (X, Y) appartient à l'hyperbole si, et seulement si, le point $(-X, -Y)$ appartient à l'hyperbole. La droite D_1 d'équation $X = 0$ est un axe de symétrie de l'hyperbole : $(X, Y) \in \mathcal{C}$ si, et seulement si, $(-X, Y) \in \mathcal{C}$. La droite D_2 d'équation $Y = 0$ est un axe de symétrie de l'hyperbole : $(X, Y) \in \mathcal{C}$ si, et seulement si, $(X, -Y) \in \mathcal{C}$. La droite D_1 est orthogonale à la directrice D et passe par le foyer F (ainsi que par le centre de symétrie O'). La droite D_2 est parallèle à la directrice D et passe par le centre de symétrie O'. Notons que l'hyperbole coupe la droite D_1 aux points $(X, Y) = (0, b)$ et $(X, Y) = (0, -b)$. L'hyperbole ne coupe pas la droite D_2. La droite D_1 s'appelle l'axe transverse de l'hyperbole et la droite D_2 s'appelle l'axe non transverse de l'hyperbole.

FIGURE III.1.10 – L'hyperbole de foyer F, de directrice D et d'excentricité 2.

L'ensemble des points (X, Y) tels que $-X^2/a^2 + Y^2/b^2 = 0$ est la réunion de deux droites d'équations $-X/a + Y/b = 0$ et $X/a + Y/b = 0$. En effet :

$$-\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 0 \quad \Longleftrightarrow \quad \left(-\frac{X}{a} + \frac{Y}{b}\right) \cdot \left(\frac{X}{a} + \frac{Y}{b}\right) = 0.$$

Ces deux droites sont appelées les asymptotes de l'hyperbole. Si les deux asymptotes sont orthogonales, ce qui se produit si, et seulement si, $|a| = |b|$, l'hyperbole est *équilatère*.
Enfin, comme O' est centre de symétrie de l'hyperbole, elle admet un autre foyer F' et une directrice D' associée à F' : F' est le symétrique de F par rapport à O' et D' est la droite symétrique de D par rapport à O'.

4.3 Équations de degré 2

Nous souhaitons maintenant étudier de manière générale les courbes caractérisées par une équation de la forme $ax^2 + 2bxy + cy^2 + dx + fy + g = 0$ avec $(a, b, c) \neq (0, 0, 0)$. Rappelons que O désigne le point de coordonnées $(0, 0)$ et que $(\vec{\imath}, \vec{\jmath})$ est la base canonique de $\mathbb{R}^2$.

Proposition 39. Soit $\mathcal{C}$ l'ensemble des points $M := (x, y) \in \mathbb{R}^2$ tels que $ax^2 + 2bxy + cy^2 + dx + fy + g = 0$. Il existe une base orthonormée $(\vec{u}, \vec{v})$ de $\mathbb{R}^2$ et des réels λ, μ, α et β tels que :

$$M = O + X\vec{u} + Y\vec{v} \in \mathcal{C} \quad \Longleftrightarrow \quad \lambda X^2 + \mu Y^2 + \alpha X + \beta Y + g = 0.$$

Les vecteurs $\vec{u}$ et $\vec{v}$ sont des vecteurs propres de l'endomorphisme $L : \mathbb{R}^2 \to \mathbb{R}^2$ défini par :

$$L : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} a & b \\ b & c \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ bx + cy \end{pmatrix}.$$

Les réels λ et μ sont les valeurs propres associées. En particulier :

$$\lambda + \mu = a + c \quad \text{et} \quad \lambda\mu = ac - b^2.$$

Enfin, les réels α et β sont définis par :

$$(d, f) = \alpha\vec{u} + \beta\vec{v}.$$

Remarque. Autrement dit, en changeant de base orthonormée, on peut se ramener au cas où le coefficient de XY est nul.

Démonstration. Commençons par montrer que L est diagonalisable dans une base orthonormée.

Lemme 40. Il existe une base orthonormée $\mathcal{B} := (\vec{u}, \vec{v})$ de vecteurs propres de L.

Démonstration. Notons que pour tous vecteurs $\vec{u} := (x, y)$ et $\vec{v} := (x', y')$, on a :

$$\begin{aligned} \big(L(\vec{v}) \mid \vec{u} \big) &= (ax' + by') \cdot x + (bx' + cy') \cdot y \\ &= x' \cdot (ax + by) + y' \cdot (bx + cy) \\ &= \big(\vec{v} \mid L(\vec{u}) \big). \end{aligned}$$

Par ailleurs, les valeurs propres de L sont les racines du polynôme caractéristique :

$$(a - \lambda)(c - \lambda) - b^2 = \lambda^2 - (a + c)\lambda + ac - b^2.$$

Le discriminant de ce polynôme est :

$$(a+c)^2 - 4(ac-b^2) = (a-c)^2 + 4b^2 \geqslant 0.$$

On pourra donc toujours trouver un vecteur propre $\vec{u}$ associé à une valeur propre λ. Si $\vec{v}$ est orthogonal à $\vec{u}$, alors :

$$\big(\, L(\vec{v}) \mid \vec{u} \,\big) = \big(\, \vec{v} \mid L(\vec{u}) \,\big) = \lambda\big(\, \vec{v} \mid \vec{u} \,\big) = 0.$$

Par conséquent, $L(\vec{v})$ est orthogonal à $\vec{u}$ et donc, colinéaire à $\vec{v}$. Autrement dit $\vec{v}$ est un vecteur propre de L. En divisant les vecteurs $\vec{u}$ et $\vec{v}$ par leurs normes, on obtient une base orthonormée de vecteurs propres. ■

En écrivant que les deux valeurs propres λ et μ sont racines du polynôme caractéristique, on obtient les relations :

$$\lambda + \mu = a + c \quad \text{et} \quad \lambda\mu = ac - b^2.$$

Observons que pour tout point $M := (x, y) \in \mathbb{R}^2$, on a :

$$ax^2 + 2bxy + cy^2 = (ax + by) \cdot x + (bx + cy) \cdot y = \big(\, L(\overrightarrow{OM}) \mid \overrightarrow{OM} \,\big).$$

Si $M = O + x\vec{\imath} + y\vec{\jmath} = O + X\vec{u} + Y\vec{v}$, on a donc :

$$\begin{aligned} ax^2 + 2bxy + cy^2 &= \big(\, L(\overrightarrow{OM}) \mid \overrightarrow{OM} \,\big) \\ &= \big(\, \lambda X\vec{u} + \mu Y\vec{v} \mid X\vec{u} + Y\vec{v} \,\big) \\ &= \lambda X^2 + \mu Y^2. \end{aligned}$$

Pour la dernière égalité, on a utilisé $\big(\, \vec{u} \mid \vec{v} \,\big) = 0$, $\big(\, \vec{u} \mid \vec{u} \,\big) = 1$ et $\big(\, \vec{v} \mid \vec{v} \,\big) = 1$.
Soient (α, β) les coordonnées du vecteur $\vec{w} := (d, f)$ dans la base $\mathcal{B}$:

$$\vec{w} = (d, f) = \alpha\vec{u} + \beta\vec{v}.$$

Alors :

$$dx + fy = \big(\, \overrightarrow{OM} \mid \vec{w} \,\big) = \big(\, X\vec{u} + Y\vec{v} \mid \alpha\vec{u} + \beta\vec{v} \,\big) = \alpha X + \beta Y.$$

Par conséquent, on a :

$$M = O + x\vec{\imath} + y\vec{\jmath} \quad \text{avec} \quad ax^2 + 2bxy + cy^2 + dx + fy + g = 0$$

si, et seulement si :

$$M = O + X\vec{u} + Y\vec{v} \quad \text{avec} \quad \lambda X^2 + \mu Y^2 + \alpha X + \beta Y + g = 0.$$

4.3.1 Cas parabolique : $\lambda\mu = 0$

Nous allons montrer que dans ce cas, la courbe $\mathcal{C}$ est :

- soit l'ensemble vide,
- soit une droite,
- soit la réunion de deux droites parallèles,
- soit une parabole.

Comme $\lambda\mu = 0$ avec $(\lambda, \mu) \neq (0, 0)$, on peut supposer que $\lambda \neq 0$ et $\mu = 0$. Quitte à diviser tous les coefficients par λ, on peut supposer que $\lambda = 1$. L'équation est alors de la forme $X^2 + \alpha X + \beta Y + g = 0$, que l'on peut également écrire sous la forme :

$$\left(X + \frac{\alpha}{2}\right)^2 + \beta Y = \gamma \quad \text{avec} \quad \gamma = \frac{\alpha^2}{4} - g.$$

- Si $\beta = 0$, l'équation prend la forme $\left(X + \frac{\alpha}{2}\right)^2 = \gamma$.
  - Si $\gamma < 0$, l'ensemble des solutions est l'ensemble vide.
  - Si $\gamma = 0$, l'ensemble des solutions est la droite d'équation :
  $$X + \alpha/2 = 0.$$
  - Si $\gamma > 0$, l'ensemble des solutions est la réunion des deux droites parallèles d'équations $X + \alpha/2 = \sqrt{\gamma}$ et $X + \alpha/2 = -\sqrt{\gamma}$.
- Si $\beta \neq 0$, l'équation prend la forme :
  $$\left(Y - \frac{\gamma}{\beta}\right) = -\frac{1}{\beta}\left(X + \frac{\alpha}{2}\right)^2$$
  qui est l'équation d'une parabole.

  Le foyer est le point de coordonnées $(X, Y) = \left(-\frac{\alpha}{2}, \frac{\gamma}{\beta} - \frac{\beta}{4}\right)$ et la directrice est la droite d'équation $Y = \frac{\gamma}{\beta} + \frac{\beta}{4}\cdot$

4.3.2 Coniques à centre : $\lambda\mu \neq 0$

Proposition 41. Si $\lambda\mu \neq 0$, et si $O' := O + \frac{\alpha}{2\lambda}\vec{u} + \frac{\beta}{2\mu}\vec{v}$, alors :

$$M = O + X\vec{u} + Y\vec{v} \quad \text{avec} \quad \lambda X^2 + \mu Y^2 + \alpha X + \beta Y + g$$

si, et seulement si :

$$M = O' + X'\vec{u} + Y'\vec{v} \quad \text{avec} \quad \lambda(X')^2 + \mu(Y')^2 = \gamma \quad \text{et} \quad \gamma := \frac{\alpha^2}{4\lambda} + \frac{\beta^2}{4\mu} - g.$$

Remarque. Le point O' est le centre de symétrie de la conique.

Démonstration. On peut écrire l'équation $\lambda X^2 + \mu Y^2 + \alpha X + \beta Y + g = 0$ sous la forme :

$$\lambda\left(X - \frac{\alpha}{2\lambda}\right)^2 + \mu\left(Y - \frac{\beta}{2\mu}\right)^2 = \frac{\alpha^2}{4\lambda} + \frac{\beta^2}{4\mu} - g.$$

On a :

$$O + X\vec{u} + Y\vec{v} = O' + X'\vec{u} + Y'\vec{v} \iff X' = X - \frac{\alpha}{2\lambda} \quad \text{et} \quad Y' = Y - \frac{\beta}{2\mu}\cdot \quad \blacksquare$$

4.3.3 Cas elliptique : $\lambda\mu > 0$

Dans ce cas, la courbe $\mathcal{C}$ est :

- soit l'ensemble vide,
- soit un point,
- soit un cercle,
- soit une ellipse.

Les réels λ et μ sont non nuls et de même signe. Quitte à multiplier tous les coefficients par -1, on peut supposer que λ et μ sont positifs.

- Si $\gamma < 0$, l'ensemble des solutions est l'ensemble vide.
- Si $\gamma = 0$, l'unique solution est le point O'.
- Si $\gamma > 0$, en divisant par γ, et en posant $a' := \sqrt{\gamma/\lambda}$ et $b' := \sqrt{\gamma/\mu}$, l'équation prend la forme :
$$\left(\frac{X'}{a'}\right)^2 + \left(\frac{Y'}{b'}\right)^2 = 1.$$
Si $a' = b'$, c'est l'équation d'un cercle centré en O' de rayon $a' = b'$. Sinon, c'est l'équation d'une ellipse. Si $a' > b'$, le grand axe est la droite d'équation $Y' = 0$ et si $b' > a'$, le grand axe de l'ellipse est la droite d'équation $X' = 0$. Le centre de symétrie de l'ellipse est le point O'.

4.3.4 Cas hyperbolique : $\lambda\mu < 0$

Dans ce cas, la courbe $\mathcal{C}$ est :

- soit la réunion de deux droites sécantes,
- soit une hyperbole.

Les réels λ et μ sont non nuls et de signes opposés.

- Si $\gamma = 0$, l'équation prend la forme :
$$\lambda(X')^2 + \mu(Y')^2 = 0.$$
En divisant par μ et en posant $a' := \sqrt{-\lambda/\mu}$, l'équation devient :
$$(a'X')^2 - (Y')^2 = 0,$$
c'est-à-dire :
$$(Y' - a'X') \cdot (Y' + a'X') = 0.$$
L'ensemble des solutions est donc la réunion des deux droites d'équations $Y' - a'X' = 0$ et $Y' + a'X' = 0$ qui se coupent au point O'.
- Si $\gamma \neq 0$, en divisant par γ, l'équation prend la forme :
$$\frac{\lambda}{\gamma}(X')^2 + \frac{\mu}{\gamma}(Y')^2 = 1.$$
L'ensemble des solutions est une hyperbole. Le centre de symétrie de l'hyperbole est le point O'. Si $\lambda/\gamma > 0$ (et donc $\mu/\gamma < 0$), l'axe transverse de l'ellipse est la droite d'équation $Y' = 0$ et l'axe non transverse est la droite d'équation $X' = 0$. Si $\lambda/\gamma < 0$ (et donc $\mu/\gamma > 0$), l'axe transverse de l'ellipse est la droite d'équation $X' = 0$ et l'axe non transverse est la droite d'équation $Y' = 0$.

Exercice type corrigé

III.1.0 Dans $\mathbb{R}^2$, soit D_1 la droite d'équation $a_1x + b_1y + c_1 = 0$, D_2 la droite d'équation $a_2x + b_2y + c_2 = 0$ et D_3 la droite d'équation $a_3x + b_3y + c_3 = 0$. Montrer que :

$$\det \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} = 0$$

si, et seulement si, les trois droites sont soit concourantes, soit deux à deux parallèles.

Solution. Supposons d'abord que le déterminant soit nul. Alors, les trois équations de droites sont liées. On peut donc trouver $(\lambda_1, \lambda_2, \lambda_3) \neq (0, 0, 0)$ tel que :

$$\lambda_1 \cdot (a_1x + b_1yc_1) + \lambda_2 \cdot (a_2x + b_2y + c_2) + \lambda_3 \cdot (a_3x + b_3y + c_3) = 0.$$

Si l'un des λ_i est nul, cela signifie que deux équations de droites sont proportionnelles et donc, que deux droites sont égales. Par exemple, si $\lambda_1 = 0$, alors $D_2 = D_3$. Puisqu'il n'y a que deux droites, elles sont soit parallèles, soit concourantes (si elles se coupent).

Si aucun λ_i n'est nul et si un point appartient à deux droites, il appartient automatiquement à la troisième. Par exemple, si $M = (x_0, y_0) \in D_1 \cap D_2$, alors $a_1x_0 + b_1y_0 + c_1 = 0$ et $a_2x_0 + b_2y_0 + c_2 = 0$. Par conséquent :

$$a_3x_0 + b_3y_0 + c_3 = -\frac{\lambda_1}{\lambda_3}(a_1x_0 + b_1y_0 + c_1) - \frac{\lambda_2}{\lambda_3} \cdot (a_2x_0 + b_2y_0 + c_2) = 0.$$

Donc, $M \in D_3$. Dans ce cas, les droites sont concourantes. La seule alternative est que deux à deux les droites soient non sécantes, c'est-à-dire parallèles.

Montrons maintenant la réciproque. Si les droites sont deux à deux parallèles, alors, les vecteurs directeurs $(b_1, -a_1)$, $(b_2, -a_2)$ et $(b_3, -a_3)$ sont colinéaires. Par conséquent, les vecteurs (a_1, a_2, a_3) et (b_1, b_2, b_3) sont colinéaires et le déterminant est nul. Si les droites sont concourantes, alors on peut trouver un point $M = (x_0, y_0)$ qui appartient aux trois droites. On a donc :

$$\begin{cases} a_1x_0 + b_1y_0 + c_1 = 0 \\ a_2x_0 + b_2y_0 + c_2 = 0 \\ a_3x_0 + b_3y_0 + c_3 = 0. \end{cases}$$

Cela implique :

$$x_0 \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + y_0 \cdot \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} + 1 \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Donc le déterminant est nul.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

III.1.1 Soit $f : \mathbb{R} \to \mathbb{R}$ une application continue. Montrer que l'ensemble des primitives $F : \mathbb{R} \to \mathbb{R}$ de f est un espace affine. Quelle est sa dimension ?

III.1.2 Soient A et B deux points distincts du plan affine $\mathbb{R}^2$. Pour $k > 0$, déterminer l'ensemble des points $M \in \mathbb{R}^2$ tels que $AM = k \cdot BM$.
Indication : pour $k \neq 1$, on introduira le barycentre G_1 de $\{(A,1),(B,-k)\}$ et le barycentre G_2 de $\{(A,1),(B,k)\}$.

III.1.3 Soient A et B deux points de $\mathbb{R}^3$. Montrer que le segment $[A,B]$ est égal à l'ensemble des barycentres des familles $\{(A,\alpha),(B,\beta)\}$ avec $\alpha \geqslant 0$ et $\beta \geqslant 0$.

III.1.4 Soient A, B, C et D quatre points non coplanaires dans $\mathbb{R}^3$. Soient A_1 le centre de gravité du triangle BCD, B_1 celui du triangle ACD, C_1 celui du triangle ABD et D_1 celui du triangle ABC. Montrer que les quatre médianes AA_1, BB_1, CC_1 et DD_1 sont concourantes.

III.1.5 Montrer que pour tout point $A \in \mathbb{R}^2$ et tout réel α, la rotation de centre A et d'angle α est une application affine.

III.1.6 Soient X et Y deux espaces affines et $f : X \to Y$ une application affine. Soit X' un sous-espace affine de X. Monter que $f(X')$ est un sous-espace affine de Y.

III.1.7 Montrer que :

$$X := \left\{ \begin{pmatrix} 1+2a+b & a-b \\ 0 & a+b \end{pmatrix} \mid (a,b) \in \mathbb{R}^2 \right\}$$

est un sous-espace affine de $M_2(\mathbb{R})$. En déterminer un point et la direction.

III.1.8 Montrer que l'ensemble des fonctions $f \in \mathcal{F}(\mathbb{R},\mathbb{R})$ telles que $f(x) + f(-x) = 1$ pour tout $x \in \mathbb{R}$, est un sous-espace affine de $\mathcal{F}(\mathbb{R},\mathbb{R})$. En déterminer un point et la direction.

Exercices sur la section 2

III.1.9 Montrer que l'ensemble des fonctions $f : \mathbb{R} \to \mathbb{R}$ telles que $f(1) = f(0) + 1$ est un hyperplan affine de $\mathcal{F}(\mathbb{R},\mathbb{R})$.

III.1.10 Montrer que l'ensemble des matrices $n \times n$ dont la somme des coefficients diagonaux est égale à 2 est un hyperplan affine de $M_n(\mathbb{R})$.

Exercices sur la section 3

III.1.11 Dans $\mathbb{R}^2$, déterminer une équation cartésienne de la droite passant par les points :

$$A := (1,2) \qquad \text{et} \qquad B := (-2,1).$$

III.1.12 Dans $\mathbb{R}^3$, déterminer une équation cartésienne du plan contenant les trois points $A := (0,0,1)$, $B := (0,0,2)$ et $C := (1,0,1)$.

III.1.13 Trouver une équation paramétrique de la droite $D \subset \mathbb{R}^2$ d'équation $3x + 2y - 1 = 0$.

III.1.14 Trouver une équation paramétrique de la droite $D \subset \mathbb{R}^3$ définie par les équations :

$$\begin{cases} 3x + y - z - 2 = 0 \\ x + y = 0. \end{cases}$$

III.1.15
Trouver une équation paramétrique du plan $P \subset \mathbb{R}^3$ défini par l'équation $3x + 2y - z = 0$.

III.1.16 Soient $A_1 := (x_1, y_1)$, $A_2 := (x_2, y_2)$ et $A_3 := (x_3, y_3)$ trois points de $\mathbb{R}^2$. Montrer que les points sont alignés si, et seulement si :

$$\det \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix} = 0.$$

III.1.17 Dans $\mathbb{R}^3$, soient P_1, P_2 et P_3 trois plans d'équations respectives :

$$a_1x + b_1y + c_1z + d_1 = 0, \qquad a_2x + b_2y + c_2z + d_2 = 0 \quad \text{et} \quad a_3x + b_3y + c_3z + d_3 = 0.$$

Donner une condition nécessaire et suffisante pour que $P_1 \cap P_2 \cap P_3$ soit un singleton.

III.1.18 Donner une condition nécessaire et suffisante pour que 4 points $A_1 := (x_1, y_1, z_1)$, $A_2 := (x_2, y_2, z_2)$, $A_3 := (x_3, y_3, z_3)$ et $A_4 := (x_4, y_4, z_4)$ soient coplanaires.
Indication : considérer le déterminant de la matrice :

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

III.1.19 Soit D_1 et D_2 deux droites de $\mathbb{R}^3$ non parallèles. Soient F_1 et F_2 les espaces qui dirigent D_1 et D_2.

1. Montrer qu'il existe un sous-espace affine $X \subset \mathbb{R}^3$ contenant D_2 et dirigé par $F_1 + F_2$.
2. On appelle f la projection orthogonale sur X. Montrer que $f(D_1)$ coupe D_2 en un unique point A_2.
3. Montrer qu'il existe un unique point $A_1 \in D_1$ tel que $f(A_1) = A_2$.
4. Montrer que pour tout $B_1 \in D_1$ et tout $B_2 \in D_2$, on a $d(B_1, B_2) \geqslant d(A_1, A_2)$ avec égalité si, et seulement si, $A_1 = B_1$ et $A_2 = B_2$.

La droite A_1A_2 s'appelle la perpendiculaire commune à D_1 et D_2.

III.1.20 Dans $\mathbb{R}^3$, on considère la droite D passant par le point $A := (0, 1, 1)$ et dirigée par le vecteur $\vec{u} := (2, -1, 0)$. Soit $B := (1, 1, -2) \in \mathbb{R}^3$. Déterminer $d(B, D)$.

Exercices sur la section 4

III.1.21 Soit $\mathcal{E}$ une ellipse de foyers F et F'. Montrer qu'il existe une constante c telle que pour tout $M \in \mathcal{E}$, $d(F, M) + d(F', M) = c$.

III.1.22 Soit $\mathcal{H}$ une hyperbole de foyers F et F'. Montrer qu'il existe une constante c telle que pour tout $M \in \mathcal{E}$, $\big|d(F, M) - d(F', M)\big| = c$.

Courbes paramétrées III.2

Ce module est indissociable du module IV.6 sur les applications vectorielles d'une variable réelle. Il en est le pendant géométrique.
Une courbe est la représentation mathématique d'objets tels que des câbles téléphoniques, le cours d'une rivière ou la trajectoire d'une planète. Par exemple, les droites, les segments, les polygones et les cercles sont des courbes.
Il existe essentiellement deux manières de définir une courbe. La première consiste à la définir comme l'ensemble des solutions d'une équation ou d'un système d'équations. C'est ce que nous avons fait dans le module III.1 lorsque nous avons étudié les droites du plan ou de l'espace, ou encore les coniques du plan.
D'un point de vue physique, on est plutôt tenté de définir une courbe comme le lieu des positions successives d'un point qui se meut suivant une loi déterminée, par exemple, la trajectoire d'un point dans le plan ou dans l'espace. On parlera alors de courbe paramétrée.

1 Courbes planes

Nous commencerons par étudier les courbes du plan affine. Un point du plan affine sera repéré par ses coordonnées dans un repère $(O;\vec{\imath},\vec{\jmath})$ (qui sera généralement orthonormé). Nous identifierons donc le plan affine avec $\mathbb{R}^2$. Le point M de coordonnées (x,y) est le point $M = O + x\vec{\imath} + y\vec{\jmath}$.
Dans toute la suite, $I \subset \mathbb{R}$ sera un intervalle contenant au moins deux points.

1.1 Notion de courbe paramétrée

Définition 1. Une *courbe* C de $\mathbb{R}^2$ est l'image d'une application continue $\gamma : I \to \mathbb{R}^2$. Nous dirons que γ est un *paramétrage* de C.

Exemples.

1. Une droite de $\mathbb{R}^2$ est une courbe. Un paramétrage possible de la droite passant par le point (x_0, y_0) et dirigée par le vecteur (u,v) est l'application $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma(t) := (x_0 + tu, y_0 + tv)$.

2. Un cercle de $\mathbb{R}^2$ est une courbe.
   Un paramétrage possible du cercle de centre (x_0, y_0) et de rayon r est l'application $\gamma : [0, 2\pi] \to \mathbb{R}^2$ définie par $\gamma(t) := (x_0 + r\cos t, y_0 + r\sin t)$.
3. Si $I \subset \mathbb{R}$ est un intervalle contenant au moins deux points, le graphe (ou courbe représentative) d'une application continue $f : I \to \mathbb{R}$ est une courbe. Un paramétrage possible de cette courbe est l'application $\gamma : I \to \mathbb{R}^2$ définie par $\gamma(t) := \big(t, f(t)\big)$.

Définition 2. Un *homéomorphisme* entre deux intervalles $I \subset \mathbb{R}$ et $J \subset \mathbb{R}$ est une application $\varphi : I \to J$ continue strictement croissante ou strictement décroissante.

Remarque. Une même courbe C admet une infinité de paramétrages.
En effet, si $\gamma : I \to C$ est un paramétrage de C et si $\varphi : J \to I$ est un homéomorphisme, alors $\gamma \circ \varphi : J \to C$ est un paramétrage de C.

Un problème se pose. Étant donné un ensemble C de points du plan, comment déterminer si C est une courbe ? Par exemple, une ellipse est-elle une courbe ? Le carré $[0, 1] \times [0, 1]$ est-il une courbe ?
Il va falloir manipuler la notion de courbe avec beaucoup de soin. D'abord, avec la définition de courbe que nous venons de donner, un ensemble réduit à un point M est automatiquement une courbe, toute application constante $\gamma : \mathbb{R} \to \{M\}$ en étant un paramétrage. Ensuite, en 1890, Giuseppe Peano a montré qu'il existe une application continue surjective $\gamma : [0, 1] \to [0, 1] \times [0, 1]$. Cela montre que contrairement à ce que notre intuition pourrait nous laisser croire, le carré $[0, 1] \times [0, 1]$ est une courbe.
C'est probablement ce qui a motivé la réflexion suivante de Felix Klein : « tout le monde sait ce qu'est une courbe, jusqu'à ce qu'il ait étudié suffisamment de mathématiques pour devenir déconcerté par la quantité innombrable d'exceptions possibles. »
Afin de ne pas nous laisser piéger par notre intuition, toutes les courbes que nous allons étudier seront données avec un paramétrage bien précis. Nous parlerons alors de courbes paramétrées.
Pour bien faire, une courbe paramétrée devrait être la donnée d'une courbe et d'un paramétrage de la courbe. Comme la donnée du paramétrage suffit à déterminer la courbe, nous utiliserons la définition suivante.

Définition 3. Une *courbe paramétrée* de $\mathbb{R}^2$ est une application continue $\gamma : I \to \mathbb{R}^2$. La courbe $C := \gamma(I)$ est appelée le *support* de γ.

Lorsque nous étudierons une courbe paramétrée $\gamma : I \to \mathbb{R}^2$, nous aurons généralement une vision cinématique du problème. Nous considérerons que le point $\gamma(t)$ est un point du plan qui évolue avec le temps $t \in I$. C'est pourquoi le paramètre sera habituellement noté t (pour temps).
Rappelons qu'une application $\gamma : I \to \mathbb{R}^2$ qui s'écrit en coordonnées $\gamma(t) = \big(x(t), y(t)\big)$ est continue si, et seulement si, les deux fonctions coordonnées $x : I \to \mathbb{R}$ et $y : I \to \mathbb{R}$ sont des fonctions continues (nous renvoyons le lecteur au module IV.6). Par conséquent, deux fonctions continues $x : I \to \mathbb{R}$ et $y : I \to \mathbb{R}$ définissent automatiquement une courbe paramétrée de $\mathbb{R}^2$.

Exemple.
Considérons les fonctions $x : [-2,2] \to \mathbb{R}$ et $y : [-2,2] \to \mathbb{R}$ définies par :

$$x(t) := t + \cos(3t) \quad \text{et} \quad y(t) := \sin(2t).$$

Ces deux fonctions sont continues sur $[-2,2]$. Elles induisent donc une courbe paramétrée $\gamma : [-2,2] \to \mathbb{R}^2$ définie par :

$$\gamma(t) = \big(t + \cos(3t), \sin(2t)\big).$$

FIGURE III.2.1 – Le support de la courbe.

Pour étudier les propriétés géométriques des courbes paramétrées, nous aurons souvent besoin d'imposer des conditions de régularité.

Définition 4. Soit $\gamma : I \to \mathbb{R}^2$ une courbe paramétrée et k un entier positif. La courbe paramétrée est dérivable (respectivement de classe $\mathcal{C}^k$ ou $\mathcal{C}^\infty$) si l'application γ est dérivable (respectivement de classe $\mathcal{C}^k$ ou $\mathcal{C}^\infty$).

1.2 Étude locale

1.2.1 Tangente

Dans tout ce paragraphe, nous étudions le comportement d'une courbe paramétrée $\gamma : I \to \mathbb{R}^2$ lorsque le paramètre t tend vers un paramètre $t_0 \in \overset{\circ}{I}$. Nous supposons que γ n'est pas constante au voisinage de t_0.
Intuitivement, la courbe paramétrée admet une tangente en $\gamma(t_0)$ si les droites passant par $\gamma(t_0)$ et $\gamma(t)$ admettent une limite quand t tend vers t_0. Le problème est de traduire mathématiquement la notion de limite de droite. Nous utiliserons la définition suivante.

Définition 5. Soit $u_0 \in \mathbb{R}^2$ un vecteur non nul. La droite $D = \gamma(t_0) + \mathrm{Vect}(u_0)$ est appelée *tangente* à γ au temps t_0 s'il existe $\lambda : I \to \mathbb{R}$ et $\varepsilon : I \to \mathbb{R}^2$ tels que :

$$\gamma(t) = \gamma(t_0) + \lambda(t)u_0 + \lambda(t)\varepsilon(t) \quad \text{avec} \quad \varepsilon(t) \underset{t \to t_0}{\longrightarrow} (0,0).$$

Remarque. Nous allons voir dans un instant que l'existence d'une tangente est liée à la dérivabilité de la fonction γ. Cependant, il se peut que la courbe paramétrée admette une tangente sans que γ ne soit dérivable.

Exemple.
La courbe paramétrée $\gamma : \mathbb{R} \to \mathbb{R}$ définie par $\gamma(t) := \big(\sqrt[3]{t}, |t|\big)$ admet une tangente horizontale au temps $t = 0$ bien que γ ne soit pas dérivable en 0. En effet :

$$\gamma(t) = \gamma(0) + \sqrt[3]{t}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \sqrt[3]{t}\begin{pmatrix} 0 \\ |t|/\sqrt[3]{(t)} \end{pmatrix}$$

$$\text{et} \quad \begin{pmatrix} 0 \\ |t|/\sqrt[3]{(t)} \end{pmatrix} \underset{t \to 0}{\longrightarrow} \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

FIGURE III.2.2 – La courbe admet une tangente horizontale

En fait, le support de la courbe admet un paramétrage dérivable : si l'on fait le changement de paramètre $u := \sqrt[3]{t}$, on a $\gamma(t) = (u, |u|^3)$ et $u \mapsto (u, |u|^3)$ est dérivable.

Venons-en maintenant au lien entre la dérivabilité de la courbe en t_0 et l'existence d'une tangente au temps t_0. Pour que la courbe représentative $\gamma : t \mapsto \big(t, f(t)\big)$ d'une fonction $f : I \to \mathbb{R}$ ait une tangente en t_0, il suffit que f soit dérivable en t_0. La tangente est alors la droite passant le point $\gamma(t_0) = \big(t_0, f(t_0)\big)$ et dirigée par le vecteur $\gamma'(t_0) = \big(1, f'(t_0)\big)$. La proposition suivante est une généralisation de ce résultat.

Proposition 1. Si γ est dérivable en t_0 et si $\gamma'(t_0) \neq 0$, alors la courbe paramétrée admet une tangente dirigée par le vecteur $\gamma'(t_0)$ au temps t_0.

Démonstration. La dérivabilité de γ en t_0 se traduit par :

$$\gamma(t) = \gamma(t_0) + (t - t_0)\gamma'(t_0) + (t - t_0)\varepsilon(t) \quad \text{avec} \quad \varepsilon(t) \underset{t \to t_0}{\longrightarrow} (0, 0).$$

Revenir à la définition d'une tangente en posant $u_0 = \gamma'(t_0)$ et $\lambda(t) = (t - t_0)$. ■

Ce résultat motive la définition suivante.

Définition 6. Supposons que γ soit dérivable en t_0. La courbe γ a un *point régulier* au temps t_0 si $\gamma'(t_0) \neq (0, 0)$. Sinon, γ a un *point stationnaire* au temps t_0.

Remarque. Si γ est dérivable en t_0, $\gamma'(t_0)$ est le *vecteur vitesse* au temps t_0. En un point stationnaire, la vitesse est nulle.

Exercice 1.
Quels sont les points réguliers de la courbe paramétrée $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma(t) := (t^2, t^3)$?
Solution. Aux temps $t \neq 0$, la courbe a des points réguliers. Au temps $t = 0$, la courbe passe par l'origine $(0, 0)$ et a un point stationnaire.

FIGURE III.2.3 – Le support d'une courbe admettant un point stationnaire

---

Définition 7. Une courbe régulière est une courbe paramétrée $\gamma : I \to \mathbb{R}^2$ dérivable telle que pour tout $t \in \overset{\circ}{I}$, $\gamma'(t) \neq (0, 0)$.

Remarque. Tous les points d'une courbe régulière sont des points réguliers et en un point régulier, une courbe dérivable admet une tangente dirigée par le vecteur vitesse $\gamma'(t_0)$.

Nous allons maintenant voir que même dans le cas d'un point stationnaire, il y a généralement une tangente.

Proposition 2.
Supposons qu'il existe un entier $p \geqslant 1$ et un vecteur $u_0 \neq (0,0)$ tels que :

$$\gamma(t) = \gamma(t_0) + (t-t_0)^p u_0 + (t-t_0)^p \varepsilon(t) \quad \text{avec} \quad \varepsilon(t) \underset{t \to t_0}{\longrightarrow} (0,0).$$

Alors, γ a une tangente dirigée par le vecteur u_0 au temps t_0.

Démonstration. Revenir à la définition d'une tangente en posant $\lambda(t) := (t-t_0)^p$. ■

Exercice 2.
La courbe de l'exercice 1 de la page précédente admet-elle une tangente au temps $t = 0$?
Solution. La courbe γ admet une tangente horizontale au temps $t = 0$. En effet :

$$\gamma(t) = (0,0) + t^2(1,0) + t^2(0,t)$$

et nous sommes dans la situation de la proposition 2 avec $u_0 = (1,0)$ et $\varepsilon(t) = (0,t)$.

Le résultat suivant est un corollaire de la proposition 2.

Corollaire 3. Si γ est de classe $\mathcal{C}^\infty$ et si les dérivées $\gamma^{(p)}(t_0)$ ne sont pas toutes nulles, alors, γ admet une tangente au temps t_0.

Démonstration. Soit $p \geqslant 1$ le plus petit entier tel que $\gamma^{(p)}(t_0) \neq (0,0)$. D'après la formule de Taylor-Young pour les fonctions vectorielles (voir module IV.8) :

$$\gamma(t) = \gamma(t_0) + \frac{(t-t_0)^p}{p!}\gamma^{(p)}(t_0) + (t-t_0)^p \varepsilon(t) \quad \text{avec} \quad \varepsilon(t) \underset{t \to t_0}{\longrightarrow} (0,0).$$ ■

L'essentiel des courbes paramétrées que nous considérerons seront des courbes de classe $\mathcal{C}^\infty$ auxquelles nous pourrons appliquer le corollaire 3. Elles admettront donc des tangentes en tout point. Cependant, il existe des courbes n'admettant de tangentes nulle part. Le premier mathématicien à avoir démontré l'existence de telles courbes est Karl Weierstrass. La plus célèbre de ces courbes est probablement la courbe introduite en 1906 par le mathématicien suédois Niels von Koch.

FIGURE III.2.4 – La courbe de von Koch n'admet de tangente en aucun point.

1.2.2 Allure de la courbe

Supposons maintenant que γ admette une tangente dirigée par le vecteur $u_0 \neq (0,0)$ au temps t_0. Nous souhaitons étudier l'allure de la courbe au voisinage de t_0.

Définition 8. Soient $\gamma : I \to \mathbb{R}^2$ une courbe paramétrée et D une droite qui passe par $\gamma(t_0)$ avec $t_0 \in \overset{\circ}{I}$. La courbe

- *bute sur la droite* D au temps t_0 si pour $t \neq t_0$ suffisamment proche de t_0, les points $\gamma(t)$ sont tous du même côté de D ;
- *traverse la droite* D au temps t_0 si pour $t > t_0$ suffisamment proche de t_0, les points $\gamma(t)$ se situent tous du même côté de D, et pour $t < t_0$ suffisamment proche de t_0, les points $\gamma(t)$ se situent tous de l'autre côté de D.

Remarque. Le fait de buter ou de traverser la droite D au temps t_0 ne donne qu'une description locale de la courbe. Cela ne présume en rien de la position globale de la courbe par rapport à la droite. Considérons par exemple une fonction dérivable $f : \mathbb{R} \to \mathbb{R}$ qui admet un maximum local en 0 et un maximum global $f(1) > f(0)$ en 1. Alors, la courbe représentative $\gamma : t \mapsto \big(t, f(t)\big)$ admet une tangente horizontale D au temps $t_0 = 0$. La courbe γ bute sur D au temps 0 : pour $t \neq 0$ suffisamment proche de 0, $\gamma(t)$ se trouve en dessous de sa tangente D. Cependant, pour $t = 1$, la courbe se trouve au dessus de D. De même, une courbe qui traverse une droite au temps t_0 peut retraverser à d'autres instants.

Exemples.

1. La courbe $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma(t) := (t, t^2)$ bute sur l'axe des abscisses au temps $t = 0$ et traverse toutes les autres droites passant par l'origine.
2. La définition que nous venons de donner ne couvre pas tous les cas de figure que l'on peut rencontrer. Considérons la courbe $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma(0) := (0,0)$ et pour $t \neq 0$, $\gamma(t) := \big(t, t\sin(\pi/t)\big)$. Au temps $t = 0$, cette courbe ne traverse pas l'axe des abscisses et ne bute pas non plus dessus.

FIGURE III.2.5 –

Définition 9. Soient $\gamma : I \to \mathbb{R}^2$ une courbe paramétrée ayant une tangente D au temps $t_0 \in \mathring{I}$. La courbe paramétrée a

- un *point ordinaire* au temps t_0 si au temps t_0, la courbe bute sur la tangente et traverse toutes les autres droites qui passent par $\gamma(t_0)$;
- un *point d'inflexion* au temps t_0 si au temps t_0, la courbe traverse toutes les droites qui passent par $\gamma(t_0)$ (y compris la tangente);
- un *point de rebroussement de première espèce* au temps t_0 si au temps t_0, la courbe traverse la tangente et bute sur toutes les autres droites qui passent par $\gamma(t_0)$;
- un *point de rebroussement de seconde espèce* au temps t_0 si au temps t_0, la courbe bute sur toutes les droites qui passent par $\gamma(t_0)$ (y compris la tangente).

Remarque. Comme précédemment, les notions que nous venons d'introduire ne concernent que le comportement local de la courbe au voisinage du temps t_0. Par exemple, une courbe qui a un point ordinaire au temps t_0 peut traverser sa tangente en un temps $t_1 \neq t_0$.

FIGURE III.2.6 –

Exemples.

1. La courbe paramétrée $\gamma_1 : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma_1(t) := (t, t^2)$ a des points ordinaires en tout temps $t \in \mathbb{R}$.
2. La courbe paramétrée $\gamma_2 : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma_2(t) := (t, t^3)$ a un point d'inflexion au temps $t = 0$. Les autres points sont des points ordinaires.
3. La courbe paramétrée $\gamma_3 : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma_3(t) := (t^2, t^3)$ a un point de rebroussement de première espèce au temps $t = 0$. Les autres points sont des points ordinaires.

4. La courbe paramétrée $\gamma_4 : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma_4(t) := (t^2, t^4 + t^5/2)$ a un point de rebroussement de seconde espèce au temps $t = 0$. Les autres points sont des points ordinaires.

Les quatre cas sont illustrés sur la figure III.2.6 de la page précédente.

Remarque. La définition que nous venons de donner ne couvre pas tous les cas de figure que l'on peut rencontrer. Par exemple, au temps $t = 0$, elle ne couvre pas le cas de la courbe $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma(0) := (0,0)$ et $\gamma(t) := \big(t, t^2 \sin(1/t)\big)$ pour $t \neq 0$. Cependant, elle couvre la majorité des cas que nous rencontrerons.

Proposition 4. Supposons qu'il existe deux entiers $q > p \geqslant 2$ et deux vecteurs non colinéaires u_0 et v_0 tels que :

$$\gamma(t) = \gamma(t_0) + (t - t_0)^p \lambda(t) u_0 + (t - t_0)^q v_0 + (t - t_0)^q \varepsilon(t)$$
$$\text{avec} \quad \lambda(t) \underset{t \to t_0}{\longrightarrow} 1 \qquad \text{et} \qquad \varepsilon(t) \underset{t \to t_0}{\longrightarrow} (0,0).$$

- Si p est impair et q est pair, la courbe a un point ordinaire au temps t_0 ;
- si p est impair et q est impair, la courbe a un point d'inflexion au temps t_0 ;
- si p est pair et q est impair, la courbe a un point de rebroussement de première espèce au temps t_0 ;
- si p est pair et q est pair, la courbe a un point de rebroussement de seconde espèce au temps t_0.

Démonstration. D'abord, nous avons :

$$\gamma(t) = \gamma(t_0) + (t - t_0)^p \big(u_0 + \varepsilon_1(t)\big) \quad \text{avec} \quad \varepsilon_1(t) \underset{t \to t_0}{\longrightarrow} (0,0).$$

D'après la proposition 2 de la page 473, la courbe admet une tangente D dirigée par le vecteur u_0 au temps t_0. Soit $D' \neq D$ une droite passant par $\gamma(t_0)$. Le vecteur u_0 ne dirige pas la droite D'.

Si p est pair, $(t - t_0)^p \lambda(t)$ est positif pour $t \neq t_0$ suffisamment proche de t_0, le point $\gamma(t)$ est du même côté de la droite D' que le point $\gamma(t_0) + u_0$. Donc, la courbe bute sur la droite D'.

Si p est impair, $(t - t_0)^p \lambda(t)$ change de signe avec $t - t_0$. Pour $t > t_0$, le point $\gamma(t)$ est du même côté de D' que le point $\gamma(t_0) + u_0$ et pour $t < t_0$, le point $\gamma(t)$ est de l'autre côté de la droite D'. La courbe traverse donc la droite D'.

Ensuite :

$$\gamma(t) = \gamma(t_0) + u(t) + v(t)$$

où $u(t) = (t - t_0)^p \lambda(t)\, u_0$ et $v(t) = (t - t_0)^q \big(v_0 + \varepsilon(t)\big)$ avec $\varepsilon(t) \underset{t \to t_0}{\longrightarrow} (0,0)$. Le vecteur $u(t)$ est colinéaire à u_0. Le point $\gamma(t_0) + u(t)$ appartient donc à la tangente D.

Si q est pair, $(t - t_0)^q$ est positif et pour $t \neq t_0$ suffisamment proche de t_0, le point $\gamma(t)$ est du même côté de D que le point $\gamma(t_0) + v_0$. La courbe bute donc sur sa tangente.

Si q est impair, $(t-t_0)^q$ change de signe avec $t-t_0$. Pour $t > t_0$, le point $\gamma(t)$ est du même côté de D que le point $\gamma(t_0)+v_0$ et pour $t < t_0$, le point $\gamma(t)$ est de l'autre côté de D. La courbe traverse donc sa tangente. ■

Définition 10. La courbe a un *point birégulier* au temps t_0 si γ est deux fois dérivable en t_0 et si $\gamma'(t_0)$ et $\gamma''(t_0)$ ne sont pas colinéaires, c'est-à-dire si $x_0'y_0''-x_0''y_0' \neq 0$ avec $(x_0', y_0') := \gamma'(t_0)$ et $(x_0'', y_0'') := \gamma''(t_0)$.

Corollaire 5. Tout point birégulier est ordinaire.

Démonstration. D'après la formule de Taylor-Young (voir module IV.8) :

$$\gamma(t) = \gamma(t_0) + (t-t_0)\gamma'(t_0) + \frac{(t-t_0)^2}{2}\gamma''(t_0) + (t-t_0)^2\varepsilon(t)$$

avec $\varepsilon(t) \underset{t\to t_0}{\longrightarrow} (0,0)$.

Appliquer alors la proposition 4 avec $u_0 = \gamma'(t_0)$, $v_0 = \gamma''(t_0)/2$, $p=1$ et $q=2$. ■

Si la courbe paramétrée γ est de classe $\mathcal{C}^\infty$, nous pouvons obtenir le développement limité de la proposition 4 de la page précédente grâce à la formule de Taylor-Young. Il en résulte les deux corollaires suivants.

Corollaire 6. Supposons que γ soit de classe $\mathcal{C}^\infty$ et ait un point régulier au temps t_0. Soit $q \geqslant 2$ le plus petit entier tel que la $q^{ème}$ dérivée $\gamma^{(q)}(t_0)$ ne soit pas colinéaire à $\gamma'(t_0)$ (on suppose qu'il existe un tel entier). Alors,

- si q est pair, la courbe paramétrée a un point ordinaire au temps t_0 ;
- si q est impair, la courbe paramétrée a un point d'inflexion au temps t_0.

Corollaire 7. Supposons que γ soit de classe $\mathcal{C}^\infty$. Soient $q > p \geqslant 1$ les deux plus petits entiers tel que $\gamma^{(p)}(t_0)$ et $\gamma^{(q)}(t_0)$ ne soient pas colinéaires (on suppose l'existence de deux tels entiers). Alors,

- Si p est impair et q est pair, la courbe a un point ordinaire au temps t_0 ;
- si p est impair et q est impair, la courbe a un point d'inflexion au temps t_0 ;
- si p est pair et q est impair, la courbe a un point de rebroussement de première espèce au temps t_0 ;
- si p est pair et q est pair, la courbe a un point de rebroussement de seconde espèce au temps t_0.

1.2.3 Comportement au bord de l'intervalle

Dans ce paragraphe, $\gamma : I \to \mathbb{R}^2$ est une courbe paramétrée et :

$$a \in \mathbb{R} \cup \{-\infty, +\infty\}$$

est une borne de l'intervalle I (il se peut que I soit ouvert en a).

Définition 11. Si $\gamma(t)$ tend vers $\gamma_0 \in \mathbb{R}^2$ quand t tend vers a, nous dirons que γ admet γ_0 comme *point d'arrêt* quand t tend vers a.
Si de plus, il existe un vecteur $u_0 \neq (0,0)$ et une fonction $\lambda : I \to \mathbb{R}$ tels que :

$$\gamma(t) = \gamma_0 + \lambda(t)u_0 + \lambda(t)\varepsilon(t) \quad \text{avec} \quad \varepsilon(t) \underset{t\to t_0}{\longrightarrow} (0,0),$$

nous dirons que la droite $D = \gamma_0 + \mathrm{Vect}(u_0)$ est tangente à γ quand t tend vers a.

Exercice 3.
Étudier les points d'arrêt de la courbe $\gamma : [0,+\infty[\to \mathbb{R}^2$ définie par :

$$\gamma(t) = \left(\frac{1-t^2}{1+t^2}, \frac{1}{1+t}\right).$$

Solution. Elle admet $(1,1)$ comme point d'arrêt quand $t \to 0$ et $(-1,0)$ comme point d'arrêt quand $t \to +\infty$. La courbe admet des tangentes verticales quand $t \to 0$ et quand $t \to +\infty$.

FIGURE III.2.7 – La courbe γ admet des tangentes verticales

Définition 12.
La courbe γ a une *branche infinie* quand t tend vers a si $\left\|\overrightarrow{O\gamma(t)}\right\| \underset{t\to a}{\longrightarrow} +\infty$.

Définition 13. Soit $u_0 \neq (0,0)$ un vecteur de $\mathbb{R}^2$. Quand t tend vers a, γ a une *direction asymptotique* dirigée par u_0 si :

$$\gamma(t) = \lambda(t)u_0 + \lambda(t)\varepsilon(t) \quad \text{avec} \quad |\lambda(t)| \underset{t\to a}{\longrightarrow} +\infty \quad \text{et} \quad \varepsilon(t) \underset{t\to t_0}{\longrightarrow} (0,0).$$

Définition 14. La courbe γ admet la droite $D = \gamma_0 + \mathrm{Vect}(u_0)$ pour *asymptote* quand t tend vers a si :

$$\gamma(t) = \gamma_0 + \lambda(t) \cdot u_0 + \varepsilon(t) \quad \text{avec} \quad |\lambda(t)| \underset{t\to a}{\longrightarrow} +\infty \quad \text{et} \quad \varepsilon(t) \underset{t\to a}{\longrightarrow} (0,0).$$

Remarque. Si quand t tend vers a, γ a une asymptote dirigée par u_0, elle a une direction asymptotique dirigée par u_0. Si γ a une direction asymptotique dirigée par u_0, elle a une branche infinie.

Exemples.

1. La courbe paramétrée $\gamma : [0, +\infty[\to \mathbb{R}^2$ définie par :
$$\gamma(t) := (t\cos t, t\sin t)$$
admet une branche infinie quand $t \to +\infty$ mais pas de direction asymptotique (voir figure III.2.8).
2. La courbe paramétrée $\gamma :]0, +\infty[\to \mathbb{R}^2$ définie par $\gamma(t) := (t, 1/t)$ admet l'axe des ordonnées pour asymptote quand $t \to 0$ et l'axe des abscisses pour asymptote quand $t \to +\infty$.

FIGURE III.2.8 –

1.3 Deux exemples

Dans ce paragraphe, nous allons présenter le plan d'étude d'une courbe paramétrée sur deux exemples. En général, on dispose de deux fonctions $x(t)$ et $y(t)$ et on souhaite tracer l'ensemble des points $M(t)$ de coordonnées $\big(x(t), y(t)\big)$. Les étapes seront essentiellement les suivantes.

1. Commencer par rechercher l'ensemble de définition commun aux fonctions x et y.
2. Essayer de réduire l'ensemble d'étude en utilisant les périodicités de x et de y, ainsi que leurs parités ou leurs symétries.
3. Dresser le tableau de variation des fonctions x et y.
4. Rechercher les points stationnaires et étudier l'allure de la courbe au voisinage de ces points.
5. Rechercher et étudier les points d'arrêt et les branches infinies.
6. Tracer la courbe.

Exercice 4.
Étudier et tracer la partie de $\mathbb{R}^2$ paramétrée par :
$$x = \cos(3t) \quad \text{et} \quad y = \sin(2t).$$

Solution.
Ensemble d'étude
Les fonctions $x : t \mapsto \cos(3t)$ et $y : t \mapsto \sin(2t)$ sont définies et continues sur $\mathbb{R}$. La fonction x est périodique de période $2\pi/3$ et la fonction y est périodique de période π. La

plus petite période commune est 2π. Il nous suffit donc de tracer le support de la courbe $\gamma : I \to \mathbb{R}^2$ définie par $\gamma(t) := \big(\cos(3t), \sin(2t)\big)$, où I est un intervalle de longueur 2π. La fonction x est paire : $x(-t) = x(t)$. La fonction y est impaire : $y(-t) = -y(t)$. On en déduit que le point de coordonnées (x, y) appartient au support de la courbe si, et seulement si, le point de coordonnées $(x, -y)$ appartient au support de la courbe. Nous pouvons donc restreindre l'ensemble d'étude à l'intervalle $[0, \pi]$. Nous tracerons le support de la restriction de la courbe à l'intervalle $[0, \pi]$, le reste du support s'obtenant par la symétrie $(x, y) \mapsto (x, -y)$, *i.e.* la symétrie par rapport à l'axe des abscisses.
Notons enfin que $x(\pi - t) = -x(t)$ et $y(\pi - t) = -y(t)$. Par conséquent, il nous suffit d'étudier la restriction de la courbe à l'intervalle $[0, \pi/2]$. Le support de la restriction à l'intervalle $[\pi/2, \pi]$ s'obtient alors par la transformation $(x, y) \mapsto (-x, -y)$, *i.e.* la symétrie centrale par rapport à $(0, 0)$. On obtient alors le reste du support en faisant la symétrie par rapport à l'axe des abscisses.

Tableau de variations

Dressons maintenant le tableau de variations des fonctions x et y sur l'intervalle $[0, \pi/2]$.

t	0		$\pi/4$		$\pi/3$		$\pi/2$
$x'(t) = -3\sin(3t)$	0	$-$		$-$	0	$+$	3
$x(t) = \cos(3t)$	1	$\searrow$	$-\sqrt{2}/2$	$\searrow$	-1	$\nearrow$	0
$y'(t) = 2\cos(2t)$	2	$+$	0	$-$		$-$	2
$y(t) = \sin(2t)$	0	$\nearrow$	1	$\searrow$	$\sqrt{3}/2$	$\searrow$	0

Points stationnaires

Les points sont tous des points réguliers puisque les dérivées de x et de y ne s'annulent jamais simultanément. Aux temps où la dérivée de x s'annule, il y a une tangente verticale, *i.e.* aux temps $t = 0$ avec $\gamma(0) = (1, 0)$ et au temps $t = \pi/3$ avec $\gamma(\pi/3) = (-1, \sqrt{3}/2)$. Au temps où la dérivée de y s'annule, il y a une tangente horizontale, *i.e.* au temps $t = \pi/4$ avec $\gamma(\pi/4) = (-\sqrt{2}/2, 1)$.

Branches infinies

Les deux fonctions x et y sont bornées. Il n'y a donc pas de branche infinie.

Tracé de la courbe

On trace la restriction de la courbe à $t \in [0, \pi/2]$, puis sa symétrique par rapport à $(0, 0)$, puis sa symétrique par rapport à l'axe des abscisses (voir la figure III.2.9 de la page ci-contre).

FIGURE III.2.9 –

Exercice 5.

Étudier et tracer la courbe paramétrée par :

$$x(t) = \frac{t^2}{t-1} \quad \text{et} \quad y(t) = \frac{(t-2)^2}{t^2-1}\cdot$$

Solution.

Ensemble d'étude. L'ensemble de définition de la fonction x est $\mathbb{R}\setminus\{1\}$ et celui de la fonction y est $\mathbb{R}\setminus\{-1,1\}$. Nous sommes donc amenés à étudier non pas une courbe paramétrée, mais trois courbes paramétrées selon que $t \in \,]-\infty,-1[$, $t \in \,]-1,1[$ ou $t \in \,]1,+\infty[$. Les fonctions n'admettent pas de symétrie particulière, et nous ne pourrons pas réduire l'ensemble d'étude.

Tableau de variations. Le lecteur calculera :

$$x'(t) = \frac{t(t-2)}{(t-1)^2} \quad \text{et} \quad y'(t) = \frac{4(t-2)(t-1/2)}{(t^2-1)^2}\cdot$$

t	$-\infty$		-1		0		$1/2$		1		2		$+\infty$
$x'(t)$		$+$		$+$	0	$-$		$-$	‖	$-$	0	$+$	
$x(t)$	$-\infty$	↗	$-1/2$	↗	0	↘	$-1/2$	↘	$-\infty$ ‖ $+\infty$	↘	4	↗	$+\infty$
$y'(t)$		$+$	‖	$+$		$+$	0	$-$	‖	$-$	0	$+$	
$y(t)$	1	↗	$+\infty$ ‖ $-\infty$	↗	-4	↗	-3	↘	$-\infty$ ‖ $+\infty$	↘	0	↗	1

Points stationnaires

Le seul point où $x'(t)$ et $y'(t)$ s'annulent simultanément est $t = 2$. Il s'agit d'un point stationnaire. Pour étudier la nature de ce point stationnaire, on va faire un développement de Taylor :

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} 4 \\ 0 \end{pmatrix} + (t-2)^2 \begin{pmatrix} 1 \\ 1/3 \end{pmatrix} + (t-2)^3 \begin{pmatrix} -1 \\ -4/9 \end{pmatrix} + (t-2)^3\varepsilon(t)$$

avec $\varepsilon(t) \to (0,0)$ quand $t \to 2$. La courbe admet donc une tangente au temps $t = 2$: la

droite passant par le point $(4,0)$ et dirigée par le vecteur $(1,1/3)$. Le vecteur $(-1,-4/9)$ n'est pas colinéaire au vecteur $(1,1/3)$. On peut donc appliquer la proposition 4 de la page 476 avec $p=2$ et $q=3$. La courbe a un point de rebroussement de première espèce au temps $t=2$.

Branches infinies

Il y a quatre branches infinies : quand $t \to -\infty$, $t \to -1$, $t \to 1$ et $t \to +\infty$. Comme $y(t) \underset{t\to\pm\infty}{\longrightarrow} 1$, la droite horizontale d'équation $y=1$ est asymptote à la courbe lorsque $t \to \pm\infty$. Comme $x(t) \underset{t\to -1}{\longrightarrow} -1/2$, la droite verticale d'équation $x=-1/2$ est donc asymptote à la courbe.

Il ne nous reste qu'à étudier la branche infinie lorsque $t \to 1$. Nous pouvons faire un développement asymptotique lorsque $t \to 1$:

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \frac{1}{t-1}\begin{pmatrix} 1 \\ 1/2 \end{pmatrix} + \begin{pmatrix} 2 \\ -5/4 \end{pmatrix} + \varepsilon(t)$$

avec $\varepsilon(t) \to (0,0)$ quand $t \to 1$. On voit donc que la droite passant par le point $(2,-5/4)$ et dirigée par le vecteur $(1,1/2)$ est asymptote à la courbe quand $t \to 1$.

Tracé de la courbe

FIGURE III.2.10 –

Dans certains cas, nous pourrons être amenés à étudier d'autre points remarquables de la courbe : les points multiples.

Définition 15. Soit $\gamma : I \to \mathbb{R}^2$ une courbe paramétrée. Le point $M \in \gamma(I)$ est un *point multiple* s'il existe $t_1 \in I$ et $t_2 \in I$ avec $t_1 \neq t_2$ et $\gamma(t_1) = \gamma(t_2)$.

Exercice 6.

Montrer que la courbe $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma(t) = (t^2 - 1, t^3 - t)$ a un point multiple.
Solution. Le point $(0,0)$ est un point multiple de la courbe γ, car :

$$\gamma(1) = \gamma(-1) = (0,0).$$

Attention, la définition que nous donnons de point multiple dépend du paramétrage de la courbe, pas seulement de son support.

FIGURE III.2.11 –

Exemple. On peut paramétrer le cercle C de centre $(0,0)$ et de rayon 1 des deux manières suivantes :

- le premier paramétrage de C est donné par l'application $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma(t) = (\cos t, \sin t)$;
- le second paramétrage de C est donné par la restriction de γ à l'intervalle $[0, 2\pi[$.

Dans les deux cas, le support de la courbe est le cercle C. Dans le second cas, le paramétrage est injectif et il n'y a aucun point multiple. Dans le premier cas, on a $\gamma(t + 2\pi) = \gamma(t)$ et tous les points de C sont des points multiples.

2 Courbes en coordonnées polaires

2.1 Définition

Un point du plan affine est généralement repéré par ses coordonnées cartésiennes : son abscisse x et son ordonnée y. Dans un repère orthonormé, il peut également être repéré par des couples (ρ, θ) tels que $x = \rho\cos\theta$ et $y = \rho\sin\theta$. Nous dirons alors que ρ et θ sont des coordonnées polaires du point. Attention, il se peut que ρ soit négatif.
Tout point du plan admet une infinité de coordonnées polaires. Par exemple, le point de coordonnées cartésiennes $(0,1)$ peut être repéré par les coordonnées polaires $\rho = 1$ et $\theta = \pi/2$. Il peut également être repéré par les coordonnées polaires $\rho = -1$ et $\theta = -\pi/2$, ou bien par $\rho = -1$ et $\theta = 3\pi/2$. Plus généralement, un même point peut avoir pour coordonnées polaires les couples (ρ, θ), $(\rho, \theta + 2\pi)$ et $(-\rho, \theta + \pi)$. L'origine O, de coordonnées cartésiennes $(0,0)$, joue un rôle particulier. Les coordonnées polaires de ce point sont de la forme $\rho = 0$ et $\theta \in \mathbb{R}$ quelconque.
La valeur absolue de ρ représente la distance du point à l'origine O du repère. Si ρ est positif, θ est égal, modulo 2π, à l'argument principal du point (x, y). Si ρ est négatif, θ diffère de cet argument principal de π modulo 2π.

Définition 16. Pour $\theta \in \mathbb{R}$, les vecteurs u_θ et v_θ sont les vecteurs :

$$u_\theta = (\cos\theta, \sin\theta) \quad \text{et} \quad v_\theta = (\sin\theta, -\cos\theta) = u_{\theta+\pi/2}.$$

Le repère orthonormé $(O; u_\theta, v_\theta)$ est dit *repère mobile*.

Le point M de coordonnée polaires (ρ, θ) est le point $M = O + \rho(\theta)u_\theta$.
Les fonctions $\theta \mapsto u_\theta$ et $\theta \mapsto v_\theta$ sont des fonctions de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$ et :

$$\frac{\mathrm{d}}{\mathrm{d}\theta}(u_\theta) = v_\theta \quad \text{et} \quad \frac{\mathrm{d}}{\mathrm{d}\theta}(v_\theta) = -u_\theta.$$

Définition 17. Une courbe paramétrée $\gamma : I \to \mathbb{R}^2$ est définie en *coordonnées polaires* par la fonction $\rho : I \to \mathbb{R}$ si ρ est une fonction continue et :

$$\gamma(\theta) = O + \rho(\theta)\, u_\theta = \big(\rho(\theta)\cos\theta, \rho(\theta)\sin\theta\big).$$

La fonction ρ peut prendre des valeurs positives ou des valeurs négatives. Le réel $|\rho(\theta)|$ est la distance de l'origine à $\gamma(\theta)$. Le point $\gamma(\theta)$ se trouve sur la droite passant par l'origine et dirigée par le vecteur u_θ. Si $\rho(\theta) > 0$, $\gamma(\theta)$ se trouve dans la même direction que le vecteur u_θ et si $\rho(\theta) < 0$, $\gamma(\theta)$ se trouve dans la direction opposée.

2.2 Tangente

Proposition 8. Soit $\rho : I \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ et $\gamma : I \to \mathbb{R}^2$ la courbe paramétrée définie par ρ. Alors, γ est de classe $\mathcal{C}^1$. En tout temps $\theta \in \overset{\circ}{I}$ tel que $\rho(\theta) \neq 0$, la courbe γ a un point régulier et admet donc une tangente.

Démonstration.
Si ρ est de classe $\mathcal{C}^1$, les fonctions $x : \theta \mapsto \rho(\theta)\cos\theta$ et $y : \theta \mapsto \rho(\theta)\sin\theta$ sont également de classe $\mathcal{C}^1$. En dérivant $\gamma(\theta) = O + \rho(\theta) \cdot u_\theta$ par rapport à θ, on obtient :

$$\gamma'(\theta) = \rho'(\theta) \cdot u_\theta + \rho(\theta) \cdot v_\theta.$$

Le vecteur vitesse $\gamma'(\theta)$ est donc nul si, et seulement si, $\rho(\theta) = \rho'(\theta) = 0$. ■

Dans la proposition précédente, nous avons requis que ρ soit une fonction de classe $\mathcal{C}^1$ tandis que dans la proposition suivante, nous ne la supposons que continue.

Proposition 9. Soit $\rho : I \to \mathbb{R}$ une fonction continue et $\gamma : I \to \mathbb{R}^2$ la courbe paramétrée définie par ρ. Supposons que $\theta_0 \in \overset{\circ}{I}$ est tel que $\rho(\theta_0) = 0$ et $\rho(\theta) \neq 0$ pour $\theta \neq \theta_0$ suffisamment proche de θ_0. Alors,

1. γ admet une tangente au temps θ_0 dirigée par le vecteur u_{θ_0} ;
2. si ρ change de signe avec $\theta - \theta_0$, γ a un point ordinaire au temps θ_0 ;
3. si ρ ne change pas de signe au voisinage de θ_0, γ a un point de rebroussement de première espèce au temps θ_0.

Démonstration. Observons tout d'abord que :

$$\gamma(\theta) = \gamma(\theta_0) + \rho(\theta) \cdot u_\theta = \gamma(\theta_0) + \rho(\theta) \cdot u_{\theta_0} + \rho(\theta)\varepsilon(\theta) \quad \text{avec} \quad \varepsilon(\theta) \underset{\theta \to \theta_0}{\longrightarrow} (0,0).$$

Cela démontre le point 1.

Ensuite, si $\rho(\theta) > 0$, l'argument principal de $\gamma(\theta)$ est égal à θ modulo 2π et si $\rho(\theta) < 0$, l'argument principal de $\gamma(\theta)$ est égal à $\theta+\pi$ modulo 2π. Si ρ change de signe avec $\theta-\theta_0$, la courbe doit traverser toutes les droites passant par $(0,0)$ sauf la tangente et γ a un point ordinaire au temps θ_0. Si ρ ne change pas de signe, γ traverse la tangente mais bute sur les autres droites passant pas $(0,0)$. Par conséquent, γ a un point de rebroussement de première espèce. ■

Exemple.
Soit $\rho : \mathbb{R} \to \mathbb{R}$ la fonction définie par $\rho(\theta) := \theta$. La courbe définie en coordonnées polaires par ρ admet une tangente horizontale et un point ordinaire au temps $\theta = 0$ (voir figure III.2.12).
Soit $\rho : \mathbb{R} \to \mathbb{R}$ la fonction définie par $\rho(\theta) := \sin^2(\theta/2)$. La courbe définie en coordonnées polaires par ρ est une cardioïde qui admet une tangente horizontale et un point de rebroussement de première espèce au temps $t = 0$ (voir figure III.2.13).

FIGURE III.2.12 –

FIGURE III.2.13 –

2.3 Branches infinies

Étudions maintenant les branches infinies d'une courbe paramétrée définie en coordonnées polaires par une fonction continue $\rho : I \to \mathbb{R}$. Soit :

$$a \in \mathbb{R} \cup \{-\infty, +\infty\}$$

une borne de I. Tout d'abord, la courbe admet une branche infinie lorsque $\theta \to a$ si, et seulement si, $|\rho(\theta)| \to +\infty$ lorsque $\theta \to a$. Si $a = +\infty$ ou $a = -\infty$, il n'y a pas de direction asymptotique. Si $a \in \mathbb{R}$, la courbe a une direction asymptotique dirigée par u_a quand θ tend vers a. On va donc se placer dans le cas où $a \in \mathbb{R}$ et étudier l'existence éventuelle d'une asymptote.

Proposition 10.
La courbe admet une *asymptote* lorsque θ tend vers $a \in \mathbb{R}$ si, et seulement si :

$$\lim_{\theta \to a} |\rho(\theta)| = +\infty \quad \text{et} \quad \lim_{\theta \to a} \rho(\theta) \cdot (\theta - a) = k \in \mathbb{R}.$$

L'*asymptote* est la droite passant par le point $O + kv_a$ et dirigée par le vecteur u_a.

Démonstration. Puisque $\dfrac{\mathrm{d}}{\mathrm{d}\theta}(u_\theta) = v_\theta$, on a :

$$u_\theta = u_a + (u_\theta - u_a) = u_a + (\theta - a)v_a + (\theta - a)\varepsilon(\theta)$$

avec $\varepsilon(\theta) \underset{\theta \to a}{\longrightarrow} (0,0)$. Donc :

$$\gamma(\theta) = O + \rho(\theta) \cdot u_\theta = O + \rho(\theta)u_a + \rho(\theta) \cdot (\theta - a)v_a + \rho(\theta) \cdot (\theta - a)\varepsilon(\theta).$$

Le résultat suit aisément de la définition d'une asymptote. ■

3 Étude métrique d'une courbe plane

Dans tout ce paragraphe, nous supposons le plan muni d'un repère orthonormé.

3.1 Longueur d'une courbe

Définition 18.
Étant donnée une courbe $\gamma : [a,b] \to \mathbb{R}^2$, la *longueur* $\ell(\gamma) \in \mathbb{R} \cup \{+\infty\}$ de γ est définie par :

$$\ell(\gamma) := \sup_{t_0 = a < t_1 < \cdots < t_n = b} \sum_{i=0}^{n-1} \left\| \overrightarrow{\gamma(t_i)\gamma(t_{i+1})} \right\|.$$

Nous dirons que $t_0 = a < t_1 < \cdots < t_n = b$ est une subdivision de $[a,b]$ et que la réunion des segments $\big[\gamma(t_i), \gamma(t_{i+1})\big]$ est une ligne polygonale inscrite. Nous venons de définir la longueur d'une courbe γ comme la borne supérieure de toutes les longueurs des lignes polygonales inscrites.

Définition 19. La courbe $\gamma : [a,b] \to \mathbb{R}^2$ est *rectifiable* si sa longueur est finie.

Proposition 11. Considérons une courbe rectifiable $\gamma : [a,b] \to \mathbb{R}^2$ et un intervalle $[t_0, t_1] \subset [a,b]$. Alors :

$$\ell\big(\gamma_{|_{[t_0,t_1]}}\big) \leqslant \ell(\gamma).$$

En particulier, la restriction de γ à l'intervalle $[t_0, t_1]$ est rectifiable.

Démonstration. À toute subdivision de $[t_0, t_1]$, nous pouvons ajouter (si nécessaire) les points a et b pour obtenir une subdivision de $[a,b]$. On en déduit que :

$$\ell(\gamma) \geqslant \left\| \overrightarrow{\gamma(a)\gamma(t_0)} \right\| + \ell\big(\gamma_{|_{[t_0,t_1]}}\big) + \left\| \overrightarrow{\gamma(t_1)\gamma(b)} \right\|.$$ ■

Proposition 12. Soit $\gamma : [a,b] \to \mathbb{R}^2$ une courbe de classe $\mathcal{C}^1$.
Alors, γ est rectifiable et :

$$\ell(\gamma) = \int_a^b \|\gamma'(t)\| \, \mathrm{d}t.$$

Démonstration. Considérons une subdivision $t_0 = a < t_1 < \cdots < t_n = b$. Comme :

$$\gamma(t_{i+1}) - \gamma(t_i) = \int_{t_i}^{t_{i+1}} \gamma'(t) \, \mathrm{d}t,$$

l'inégalité triangulaire (théorème 26 de la page 797) implique :

$$\left\|\overrightarrow{\gamma(t_i)\gamma(t_{i+1})}\right\| = \left\|\int_{t_i}^{t_{i+1}} \gamma'(t) \, \mathrm{d}t\right\| \leqslant \int_{t_i}^{t_{i+1}} \|\gamma'(t)\| \, \mathrm{d}t.$$

Par conséquent :

$$\sum_{i=0}^{n-1} \left\|\overrightarrow{\gamma(t_i)\gamma(t_{i+1})}\right\| \leqslant \int_a^b \|\gamma'(t)\| \, \mathrm{d}t \ ;$$

en prenant la borne supérieure sur toutes les subdivisions de $[a,b]$, nous obtenons :

$$\ell(\gamma) \leqslant \int_a^b \|\gamma'(t)\| \, \mathrm{d}t.$$

En particulier, la courbe est rectifiable.
Définissons maintenant une fonction $\varphi : [a,b] \to \mathbb{R}$ par :

$$\forall t \in [a,b] \quad \varphi(t) := \ell\big(\gamma_{|_{[a,t]}}\big).$$

On a $\varphi(a) = 0$ et $\varphi(b) = \ell(\gamma)$. On va montrer que φ est dérivable de dérivée $\|\gamma'(t)\|$, ce qui conclura la preuve.
Dans l'encadrement suivant valable pour $h > 0$:

$$\frac{1}{h}\left\|\overrightarrow{\gamma(t)\gamma(t+h)}\right\| \leqslant \frac{\varphi(t+h) - \varphi(t)}{h} \leqslant \frac{1}{h}\int_t^{t+h} \|\gamma'(u)\| \, \mathrm{d}u,$$

le membre de gauche et le membre de droite ont la même limite quand $h \to 0$. Cette limite vaut $\|\gamma'(t)\|$, donc $\varphi'(t) = \|\gamma'(t)\|$. ∎

Exemples.

1. Soit γ_0 un point quelconque du plan et v_0 un vecteur non nul de $\mathbb{R}^2$. Le support de la courbe $\gamma : [0,1] \to \mathbb{R}^2$ définie par $\gamma(t) := \gamma_0 + t \cdot v_0$ est le segment $[\gamma_0, \gamma_0 + v_0]$. La longueur de γ est :

$$\ell(\gamma) = \int_0^1 \|v_0\| \, \mathrm{d}t = \|v_0\|.$$

2. Le support de la courbe $\gamma : [0, 2\pi] \to \mathbb{R}^2$ définie par :

$$\gamma(t) := (x_0 + r\cos t, y_0 + r\sin t)$$

est le cercle de centre (x_0, y_0) et de rayon r. La longueur de γ est :

$$\ell(\gamma) = \int_0^{2\pi} \sqrt{(-r\sin t)^2 + (r\cos t)^2} \, \mathrm{d}t = 2\pi r.$$

Proposition 13. Soit $\gamma : [a,b] \to \mathbb{R}^2$ une courbe définie en coordonnées polaires par une fonction $\rho : [a,b] \to \mathbb{R}$ de classe $\mathcal{C}^1$. La longueur de γ est donnée par :

$$\ell(\gamma) = \int_a^b \sqrt{\rho^2 + (\rho')^2}\,(t)\,\mathrm{d}t.$$

Démonstration. Posons $x(t) = \rho(t)\cos t$ et $y(t) = \rho(t)\sin t$. Nous avons vu dans la démonstration de la proposition 8 de la page 484 que :

$$\gamma'(t) = \rho'(t)\cdot u_t + \rho(t)\cdot v_t.$$

Comme le repère $(O; u_t, v_t)$ est orthonormé, on a donc :

$$\|\gamma'\| = \sqrt{\rho^2 + (\rho')^2}.$$

Le résultat découle de la proposition 12 de la page précédente. ■

3.2 Paramétrage normal

Nous allons maintenant montrer qu'une courbe régulière peut être reparamétrée de manière à la parcourir à vitesse constante égale à 1. Nous dirons alors qu'on a un paramétrage normal de la courbe ou bien que la courbe est paramétrée par une abscisse curviligne.
Rappelons que les homéomorphismes entre deux intervalles $I_1 \subset \mathbb{R}$ et $I_2 \subset \mathbb{R}$ sont strictement monotones : soit ils sont strictement croissants, soit ils sont strictement décroissants.

Définition 20. Deux courbes $\gamma_1 : I_1 \to \mathbb{R}^2$ et $\gamma_2 : I_2 \to \mathbb{R}^2$ représentent la même *courbe géométrique* s'il existe un homéomorphisme $h : I_1 \to I_2$ tel que $\gamma_2 = \gamma_1 \circ h$. On définit ainsi une relation d'équivalence sur les courbes paramétrées. Les classes d'équivalence sont appelées des courbes géométriques, et les représentants d'une courbe géométrique sont appelés des paramétrages de la courbe géométrique.
Si de plus h est croissant, nous dirons que γ_1 et γ_2 représentent la même courbe géométrique orientée. On définit ainsi une relation d'équivalence dont les classes d'équivalence sont appelées des *courbes géométriques orientées*.

Proposition 14. Si $\gamma_1 : [a_1,b_1] \to \mathbb{R}^2$ et $\gamma_2 : [a_2,b_2] \to \mathbb{R}^2$ représentent la même courbe géométrique, alors $\ell(\gamma_1) = \ell(\gamma_2)$.

Démonstration. Soit $t_0 = a_1 < t_1 \cdots < t_n = b_1$ une subdivision de $[a_1,b_1]$. Si $h(a_1) = a_2$, alors $h(t_0) = a_2 < h(t_1) < \cdots < h(t_n) = b_2$ est une subdivision de $[a_2,b_2]$. Il en découle aisément que :

$$\ell(\gamma_2) \geqslant \sum_{i=0}^{n-1} \left\|\overrightarrow{\gamma_2\big(h(t_i)\big)\gamma_2\big(h(t_{i+1})\big)}\right\| = \sum_{i=0}^{n-1} \left\|\overrightarrow{\gamma_1(t_i)\gamma_1(t_{i+1})}\right\|.$$

En prenant la borne supérieure sur toutes les subdivisions de $[a_1,b_1]$, nous obtenons :

$$\ell(\gamma_2) \geqslant \ell(\gamma_1).$$

Le problème étant symétrique par rapport à γ_1 et à γ_2, nous en déduisons que :

$$\ell(\gamma_2) = \ell(\gamma_1).$$

Si $h(a_1) = b_2$, alors $h(t_n) = a_2 < h(t_{n-1}) < \cdots < h(t_0) = b_2$ est une subdivision de $[a_2, b_2]$ et nous pouvons faire le même raisonnement que précédemment. ■

Définition 21. Soit $\gamma : I \to \mathbb{R}^2$ une courbe paramétrée. Une application $\varphi : I \to \mathbb{R}$ est une *abscisse curviligne* de γ si pour tout intervalle $[a, b] \subset I$:

$$\ell\big(\gamma_{|_{[a,b]}}\big) = \big|\varphi(b) - \varphi(a)\big|.$$

Définition 22. Un paramétrage $\gamma : I \to \mathbb{R}^2$ d'une courbe géométrique est *normal* si pour tout intervalle $[a, b] \subset I$:

$$\ell\big(\gamma_{|_{[a,b]}}\big) = b - a.$$

Remarque. Chaque fois que nous aurons à faire à un paramétrage normal $\gamma : I \to \mathbb{R}^2$, nous noterons $s \in I$ le paramètre. Par abus de langage, nous dirons que s est une abscisse curviligne. Nous dirons également que la courbe est paramétrée par une abscisse curviligne.

Proposition 15. Soit $\gamma : I \to \mathbb{R}^2$ une courbe de classe $\mathcal{C}^1$. Alors, γ est un paramétrage normal si, et seulement si, pour tout $s \in I$:

$$\big\|\gamma'(s)\big\| = 1.$$

Démonstration. Si $\big\|\gamma'(s)\big\| = 1$, alors :

$$\ell\big(\gamma_{|_{[a,b]}}\big) = \int_a^b \|\gamma'(s)\| \,\mathrm{d}s = b - a.$$

Réciproquement, si γ est un paramétrage normal d'une courbe géométrique, fixons $s_0 \in \overset{\circ}{I}$ et définissons $\varphi : I \to \mathbb{R}$ par :

$$\varphi(s) := \int_{s_0}^{s} \|\gamma'(u)\| \,\mathrm{d}u.$$

La fonction φ est dérivable de dérivée égale à $\|\gamma'(s)\|$. Pour $s > s_0$, nous avons :

$$\varphi(s) = \ell\big(\gamma_{|_{[s_0,s]}}\big) = s - s_0$$

et donc $\|\gamma'(s_0)\| = \varphi'(s_0) = 1$. ■

Proposition 16. Toute courbe géométrique (orientée) admettant un paramétrage par une courbe régulière de classe $\mathcal{C}^1$, admet un paramétrage normal de classe $\mathcal{C}^1$.

Démonstration. Soit $\gamma : I \to \mathbb{R}^2$ un paramétrage de la courbe géométrique. Par hypothèse, nous pouvons supposer que γ est de classe $\mathcal{C}^1$ et que γ' ne s'annule pas sur I.

Choisissons $t_0 \in I$ et définissons $\varphi : I \to \mathbb{R}$ par :

$$\varphi(t) := \int_{t_0}^{t} \|\gamma'(t)\| \, \mathrm{d}t.$$

Posons $J = \varphi(I)$. Nous allons montrer que $\varphi : I \to J$ est un difféomorphisme de classe $\mathcal{C}^1$ et que $\gamma_1 = \gamma \circ \varphi^{-1} : J \to \mathbb{R}^2$ est un paramétrage normal de la même courbe géométrique que γ.

L'application φ est de classe $\mathcal{C}^1$ et de dérivée :

$$\varphi'(t) = \|\gamma'(t)\| > 0.$$

Elle réalise donc un difféomorphisme croissant de classe $\mathcal{C}^1$ de I sur J.

En dérivant $\gamma(t) = \gamma_1 \circ \varphi(t)$, nous obtenons :

$$\gamma'(t) = \gamma_1'\big(\varphi(t)\big) \cdot \varphi'(t).$$

En prenant la norme de chaque côté, il s'en suit :

$$\|\gamma'(t)\| = \big\|\gamma_1'\big(\varphi(t)\big)\big\| \cdot \|\gamma'(t)\|.$$

Nous pouvons simplifier à gauche et à droite par $\|\gamma'(t)\|$ qui ne s'annule pas. Nous obtenons alors que pour tout $s = \varphi(t) \in J$, $\|\gamma_1'(s)\| = 1$. ■

Proposition 17. Si $\gamma_1 : J_1 \to \mathbb{R}^2$ et $\gamma_2 : J_2 \in \mathbb{R}^2$ sont deux paramétrages normaux de la même courbe géométrique, alors, il existe $\varepsilon \in \{+1, -1\}$ et $c \in \mathbb{R}$ tels que $s \in J_1$ si, et seulement si, $\varepsilon s + c \in J_2$ et $\gamma_1(s) = \gamma_2(\varepsilon s + c)$. Si de plus γ_1 et γ_2 représentent la même courbe géométrique orientée, alors $\varepsilon = +1$.

Démonstration. Comme γ_1 et γ_2 représentent la même courbe géométrique, il existe un homéomorphisme $\varphi : J_1 \to J_2$ tel que $\gamma_1 = \gamma_2 \circ \varphi$. Si $[a, b] \subset I$ nous avons, d'après la proposition 14 de la page 488 :

$$\ell\left(\gamma_2|_{\varphi([a,b])}\right) = \ell\big(\gamma_1|_{[a,b]}\big). \tag{1}$$

Si φ est croissant, $\varphi\big([a, b]\big) = \big[\varphi(a), \varphi(b)\big]$ et donc l'égalité (1) devient :

$$\varphi(b) - \varphi(a) = b - a.$$

Si l'on fixe $t_0 \in I$, alors pour tout $t \in I$, nous aurons :

$$\varphi(t) - \varphi(t_0) = t - t_0 \quad \text{et donc} \quad \varphi(t) = t + \varphi(t_0) - t_0 = t + c.$$

Si φ est décroissant, $\varphi\big([a, b]\big) = \big[\varphi(b), \varphi(a)\big]$ et donc l'égalité (1) devient :

$$\varphi(a) - \varphi(b) = b - a.$$

Si l'on fixe $t_0 \in I$, alors pour tout $t \in I$, nous aurons :

$$\varphi(t) - \varphi(t_0) = t_0 - t \quad \text{et donc} \quad \varphi(t) = -t + \varphi(t_0) + t_0 = -t + c.$$ ■

3.3 Courbure

Dans ce paragraphe, nous allons introduire une notion fondamentale dans l'étude géométrique des courbes : la courbure. La courbure en un point de la courbe traduit la rapidité avec laquelle la courbe s'éloigne de sa tangente. Plus précisément, imaginons un point parcourant la courbe à vitesse constante égale à 1. Si la courbe dévie de sa tangente, le point subira une accélération. La courbure mesure cette accélération. Si le point se déplace en ligne droite, la courbure sera nulle. Plus la courbe dévie de sa tangente, plus la courbure sera grande. Si le plan est orienté, la courbure aura un signe : positif si la courbe tourne à gauche et négatif si la courbe tourne à droite.
Nous supposons maintenant le plan affine muni d'une orientation. Dans un premier temps, nous nous plaçons dans le cadre où $\gamma : I \to \mathbb{R}^2$ est un paramétrage normal de classe $\mathcal{C}^2$.

Définition 23. Le *repère de Serret-Frenet* de γ en $s \in I$ est le repère orthonormé direct $\big(\gamma(s); \overrightarrow{T}(s), \overrightarrow{N}(s)\big)$ où $\overrightarrow{T}(s) := \gamma'(s)$.

FIGURE III.2.14 – Le repère de Serret-Frenet de γ en s.

Proposition 18. Il existe une fonction continue $\kappa : \overset{\circ}{I} \to \mathbb{R}$ telle que :

$$\overrightarrow{T}' = \kappa \cdot \overrightarrow{N}.$$

Démonstration. Comme γ est un paramétrage normal :

$$(\overrightarrow{T} \mid \overrightarrow{T}) = \|\gamma'\|^2 = 1.$$

En dérivant, nous obtenons :

$$2\,(\overrightarrow{T}' \mid \overrightarrow{T}) = 0,$$

ce qui montre que pour tout $s \in \overset{\circ}{I}$, le vecteur $\overrightarrow{T}(s)$ est colinéaire au vecteur $\overrightarrow{N}(s)$. Ceci montre l'existence de la fonction κ. Nous avons alors $\overrightarrow{T}' = \kappa \cdot \overrightarrow{N}$, et :

$$\kappa = (\overrightarrow{T}' \mid \overrightarrow{N})$$

est donc une fonction continue. ∎

Définition 24. La fonction $\kappa = (\overrightarrow{T}' \mid \overrightarrow{N}) : \overset{\circ}{I} \to \mathbb{R}$ est la *courbure algébrique* de γ. La courbure de γ est la fonction $|\kappa|$.

Proposition 19. Pour tout $s_0 \in \overset{\circ}{I}$:

$$\gamma(s) = \gamma(s_0) + (s - s_0)\overrightarrow{T}(s_0) + \frac{\kappa(s_0)}{2}(s - s_0)^2\overrightarrow{N}(s_0) + (s - s_0)^2\varepsilon(s)$$

avec $\varepsilon(s) \underset{s \to s_0}{\longrightarrow} (0, 0)$.

Démonstration. Il s'agit de faire un développement de Taylor à l'ordre 2. Pour s_0, la courbe passe par $\gamma(s_0)$. Le vecteur vitesse est $\gamma'(s_0) = \overrightarrow{T}(s_0)$. Enfin :

$$\gamma''(s_0) = \overrightarrow{T}'(s_0) = \kappa(s_0) \cdot \overrightarrow{N}(s_0).$$

Remarque. Considérons la courbe auxilliaire :

$$s \mapsto \gamma(s_0) + (s - s_0)\overrightarrow{T}(s_0) + \frac{\kappa(s_0)}{2}(s - s_0)^2\overrightarrow{N}(s_0).$$

C'est la courbe représentative de la fonction $x \mapsto \kappa x^2/2$ dans le repère de Serret-Frenet au temps s_0. C'est une parabole. Plus $|\kappa|$ est grand, plus la parabole s'éloigne rapidement de sa tangente. Si l'on parcourt la courbe dans le sens des s croissants, en un point où la courbure est strictement positive, on tourne vers la gauche. En un point où la courbure est strictement négative, on tourne vers la droite.

Exemples.

1. Soit M un point de $\mathbb{R}^2$ et u un vecteur unitaire. La courbe paramétrée $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par $\gamma(s) := M + s \cdot u$ est un paramétrage normal. Le support de γ est une droite. La courbure algébrique de γ est la fonction nulle. En effet, $\overrightarrow{T} = u$ et $\overrightarrow{T}' = (0, 0)$.
2. Soit (x_0, y_0) un point de $\mathbb{R}^2$ et $R > 0$ un réel. La courbe paramétrée $\gamma : \mathbb{R} \to \mathbb{R}^2$ définie par :
$$\gamma(s) := \big(x_0 + R\cos(s/R), y_0 + R\sin(s/R)\big)$$
est un paramétrage normal. Le support de γ est un cercle de rayon R. On a :
$$\overrightarrow{T} = \big(-\sin(s/R), \cos(s/R)\big), \quad \overrightarrow{N} = \big(-\cos(s/R), -\sin(s/R)\big)$$
et :
$$\overrightarrow{T}'(s) = \frac{1}{R} \cdot \big(-\cos(s/R), -\sin(s/R)\big) = \frac{1}{R} \cdot \overrightarrow{N}(s).$$
La courbure algébrique de γ est donc la fonction constante égale à $1/R$.

Proposition 20 (Formules de Serret-Frenet). Si $\gamma : I \to \mathbb{R}^2$ est un paramétrage normal de classe $\mathcal{C}^2$, si $\big(\gamma(s); \overrightarrow{T}(s), \overrightarrow{N}(s)\big)$ est le repère de Serret-Frenet en s et si $\kappa : \overset{\circ}{I} \to R$ est la courbure algébrique de γ, alors :

$$\overrightarrow{T}' = \kappa \cdot \overrightarrow{N} \qquad \text{et} \qquad \overrightarrow{N}' = -\kappa \cdot \overrightarrow{T}.$$

Démonstration. Si $\gamma(s) = \big(x(s), y(s)\big)$, alors :

$$\overrightarrow{T} = (x', y') \quad \text{et} \quad \overrightarrow{N} = (-y', x').$$

Nous venons de voir dans la démonstration précédente que :

$$\overrightarrow{T}' = (x'', y'') = \kappa \cdot \overrightarrow{N} = (-\kappa y', \kappa x').$$

Nous avons donc :

$$\overrightarrow{N}' = (-y'', x'') = (-\kappa x', -\kappa y') = -\kappa \cdot \overrightarrow{T}.$$

■

Proposition 21. La courbe paramétrée γ a un point birégulier au temps s_0 si, et seulement si, $\kappa(s_0) \neq 0$.

Démonstration. Nous savons que $\gamma'(s_0) = \overrightarrow{T}(s_0)$ n'est pas nul et que le vecteur $\gamma''(s_0) = \kappa(s_0) \cdot \overrightarrow{N}(s_0)$ est orthogonal à $\gamma'(s_0)$. Par conséquent, les vecteurs $\gamma'(s_0)$ et $\gamma''(s_0)$ ne sont pas colinéaires si, et seulement si, $\gamma''(s_0) \neq (0,0)$, donc si, et seulement si, $\kappa(s_0) \neq 0$. ■

Corollaire 22. Si γ a un point d'inflexion au temps s_0, alors $\kappa(s_0) = 0$.

Démonstration. Un point birégulier est un point ordinaire. Si γ a un point d'inflexion au temps s_0, ce n'est pas un point birégulier et donc, $\kappa(s_0) = 0$. ■

Définition 25. Si γ a un point birégulier au temps s_0 (c'est-à-dire si $\kappa(s_0) \neq 0$), le réel $R(s_0) := 1/\kappa(s_0)$ est le *rayon de courbure algébrique* en s_0 et le point $\gamma(s_0) + R(s_0) \cdot \overrightarrow{N}(s_0)$ est le *centre de courbure* en s_0. Le *rayon de courbure* de γ en s_0 est $|R(s_0)|$.

FIGURE III.2.15 –

Considérons maintenant une courbe régulière $\gamma : I \to \mathbb{R}^2$ de classe $\mathcal{C}^2$. Nous allons voir que nous pouvons encore définir le repère de Serret-Frenet, la courbure algébrique et la courbure, et en un point birégulier, le rayon de courbure algébrique, le rayon de courbure et le centre de courbure.

Définition 26. Le *repère de Serret-Frenet* de γ au temps $t \in I$ est le repère orthonormé direct $\big(\gamma(t); \overrightarrow{T}(t), \overrightarrow{N}(t)\big)$ où $\overrightarrow{T} = \gamma'/\|\gamma'\|$.

Soit maintenant $\gamma_1 : J_1 \to \mathbb{R}^2$ un paramétrage normal de la même courbe géométrique orientée que $\gamma : I \to \mathbb{R}^2$. Dans ce cas, il existe un homéomorphisme croissant (une abscisse curviligne) $\varphi : I \to J_1$ tel que $\gamma = \gamma_1 \circ \varphi$. Comme γ_1 est un paramétrage normal, nous pouvons définir la courbure algébrique $\kappa_1 : J_1 \to \mathbb{R}$ de γ_1.

Définition 27. La *courbure algébrique* de γ est la fonction $\kappa := \kappa_1 \circ \varphi : I \to \mathbb{R}$. La *courbure* de γ est la fonction $|\kappa| : I \to \mathbb{R}$.

Pour que cette définition ait un sens, il nous faut vérifier qu'elle ne dépend pas du choix du paramétrage normal γ_1. D'après la proposition 17 de la page 490, si $\gamma_2 : J_2 \to \mathbb{R}^2$ est un autre paramétrage normal de la même courbe géométrique orientée que γ, il existe une constante c telle que $s \in J_1$ si, et seulement si, $s + c \in J_2$ et $\gamma_1(s) = \gamma_2(s + c)$. Il s'en suit que :

$$\overrightarrow{T}_1(s) = \overrightarrow{T}_2(s + c), \quad \overrightarrow{N}_1(s) = \overrightarrow{N}_2(s + c)$$

et donc :

$$\kappa_1(s) = \big(\, \overrightarrow{T}'_1(s) \mid \overrightarrow{N}_1(s) \,\big) = \big(\, \overrightarrow{T}'_2(s + c) \mid \overrightarrow{N}_2(s + c) \,\big) = \kappa_2(s + c).$$

Proposition 23. La courbe paramétrée γ a un point birégulier au temps t_0 si, et seulement si, γ_1 a un point birégulier au temps $s_0 = \varphi(t_0)$.

Démonstration. En dérivant $\gamma = \gamma_1 \circ \varphi$, nous obtenons :

$$\gamma'(t_0) = \varphi'(t_0) \cdot \gamma'_1(s_0)$$

$$\text{et} \quad \gamma''(t_0) = \big(\varphi'(t_0)\big)^2 \cdot \gamma''_1(s_0) + \varphi''(t_0) \cdot \gamma'_1(s_0)$$

avec $\varphi'(t_0) \neq 0$. Par conséquent, $\gamma'(t_0)$ et $\gamma''(t_0)$ sont colinéaires si, et seulement si, $\gamma'_1(s_0)$ et $\gamma''_1(s_0)$ sont colinéaires. ∎

Corollaire 24. Si γ a un point birégulier au temps $t_0 \in \overset{\circ}{I}$, alors $\kappa(t_0) \neq 0$.

La définition suivante a donc un sens.

Définition 28. Si γ a un point birégulier au temps $t_0 \in \overset{\circ}{I}$, $R(t_0) := 1/\kappa(t_0)$ est le rayon de courbure algébrique au temps t_0 et le point $\gamma(t_0) + R(t_0) \cdot \overrightarrow{N}(t_0)$ est le *centre de courbure* au temps t_0. Le *rayon de courbure* de γ au temps t_0 est $|R(t_0)|$.

Remarque. Ce rayon de courbure algébrique et ce centre de courbure coïncident avec le rayon de courbure algébrique et le centre de courbure de γ_1 au temps $s_0 = \varphi(t_0)$.

Nous terminerons ce paragraphe sur la courbure par une formule fondamentale qui montre que la courbure mesure la vitesse de rotation du vecteur unitaire tangent $\overrightarrow{T}$.
L'application $\overrightarrow{T} : I \to \mathbb{R}^2$ est une application de classe $\mathcal{C}^1$ et $\|\overrightarrow{T}\| = 1$. Il existe donc une application $\alpha : I \to \mathbb{R}$ de classe $\mathcal{C}^1$ telle que :

$$\overrightarrow{T} = \cos\alpha\,\vec{\imath} + \sin\alpha\,\vec{\jmath}.$$

Proposition 25. Si $\gamma : I \to \mathbb{R}$ est une courbe de classe $\mathcal{C}^2$, si $s : I \to \mathbb{R}$ est une abscisse curviligne croissante et si $\alpha : I \to \mathbb{R}$ est une application de classe $\mathcal{C}^1$ telle que $\overrightarrow{T} = \cos\alpha\vec{\imath} + \sin\alpha\vec{\jmath}$, alors la courbure algébrique $\kappa : \overset{\circ}{I} \to \mathbb{R}$ est donnée par :

$$\kappa(t) = \frac{\alpha'(t)}{s'(t)}.$$

En particulier, si γ est un paramétrage normal d'abscisse curviligne s, alors :

$$\kappa = \frac{\mathrm{d}\alpha}{\mathrm{d}s}.$$

Démonstration. Soit $s : I \to \mathbb{R}$ une abscisse curviligne croissante. Alors, $\gamma_1 := \gamma \circ s$ est un paramétrage normal paramétrant la même courbe géométrique orientée que γ. On a donc :

$$\overrightarrow{T} = \overrightarrow{T}_1 \circ s, \quad \overrightarrow{N} = \overrightarrow{N}_1 \circ s, \quad \kappa = \kappa_1 \circ s \quad \text{et} \quad \overrightarrow{T}'_1 = \kappa_1 \cdot \overrightarrow{N}_1.$$

En dérivant $\overrightarrow{T} = \cos\alpha\,\vec{\imath} + \sin\alpha\,\vec{\jmath}$, on obtient :

$$\overrightarrow{T}' = (-\sin\alpha\,\vec{\imath} + \cos\alpha\,\vec{\jmath}) \cdot \alpha' = \alpha' \cdot \overrightarrow{N}.$$

Par conséquent :

$$\alpha' \cdot \overrightarrow{N} = \overrightarrow{T}' = (\overrightarrow{T}'_1 \circ s) \cdot s' = \big((\kappa_1 \cdot \overrightarrow{N}_1) \circ s\big) \cdot s' = (\kappa_1 \circ s) \cdot (\overrightarrow{N}_1 \circ s) \cdot s' = \kappa s' \cdot \overrightarrow{N}.$$

D'où $\alpha' = \kappa s'$. ∎

3.4 Théorème fondamental

L'un des intérêts majeurs de la courbure algébrique d'une courbe régulière γ est que cette quantité décrit bien la géométrie de la courbe :

- si la courbure algébrique est positive, la courbe « tourne à gauche » ;
- si la courbure algébrique est négative, la courbe « tourne à droite » ;
- plus la courbure est grande, plus la courbe « tourne vite ».

On peut donc s'imaginer pouvoir tracer la courbe en ne connaissant que sa courbure. C'est essentiellement ce que dit le théorème suivant. Cette démonstration est basée sur le théorème de Cauchy-Lipschitz qui ne sera démontré que dans le volume de L2. En particulier, nous utiliserons le résultat suivant.
Si $A : I \to M_n(\mathbb{R})$ est une application continue, alors pour tout $t_0 \in I$ et tout $Y_0 \in \mathbb{R}^n$, l'équation différentielle $Y' = A(t) \cdot Y$ admet une unique solution maximale $f : J \to \mathbb{R}^n$ définie sur un intervalle $J \subset I$ contenant t_0 et vérifiant $f(t_0) = Y_0$. Si de plus, cette solution est bornée, alors $I = J$.

Théorème 26. Étant donnés une fonction continue $\kappa : I \to \mathbb{R}$, un temps $s_0 \in \mathring{I}$, un point $\gamma_0 \in \mathbb{R}^2$ et un vecteur $u_0 \in \mathbb{R}^2$ de norme $\|u_0\| = 1$, il existe un unique paramétrage normal $\gamma : I \to \mathbb{R}^2$ de classe $\mathcal{C}^2$, de courbure κ, telle que $\gamma(s_0) = \gamma_0$ et $\gamma'(s_0) = u_0$.

Démonstration.

<u>*Existence d'une solution*</u>. Soit v_0 l'unique vecteur de $\mathbb{R}^2$ tel que (u_0, v_0) soit une base orthonormée directe de $\mathbb{R}^2$.

Considérons la solution maximale $s \mapsto \big(\overrightarrow{T}(s), \overrightarrow{N}(s)\big) \in \mathbb{R}^2 \times \mathbb{R}^2$ de l'équation différentielle linéaire à coefficients continus :

$$Y' = \begin{pmatrix} 0 & \kappa(s) \\ \kappa(s) & 0 \end{pmatrix} \cdot Y$$

qui vérifie les conditions initiales $\overrightarrow{T}(s_0) = u_0$ et $\overrightarrow{N}(s_0) = v_0$. Cette solution n'est *a priori* définie que dans un voisinage J de s_0.

Lemme 27. Pour tout $s \in J$, $\big(\overrightarrow{T}(s), \overrightarrow{N}(s)\big)$ est une base orthonormée directe de $\mathbb{R}^2$.

Démonstration. Considérons la fonction vectorielle :

$$G := \big((\overrightarrow{T} \mid \overrightarrow{T}), (\overrightarrow{T} \mid \overrightarrow{N}), (\overrightarrow{N} \mid \overrightarrow{N})\big).$$

Alors, G est solution de l'équation différentielle linéaire à coefficients continus :

$$Y' = \begin{pmatrix} 0 & 2\kappa(s) & 0 \\ -\kappa(s) & 0 & \kappa(s) \\ 0 & -2\kappa(s) & 0 \end{pmatrix} \cdot Y.$$

Il y a une unique solution telle que $G(s_0) = (1, 0, 1)$. La fonction constante égale à $(1, 0, 1)$ est solution. Donc, pour tout $s \in J$, $G(s) = (1, 0, 1)$. Cela montre que la base $\big(\overrightarrow{T}(s), \overrightarrow{N}(s)\big)$ est orthonormée.

Le déterminant de la matrice $\big(\overrightarrow{T}(s), \overrightarrow{N}(s)\big)$ est une fonction continue qui ne peut prendre que les valeurs $+1$ ou -1 et prend la valeur $+1$ en s_0. Elle est donc constante égale à $+1$, et la base $\big(\overrightarrow{T}(s), \overrightarrow{N}(s)\big)$ est orthonormée directe. ∎

Comme la longueur du vecteur $\big(\overrightarrow{T}(s), \overrightarrow{N}(s)\big)$ est bornée sur J (en fait constante égale à $\sqrt{2}$), la solution maximale est définie sur I, *i.e.* $J = I$. Définissons $\gamma : I \to \mathbb{R}^2$ par :

$$\gamma(s) := \gamma_0 + \int_{s_0}^{s} \overrightarrow{T}(t)\,\mathrm{d}t.$$

Alors, γ est un paramétrage normal de courbure algébrique κ. En effet, comme $\overrightarrow{T}' = \kappa \cdot \overrightarrow{N}$, la fonction $\overrightarrow{T}$ est de classe $\mathcal{C}^1$. La primitive γ est donc de classe $\mathcal{C}^2$.

De plus, comme $\overrightarrow{T}(s) = \gamma'(s)$,

- $\|\gamma'(s)\| = 1$, ce qui montre que la courbe γ est un paramétrage normal ;
- le repère de Serret-Frenet de γ au temps s est $\big(\gamma(s), \overrightarrow{T}(s), \overrightarrow{N}(s)\big)$;
- $\overrightarrow{T}' = \kappa \cdot \overrightarrow{N}$.

Cela montre que κ est la courbure de γ.

<u>*Unicité de la solution*</u>. Supposons qu'il existe deux paramétrages normaux $\gamma : I \to \mathbb{R}^2$ et $\gamma_2 : I \to \mathbb{R}^2$ de classe $\mathcal{C}^2$, de courbure κ, telles que $\gamma_1(s_0) = \gamma_2(s_0) = \gamma_0$, et $\gamma_1'(s_0) = \gamma_2'(s_0) = u_0$. Soient $(\gamma_1, \overrightarrow{T}_1, \overrightarrow{N}_1)$ et $(\gamma_2, \overrightarrow{T}_2, \overrightarrow{N}_2)$ les repères de Serret-Frenet associés. Considérons la fonction :

$$f := (\overrightarrow{T}_1 \mid \overrightarrow{T}_2) + (\overrightarrow{N}_1 \mid \overrightarrow{N}_2) : \mathring{I} \to \mathbb{R}.$$

Alors, $f(s_0) = (u_0 \mid u_0) + (v_0 \mid v_0) = 2$ et

$$\begin{aligned} f' &= (\overrightarrow{T}_1' \mid \overrightarrow{T}_2) + (\overrightarrow{T}_1 \mid \overrightarrow{T}_2') + (\overrightarrow{N}_1' \mid \overrightarrow{N}_2) + (\overrightarrow{N}_1 \mid \overrightarrow{N}_2') \\ &= \kappa(\overrightarrow{N}_1 \mid \overrightarrow{T}_2) + \kappa(\overrightarrow{T}_1 \mid \overrightarrow{N}_2) - \kappa(\overrightarrow{T}_1 \mid \overrightarrow{N}_2) - \kappa(\overrightarrow{N}_1 \mid \overrightarrow{T}_2) \\ &= 0. \end{aligned}$$

Par conséquent, la fonction f est constante égale à 2. Comme $\overrightarrow{T}_1$, $\overrightarrow{T}_2$, $\overrightarrow{N}_1$ et $\overrightarrow{N}_2$ sont des vecteurs unitaires, $(\overrightarrow{T}_1 \mid \overrightarrow{T}_2) \leqslant 1$ et $(\overrightarrow{N}_1 \mid \overrightarrow{N}_2) \leqslant 1$ avec égalité si, et seulement si, $\overrightarrow{T}_1 = \overrightarrow{T}_2$ et $\overrightarrow{N}_1 = \overrightarrow{N}_2$. Par conséquent, $\overrightarrow{T}_1 = \overrightarrow{T}_2$ et :

$$\gamma_1(s) = \gamma_0 + \int_{s_0}^{s} \overrightarrow{T}_1(t)\,\mathrm{d}t = \gamma_0 + \int_{s_0}^{s} \overrightarrow{T}_2(t)\,\mathrm{d}t = \gamma_2(s).$$

Exemple. Le support d'une courbe paramétrée de courbure constante égale à 0 est contenue dans une droite. Le support d'une courbe paramétrée de courbure constante égale à $1/R$ est contenue dans un cercle de rayon R.

4 Courbes de l'espace

Nous allons maintenant nous intéresser aux courbes de l'espace affine $\mathbb{R}^3$. Il s'agit d'une généralisation de ce que nous venons d'étudier pour les courbes planes. Nous introduirons les notions de tangente et de plan osculateur. Puis nous étudierons les propriétés métriques des courbes de l'espace. En particulier, nous introduirons la notion de courbure d'une courbe de l'espace, puis une nouvelle notion : celle de torsion.

Définition 29. Une *courbe paramétrée* de $\mathbb{R}^3$ est une application continue $\gamma : I \to \mathbb{R}^3$ où $I \subset \mathbb{R}$ est un intervalle non réduit à un point. La courbe $C = \gamma(I)$ est appelée le *support* de γ.

Exemples.

1. Le support de la courbe paramétrée $\gamma : \mathbb{R} \to \mathbb{R}^3$ définie par :
$$\gamma(t) := (x_0 + tu, y_0 + tv, z_0 + tw)$$
est la droite de $\mathbb{R}^3$ passant par le point (x_0, y_0, z_0) et dirigée par le vecteur (u, v, w).
2. Le support de la courbe paramétrée $\gamma : \mathbb{R} \to \mathbb{R}^3$ définie par :
$$\gamma(t) := (r\cos t, r\sin t, at)$$
avec $r > 0$ et $a > 0$ est une hélice circulaire (voir figure III.2.16).

FIGURE III.2.16 –

Comme dans le cas des courbes planes, une courbe $\gamma : I \to \mathbb{R}^3$ est dérivable, de classe $\mathcal{C}^k$ ou de classe $\mathcal{C}^\infty$ suivant que γ est une application dérivable, de classe $\mathcal{C}^k$ ou de classe $\mathcal{C}^\infty$. Une courbe régulière est une courbe paramétrée dérivable dont la dérivée ne s'annule pas.

4.1 Tangente et plan osculateur

Dans ce paragraphe, nous étudions le comportement de la courbe de l'espace $\gamma : I \to \mathbb{R}^3$ lorsque le paramètre t tend vers $t_0 \in \overset{\circ}{I}$. Nous supposerons que le support de la courbe n'est pas contenu dans une droite.

La définition d'une tangente pour une courbe de l'espace est la même que pour une courbe plane.

Définition 30 (Tangente). Soit $u_0 \in \mathbb{R}^3$ un vecteur non nul.
La droite $D = \gamma(t_0) + \mathrm{Vect}(u_0)$ est tangente à $\gamma : I \to \mathbb{R}^3$ au temps $t_0 \in I$ s'il existe $\lambda : I \to \mathbb{R}$ et $\varepsilon : I \to \mathbb{R}^3$ tels que :
$$\gamma(t) = \gamma(t_0) + \lambda(t)u_0 + \lambda(t)\varepsilon(t) \quad \text{avec} \quad \varepsilon(t) \underset{t\to t_0}{\longrightarrow} (0,0,0).$$

Nous allons maintenant introduire la notion de plan osculateur. Lorsqu'une courbe admet une tangente au temps t_0, il s'agit de définir, parmi les plans contenant la tangente, celui qui « approche le mieux » la courbe au temps t_0.

Définition 31 (Plan osculateur). Le plan P est osculateur à $\gamma : I \to \mathbb{R}$ au temps $t_0 \in I$ s'il existe $u_0 \in \mathbb{R}^3$, $v_0 \in \mathbb{R}^3$, $\lambda : I \to \mathbb{R}$ et $\mu : I \to \mathbb{R}$ et $\varepsilon : I \to \mathbb{R}^3$ tels que :

- $P = \gamma(t_0) + \mathrm{Vect}(u_0, v_0)$ et
- $\gamma(t) = \gamma_0 + \lambda(t)u_0 + \lambda(t)\mu(t)v_0 + \lambda(t)\mu(t)\varepsilon(t)$ avec $\lambda(t) \underset{t\to t_0}{\longrightarrow} 0$, $\mu(t) \underset{t\to t_0}{\longrightarrow} 0$ et $\varepsilon(t) \underset{t\to t_0}{\longrightarrow} (0,0,0)$.

Remarque. Si γ est une courbe contenue dans un plan P et si γ admet une tangente au temps t_0, alors P est osculateur à γ au temps t_0. Cependant, selon la définition que nous venons de donner, une courbe ayant un plan osculateur admet nécessairement une tangente : la droite $\gamma(t_0) + \mathrm{Vect}(u_0)$. Il existe donc des courbes planes qui n'ont pas de plan osculateur.

Exercice 7.
Déterminer le plan osculateur à la courbe paramétrée $\gamma : \mathbb{R} \to \mathbb{R}$ définie par $\gamma(t) = (t, t^2, t^3)$ au temps $t = 0$.
Solution. C'est le plan horizontal d'équation $z = 0$ car :

$$\gamma(t) = O + t\vec{\imath} + t^2\vec{\jmath} + t^2\varepsilon(t)$$

avec $\varepsilon(t) = t\vec{k} \underset{t\to 0}{\longrightarrow} \overrightarrow{0}$.

Nous avons alors une proposition analogue à la proposition 2 de la page 473.

Proposition 28. Supposons qu'il existe des entiers $q > p \geqslant 1$ et des vecteurs u_0 et v_0 non colinéaires tels que :

$$\gamma(t) = \gamma(t_0) + (t - t_0)^p \lambda(t) u_0 + (t - t_0)^q v_0 + (t - t_0)^q \varepsilon(t) \quad \text{avec}$$
$$\lambda(t) \underset{t\to t_0}{\longrightarrow} 1 \qquad \varepsilon(t) \underset{t\to t_0}{\longrightarrow} (0, 0, 0).$$

Alors, la droite $D = \gamma(t_0) + \mathrm{Vect}(u_0)$ est tangente à γ au temps t_0 et le plan $P = \gamma(t_0) + \mathrm{Vect}(u_0, v_0)$ est osculateur à γ au temps t_0.

Démonstration. Il suffit de revenir aux définitions de tangente et de plan osculateur en posant $\alpha(t) = (t - t_0)^p$ et $\beta(t) = (t - t_0)^q$. ∎

Définition 32. Une courbe dérivable γ a un *point régulier* au temps t_0 si $\gamma'(t_0) \neq (0, 0, 0)$. Si de plus γ est deux fois dérivable et si $\gamma'(t_0)$ et $\gamma''(t_0)$ ne sont pas colinéaires, γ a un *point birégulier* au temps t_0.

Corollaire 29. Si γ a un point birégulier au temps t_0, γ admet pour plan osculateur $P = \gamma(t_0) + \mathrm{Vect}\big(\gamma'(t_0), \gamma''(t_0)\big)$ au temps t_0.

Les notions de longueur et de paramétrage normal définies dans le cadre des courbes planes restent valides dans le cas des courbes de l'espace. Nous rappelons les définitions et les propriétés les plus importantes, sans en donner les démonstrations qui sont en tous points analogues à celles du paragraphe 3.
Tout d'abord, la longueur d'une courbe est définie comme la borne supérieure des longueurs des lignes polygonales inscrites :

$$\ell(\gamma) = \sup_{t_0 = a < t_1 < \cdots < t_n = b} \sum_{i=0}^{n-1} \left\| \overrightarrow{\gamma(t_i)\gamma(t_{i+1})} \right\|.$$

Les notions de courbes géométriques (respectivement courbes géométriques orientées) sont analogues. Ce sont les classes d'équivalence pour la relation $\gamma_1 : I_1 \to \mathbb{R}^3$ est équivalente

à $\gamma_2 : I_2 \to \mathbb{R}^3$ si, et seulement s'il existe un homéomorphisme (respectivement un homéomorphisme croissant) $h : I_1 \to I_2$ tel que $\gamma_1 = \gamma_2 \circ h$. Les représentants d'une courbe géométrique sont appelés les paramétrages de la courbe.
Un paramétrage $\gamma : I \to \mathbb{R}^3$ d'une courbe géométrique est normal si, et seulement si, pour tout segment $[a, b] \subset I$, $\ell\big(\gamma_{|_{[a,b]}}\big) = b - a$.
Si $\gamma : I \to \mathbb{R}^3$ est une courbe de classe $\mathcal{C}^1$, alors γ est un paramétrage normal d'une courbe géométrique si, et seulement si, pour tout $s \in I$, $\|\gamma'(s)\| = 1$.
Toute courbe géométrique (orientée) admettant un paramétrage par une courbe régulière de classe $\mathcal{C}^1$, admet un paramétrage normal de classe $\mathcal{C}^1$.

4.2 Courbure, torsion

Le but de ce paragraphe est de définir les notions de courbure et de torsion pour une courbe géométrique de l'espace. Nous supposerons que l'espace affine est orienté. Nous ferons l'hypothèse que la courbe admet un paramétrage normal $\gamma : I \to \mathbb{R}^3$ de classe $\mathcal{C}^3$. Le lecteur vérifiera que comme dans le paragraphe 3.3, les définitions ne dépendent pas du paramétrage normal de la courbe géométrique.
Dans le cas d'une courbe de l'espace, comme dans le cas d'une courbe plane, la courbure mesure la rapidité à laquelle une courbe s'éloigne de sa tangente. Cependant, dans le cas d'une courbe de l'espace, il n'est plus possible de distinguer si un point se déplaçant le long de la courbe tourne vers la droite, le haut, la gauche ou le bas. On ne peut donc plus attribuer de signe à la courbure, *i.e.* on ne peut plus parler de courbure algébrique.

Définition 33. La *courbure* d'un paramétrage normal $\gamma : I \to \mathbb{R}$ est la fonction continue $\kappa : \overset{\circ}{I} \to \mathbb{R}$ définie par $\kappa(t) := \big\|\gamma''(t)\big\|$.

Comme dans le cas d'une courbe plane, $\kappa(s_0) \neq 0$ si, et seulement si, la courbe γ a un point birégulier au temps s_0. Notons que vecteur $\gamma''(s)$ est orthogonal au vecteur $\gamma'(s)$. En effet, en dérivant, $(\gamma' \mid \gamma') = 1$, nous obtenons $2\,(\gamma'' \mid \gamma') = 0$. En un point birégulier, le vecteur $\overrightarrow{N}(s) = \gamma''(s)/\big\|\gamma''(s)\big\|$ est donc orthogonal au vecteur $\overrightarrow{T}(s) = \gamma'(s)$. Nous pouvons alors définir le rayon de courbure et le centre de courbure de la manière suivante.

Définition 34.
Si γ a un point birégulier au temps s_0 (*i.e.* $\kappa(s_0) \neq 0$), le réel positif $R(s_0) := 1/\kappa(s_0)$ est le *rayon de courbure* au temps s_0 et le point $\gamma(s_0) + R(s_0)\,\overrightarrow{N}(s_0)$ est le *centre de courbure* au temps s_0.

Introduisons maintenant la notion de torsion. Cette quantité n'est définie que pour des courbes de classe $\mathcal{C}^3$ n'admettant que des point biréguliers. Nous dirons que la courbe est birégulière. Rappelons qu'en un point birégulier, la courbe admet un plan osculateur.
La torsion d'une courbe de l'espace est une quantité qui mesure la vitesse de rotation du plan osculateur. En particulier, la torsion d'une courbe plane birégulière sera nulle.
Pour définir la torsion d'une courbe de l'espace birégulière, il nous faut d'abord introduire le repère mobile de Serret-Frenet.

Définition 35. Le *repère de Serret-Frenet* de γ au temps $s \in I$ est le repère orthonormé direct $\big(\gamma(s), \overrightarrow{T}(s), \overrightarrow{N}(s), \overrightarrow{B}(s)\big)$ où $\overrightarrow{T} := \gamma'$, $\overrightarrow{N} := \overrightarrow{T}'/\|\overrightarrow{T}'\|$.

L'application γ est de classe $\mathcal{C}^3$, l'application $\overrightarrow{T}$ est donc de classe $\mathcal{C}^2$ et ne s'annule pas. Par conséquent, l'application $\overrightarrow{N}$ est de classe $\mathcal{C}^1$. Comme la base $\big(\overrightarrow{T}, \overrightarrow{N}, \overrightarrow{B}\big)$ est orthonormée directe, le vecteur $\overrightarrow{B}$ peut être défini par :

$$\overrightarrow{B} = \overrightarrow{T} \wedge \overrightarrow{N},$$

où $u \wedge v$ désigne le produit vectoriel de u et de v. L'application $\overrightarrow{B}$ est donc elle aussi de classe $\mathcal{C}^1$.

Proposition 30. Il existe une fonction continue $\tau : \overset{\circ}{I} \to \mathbb{R}$ telle que :

$$\overrightarrow{B}' = \tau\, \overrightarrow{N}.$$

Démonstration. En dérivant $\overrightarrow{B} = \overrightarrow{T} \wedge \overrightarrow{N}$, nous obtenons :

$$\overrightarrow{B}' = \overrightarrow{T}' \wedge \overrightarrow{N} + \overrightarrow{T} \wedge \overrightarrow{N}' = \overrightarrow{T} \wedge \overrightarrow{N}'$$

car par définition, pour tout $s \in \overset{\circ}{I}$, le vecteur $\overrightarrow{T}'(s)$ est colinéaire au vecteur $\overrightarrow{N}(s)$. Nous en déduisons que le vecteur $\overrightarrow{B}'(s)$ est orthogonal au vecteur $\overrightarrow{T}(s)$.
De plus, en dérivant $(\overrightarrow{B} \mid \overrightarrow{B}) = \|\overrightarrow{B}\|^2 = 1$, nous obtenons $2\,(\overrightarrow{B}' \mid \overrightarrow{B}) = 0$. Par conséquent le vecteur $\overrightarrow{B}'(s)$ est orthogonal au vecteur $\overrightarrow{B}(s)$.
Un vecteur orthogonal à la fois à $\overrightarrow{T}(s)$ et à $\overrightarrow{B}(s)$ est colinéaire à $\overrightarrow{N}(s)$. ∎

Définition 36. La *torsion* de γ est la fonction $\tau := (\overrightarrow{B}' \mid \overrightarrow{N}) : \overset{\circ}{I} \to \mathbb{R}$.

Proposition 31 (Formules de Serret-Frenet). Si $\gamma : I \to \mathbb{R}^3$ est un paramétrage normal birégulier de classe $\mathcal{C}^3$, si $\big(\gamma(s); \overrightarrow{T}(s), \overrightarrow{N}(s), \overrightarrow{B}(s)\big)$ est le repère de Serret-Frenet au temps s, si $\kappa : \overset{\circ}{I} \to \mathbb{R}$ est la courbure de γ et si $\tau : \overset{\circ}{I} \to \mathbb{R}$ est la torsion de γ, alors :

$$\begin{cases} \overrightarrow{T}' = \kappa\, \overrightarrow{N} \\ \overrightarrow{N}' = -\kappa\, \overrightarrow{T} - \tau\, \overrightarrow{B} \\ \overrightarrow{B}' = \tau\, \overrightarrow{N}. \end{cases}$$

Démonstration. La seule formule manquante est la deuxième qui concerne $\overrightarrow{N}'$. En dérivant $\overrightarrow{N} = -\overrightarrow{T} \wedge \overrightarrow{B}$, nous obtenons

$$\overrightarrow{N}' = -\overrightarrow{T}' \wedge \overrightarrow{B} - \overrightarrow{T} \wedge \overrightarrow{B}' = -\kappa\, \overrightarrow{N} \wedge \overrightarrow{B} - \tau\, \overrightarrow{T} \wedge \overrightarrow{N} = -\kappa\, \overrightarrow{T} - \tau\, \overrightarrow{B}.$$ ∎

Dans le cas où γ est une courbe birégulière mais n'est pas un paramétrage normal, nous pouvons écrire $\gamma = \gamma_1 \circ h$ avec γ_1 un paramétrage normal birégulier. La courbure κ et la torsion τ de γ sont définies par :

$$\kappa = \kappa_1 \circ h \quad \text{et} \quad \tau = \tau_1 \circ h,$$

où κ_1 et τ_1 sont la courbure et la torsion de γ_1.

Exercice 8.
Déterminer la courbure et la torsion de l'hélice $\gamma : \mathbb{R} \to \mathbb{R}^3$ définie par :

$$\gamma(t) := (r\cos t, r\sin t, at) \quad \text{avec} \quad r > 0 \quad \text{et} \quad a > 0.$$

Solution. C'est une courbe de courbure constante $\kappa = r/(r^2 + a^2)$ et de torsion constante $\gamma = -a/(r^2 + a^2)$.

4.3 Théorème fondamental

Le résultat qui suit affirme qu'une courbe de l'espace birégulière est essentiellement caractérisée par sa courbure et sa torsion.

Théorème 32. Étant donnés deux fonctions continues $\kappa : I \to]0, +\infty[$ et $\tau : I \to \mathbb{R}$, un temps $s_0 \in \overset{\circ}{I}$, un point $\gamma_0 \in \mathbb{R}^2$ et deux vecteurs unitaires orthogonaux $u_0 \in \mathbb{R}^3$ et $v_0 \in \mathbb{R}^3$, il existe un unique paramétrage normal $\gamma : I \to \mathbb{R}^3$ de classe $\mathcal{C}^3$, de courbure κ et de torsion τ, telle que $\gamma(s_0) = \gamma_0$ et $\gamma'(s_0) = u_0$ et $\gamma''(s_0) = \kappa(s_0) \cdot v_0$.

Démonstration. La preuve est similaire à celle du théorème 26 de la page 496 (voir exercice III.2.15 de la page 505). ■

Exemple. Les seules courbes de l'espace birégulières de courbure constante et de torsion constante sont des arcs d'hélice. En effet, par tout point de l'espace, on peut faire passer une hélice, et il y a unicité de la courbe à courbure et torsion prescrites.

Exercice type corrigé

III.2.0 Soient $a \in \mathbb{R}$ et $\gamma_a : \mathbb{R} \to \mathbb{R}^2$ la courbe définie par $\gamma_a(t) = \big(t - a\,\sin(t), 1 - a\,\cos(t)\big)$. Déterminer l'allure de la courbe γ_a au temps $t = 0$.
Solution. Pour tout $a \in \mathbb{R}$, la courbe γ_a est définie, continue et de classe $\mathcal{C}^\infty$ au voisinage de $t = 0$ (et même sur $\mathbb{R}$ tout entier). On a :

$$t - a\,\sin(t) = (1-a)t + a\,\frac{t^3}{6} + \mathrm{o}(t^3) \quad \text{et} \quad 1 - a\,\cos(t) = (1-a) + a\,\frac{t^2}{2} + \mathrm{o}(t^3).$$

Par conséquent :

$$\gamma(t) = \begin{pmatrix} 0 \\ 1-a \end{pmatrix} + t\begin{pmatrix} 1-a \\ 0 \end{pmatrix} + a\,t^2\begin{pmatrix} 0 \\ 1/2 \end{pmatrix} + a\,t^3\begin{pmatrix} 1/6 \\ 0 \end{pmatrix} + t^3\varepsilon(t)$$

avec $\varepsilon(t) \underset{t\to 0}{\longrightarrow} (0,0)$.

La courbe γ_a passe par le point $(0, 1-a)$ au temps $t = 0$. Le cas $a = 0$ étant évident (la courbe est une droite), nous pouvons supposer $a \neq 0$.

- Si $1 - a \neq 0$, c'est-à-dire $a \neq 1$, la dérivée $\gamma_a'(0)$ est non nulle. La courbe a alors un point régulier au temps $t = 0$. La tangente est dirigée par le vecteur $(1-a, 0)$. C'est une tangente horizontale. Le vecteur $\gamma_a''(0)$ est non nul et colinéaire au vecteur $(0, 1/2)$ (le coefficient de t^2). Les vecteurs $\gamma_a'(0)$ et $\gamma_a''(0)$ ne sont donc pas colinéaires et la courbe admet un point birégulier (donc ordinaire) au temps $t = 0$.
- Si $a = 1$, la dérivée $\gamma_1'(0)$ s'annule. La courbe a un point stationnaire au temps $t = 0$. Le coefficient de t^2 est $u_0 := (0, a/2)$ qui est non nul. La courbe admet une tangente dirigée par le vecteur u_0. Cette tangente est l'axe des ordonnées. Le coefficient de t^3 est $v_0 := (a/6, 0)$ qui n'est pas colinéaire à u_0. Les deux plus petits entiers $q > p \geqslant 1$ tels que $\gamma_1^{(p)}(0)$ et $\gamma_1^{(q)}(0)$ ne sont pas colinéaires sont donc $p = 2$ et $q = 3$. Comme p est pair et q impair, la courbe a un point de rebroussement de première espèce au temps $t = 0$.

FIGURE III.2.17 – Courbes cycloïdales :
$a = 1$ en trait plein, $a = 1/2$ en tiret, $a = 3/2$ en pointillé.

Remarque. Ces courbes sont des *courbes cycloïdales* : ce sont les trajectoires d'un point d'une roue lorsque celle-ci se déplace en ligne droite. La roue a un rayon 1 et le point se trouve à distance $|a|$ du centre (elle est « à l'extérieur » de la roue lorsque $|a| > 1$). Le cas $|a| = 1$ correspond à un point de la circonférence de la roue ; la courbe correspondante s'appelle une *cycloïde*.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

III.2.1 Déterminer l'allure des courbes paramétrées suivantes au voisinage de $t = 0$:

$$\gamma_1 : t \mapsto (t^3 + t^4, t^6), \quad \gamma_2 : t \mapsto (t^3, t^5 + t^6), \quad \gamma_3 : t \mapsto \left(\frac{t^2}{1+t^2}, \frac{t^2}{1+t}\right).$$

III.2.2 Déterminer les branches infinies des courbes paramétrées suivantes :

$$\gamma_1 : t \mapsto \left(\frac{e^t}{t+1}, \frac{t^2}{t-1}\right), \quad \gamma_2 : t \mapsto \left(\frac{e^{-t}}{t+1}, \frac{te^{-t}}{t+1}\right), \quad \gamma_3 : t \mapsto \left(t^2 + \frac{2}{t}, t^2 + \frac{1}{t^2}\right).$$

III.2.3 Étudier et tracer la partie de $\mathbb{R}^2$ paramétrée par :

$$x = \frac{t^3}{t^2+t-2} \quad \text{et} \quad y = \frac{t^2-2t}{t-1}.$$

III.2.4 Étudier et tracer la partie de $\mathbb{R}^2$ paramétrée par :

$$x = \sin t \quad \text{et} \quad y = \frac{\cos^2 t}{2-\cos t}.$$

III.2.5 Soit C le cercle de centre 0 et de rayon 1, P le point de coordonnées $(-1,0)$ et M le point de coordonnées $(1,2t)$, $t \in \mathbb{R}$. Soit $M' \neq P$ le point qui se trouve à l'intersection de la droite (PM) avec le cercle C. On dit que M' est le projeté stéréographique de M sur C par rapport à P. Soient (x,y) les coordonnées cartésiennes de M'.

1. On suppose $t \neq 0$. Montrer que $x = y/t - 1$. En déduire que :

$$\frac{y}{t^2}\big((1+t^2)y - 2t\big) = 0.$$

2. Montrer que pour tout $t \in \mathbb{R}$:

$$x = \frac{1-t^2}{1+t^2} \quad \text{et} \quad y = \frac{2t}{1+t^2}.$$

3. En déduire que lorsque t parcourt $\mathbb{R}$ en croissant :

$$\gamma(t) = \left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right)$$

parcourt $C \setminus P$ dans le sens trigonométrique.

Exercices sur la section 2

III.2.6 Tracer les courbes définies en coordonnées polaires par :

1. $\rho(\theta) = 1 + \cos(\theta)$,
2. $\rho(\theta) = \sqrt{2\cos(2\theta)}$
3. $\rho(\theta) = 1 + \cos(3\theta) + \sin^2(3\theta)$.

III.2.7 Montrer que la droite d'équation cartésienne $\alpha x + \beta y = 1$, avec $(\alpha,\beta) \neq (0,0)$, admet le paramétrage en coordonnées polaires :

$$\rho(\theta) = \frac{1}{\alpha\cos(\theta) + \beta\sin(\theta)}.$$

III.2.8 Montrer que le cercle centré en $(\alpha,\beta) \neq (0,0)$ qui passe par l'origine $(0,0)$ admet le paramétrage en coordonnées polaires :

$$\rho(\theta) = 2\alpha\cos(\theta) + 2\beta\sin(\theta).$$

III.2.9 Soit p et e deux réels strictement positifs. Considérons la courbe définie en coordonnées polaires par :

$$\rho(\theta) = \frac{p}{1 + e\cos(\theta)}.$$

Montrer que selon que $e = 1$, $e < 1$ ou $e > 1$, la courbe est une parabole, une ellipse ou une hyperbole dont l'excentricité est e, dont un foyer est l'origine, la directrice associée étant la droite définie en coordonnées polaires par :

$$\rho(\theta) = \frac{p}{e\cos(\theta)}.$$

Exercices sur la section 3

III.2.10 Calculer la longueur des courbes suivantes :

1. la courbe paramétrée par $t \in [0,1] \mapsto (3t, 3t^2)$,
2. la courbe paramétrée par $t \in [0,\pi] \mapsto (\cos t + \cos^2 t, \sin t + \sin t \cos t)$.

III.2.11 Soit $x : I \to \mathbb{R}$ et $y : I \to \mathbb{R}$ deux fonctions de classe $\mathcal{C}^2$ et $\gamma : I \to \mathbb{R}^2$ la courbe paramétrée définie par $\gamma(t) = \big(x(t), y(t)\big)$. Montrer que la courbure de γ est donnée par $\kappa = \dfrac{x'y'' - y'x''}{(x'^2 + y'^2)^{3/2}}$.

III.2.12 Soit $f : I \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ et $\gamma : I \to \mathbb{R}^2$ la courbe représentative de f définie par $\gamma(t) = \big(t, f(t)\big)$. Montrer que la courbure de γ est donnée par $\kappa = \dfrac{f''}{(1 + f'^2)^{3/2}}$.

Exercices sur la section 4

III.2.13 Soit $\gamma : \mathbb{R} \to \mathbb{R}^3$ une courbe de classe $\mathcal{C}^3$.

1. Montrer que la courbure est donnée par $\kappa = \dfrac{\|\gamma' \wedge \gamma''\|}{\|\gamma'\|^3}$.
2. Montrer que la torsion est donnée par $\tau = \dfrac{\big(\gamma' \wedge \gamma'' \mid \gamma'''\big)}{\|\gamma' \wedge \gamma''\|^2}$.

III.2.14
Déterminer la courbure et la torsion de la courbe $\gamma : \mathbb{R} \to \mathbb{R}^3$ définie par $\gamma(t) = (t, t^2, t^3)$.

III.2.15 Soient $\kappa : I \to]0, +\infty[$ et $\tau : I \to \mathbb{R}$ deux fonctions continues, soit $s_0 \in \mathring{I}$, soit $\gamma_0 \in \mathbb{R}^2$ et soient $u_0 \in \mathbb{R}^3$ et $v_0 \in \mathbb{R}^3$ deux vecteurs unitaires orthogonaux.

1. Soit $w_0 = u_0 \wedge v_0$ l'unique vecteur de $\mathbb{R}^3$ tel que (u_0, v_0, w_0) soit une base orthonormée directe de $\mathbb{R}^3$. Soit $s \mapsto \big(\overrightarrow{T}(s), \overrightarrow{N}(s), \overrightarrow{B}(s)\big) \in \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3$ la solution maximale de l'équation différentielle linéaire à coefficients continus :

$$Y' = \begin{pmatrix} 0 & \kappa(s) & 0 \\ \kappa(s) & 0 & -\tau(s) \\ 0 & \tau(s) & 0 \end{pmatrix} \cdot Y$$

qui vérifie les conditions initiales $\overrightarrow{T}(s_0) = u_0$, $\overrightarrow{N}(s_0) = v_0$ et $\overrightarrow{B}(s_0) = w_0$. Soit G la fonction vectorielle définie par :

$$G := \big((\overrightarrow{T} \mid \overrightarrow{T}), (\overrightarrow{T} \mid \overrightarrow{N}), (\overrightarrow{T} \mid \overrightarrow{B}), (\overrightarrow{N} \mid \overrightarrow{N}), (\overrightarrow{N} \mid \overrightarrow{B}), (\overrightarrow{B} \mid \overrightarrow{B})\big).$$

(a) Montrer que G est solution d'une équation différentielle linéaire à coefficients continus.

(b) Montrer que la fonction constante égale à $(1, 0, 0, 1, 0, 1)$ est solution de la même équation différentielle.

(c) En déduire que pour tout $s \in J$, $\big(\overrightarrow{T}(s), \overrightarrow{N}(s), \overrightarrow{B}(s)\big)$ est une base orthonormée de $\mathbb{R}^3$.

(d) Montrer que le déterminant de la matrice $\big(\overrightarrow{T}(s), \overrightarrow{N}(s), \overrightarrow{B}(s)\big)$ est une fonction continue qui ne peut prendre que les valeurs $+1$ ou -1.

(e) Calculer la valeur de ce déterminant en s_0.

(f) En déduire que la base $\big(\overrightarrow{T}(s), \overrightarrow{N}(s), \overrightarrow{B}(s)\big)$ est orthonormée directe.

(g) Montrer que $J = I$ et que la courbe $\gamma : I \to \mathbb{R}^3$ définie par :

$$\gamma(s) = \gamma_0 + \int_{s_0}^{s} \overrightarrow{T}(t)\,\mathrm{d}t$$

est un paramétrage normal birégulier de classe $\mathcal{C}^3$ de courbure κ et de torsion τ.

2. Supposons maintenant qu'il existe deux paramétrages normaux biréguliers $\gamma_1 : I \to \mathbb{R}^3$ et $\gamma_2 : I \to \mathbb{R}^3$ de classe $\mathcal{C}^3$, de courbure κ, de torsion τ, tels que :

$$\gamma_1(s_0) = \gamma_2(s_0) = \gamma_0, \qquad \gamma_1'(s_0) = \gamma_2'(s_0) = u_0 \quad \text{et} \quad \gamma_1''(s_0) = \gamma_2''(s_0) = \kappa(s_0) \cdot v_0.$$

Soient $(\gamma_1, \overrightarrow{T}_1, \overrightarrow{N}_1, \overrightarrow{B}_1)$ et $(\gamma_2, \overrightarrow{T}_2, \overrightarrow{N}_2, \overrightarrow{B}_2)$ les repères de Serret-Frenet associés. Soit $f : \overset{\circ}{I} \to \mathbb{R}$ la fonction définie par :

$$f = (\overrightarrow{T}_1 \mid \overrightarrow{T}_2) + (\overrightarrow{N}_1 \mid \overrightarrow{N}_2) + (\overrightarrow{B}_1 \mid \overrightarrow{B}_2).$$

(a) Montrer que $f(s_0) = 3$ et que f' est identiquement nulle. En déduire que f est constante égale à 3.

(b) Montrer que $(\overrightarrow{T}_1 \mid \overrightarrow{T}_2) \leqslant 1$, $(\overrightarrow{N}_1 \mid \overrightarrow{N}_2) \leqslant 1$ et $(\overrightarrow{B}_1 \mid \overrightarrow{B}_2) \leqslant 1$ avec égalité si, et seulement si, $\overrightarrow{T}_1 = \overrightarrow{T}_2$ et $\overrightarrow{N}_1 = \overrightarrow{N}_2$ et $\overrightarrow{B}_1 = \overrightarrow{B}_2$. En déduire que $\overrightarrow{T}_1 = \overrightarrow{T}_2$ et que $\gamma_1 = \gamma_2$.

Partie IV

Analyse

L'analyse, science des infinis et des limites, est née vers 1650. Après Cavalieri parlant d'indivisibles (1635), Descartes, Fermat et Roberval déterminèrent des tangentes, puis Pascal écrivit sur la cycloïde en calculant des aires. Influencés par ces travaux, Newton (1665), puis indépendamment Leibniz (1675), créèrent les calculs différentiel et intégral.

Quelques bribes d'une analyse plus fruste les précédèrent : propriétés des nombres réels dues à Eudoxe, réduction en fraction continue chez Euclide, calculs archimédiens d'aires et d'une tangente (grâce à la *neusis*, insertion entre deux arcs d'un segment de longueur donnée de support passant par un point connu), ébauche de géométrie différentielle arabe vers le Xe siècle.

L'âge d'or de l'analyse classique débute donc à la fin du XVIIe siècle, justifié par la cinématique. Au siècle suivant, les Bernoulli et Euler, calculateur talentueux, forgeront des outils toujours efficaces (fonctions transcendantes, résolutions exactes ou approchées des équations différentielles...). Puis Fourier, étudiant des problèmes de physique, introduira une analyse dont l'importance théorique et pratique ne saurait être surestimée.

La formalisation de l'analyse, secondaire pour Euler plus préoccupé d'avancer que de fonder les calculs dans un cadre général, sera surtout l'œuvre du XIXe siècle : Cauchy, Bolzano, Dirichlet, Dedekind, Stokes, Weierstrass. L'intégrale a été intuitive au départ, affermie et élargie par Cauchy et Riemann, puis au siècle suivant par Lebesgue et enfin Denjoy et Perron (dont les travaux ont été simplifiés récemment par Kurzweil et Henstock).

Vinrent ensuite de fécondes extensions hors des fonctions réelles ou complexes d'une ou plusieurs variables : l'analyse fonctionnelle (Riemann, Hilbert, Volterra, Fredholm, Fréchet, Riesz, Banach, von Neumann) où les fonctions sont vues comme des points, motivée par des problèmes variationnels, puis la mécanique quantique, les espaces vectoriels topologiques ou normés (Hilbert, Banach), les espaces de Sobolev, les distributions (L. Schwarz), rendant l'analyse universelle. Ces techniques ont aujourd'hui, dans le cadre de la théorie des équations aux dérivées partielles, d'importantes applications industrielles et en physique et chimie (développées en particulier par l'école française autour de J.-L. Lions).

L'étudiant devra d'abord *apprendre à calculer* : dériver, intégrer, manipuler suites et séries, tracer des courbes. Il est totalement *illusoire* de penser que les progrès des ordinateurs et l'existence de logiciels performants de calcul numérique ou formel dispensent de cet apprentissage « sur le tas ». Il devra aussi se former graduellement à l'apprentissage de la rigueur, cruciale en analyse, à laquelle il a été peu préparé par l'enseignement secondaire.

Nombres réels suites numériques

IV.1

L'ensemble des nombres rationnels $\mathbb{Q}$ est de construction aisée (*cf.* page 99), mais il n'est pas adapté pour certaines opérations mathématiques nécessaires en géométrie élémentaire, ni pour toutes celles que nous regroupons de nos jours dans le domaine de *l'analyse*. Il lui « manque » certains nombres « simples », comme la racine de 2 ou π ou la base e des logarithmes népériens. En un certain sens « il est plein de trous », comme nous allons l'expliquer.

Si nous partons de l'idée intuitive de « droite D de la géométrie plane élémentaire », il est facile « d'identifier » $\mathbb{Q}$ à une partie Q de D en reprenant, pour l'essentiel, ce que faisaient les mathématiciens de l'antiquité grecque. On choisit une origine O dans D, une unité de longueur L et une orientation et l'on reporte sur D, à partir de O dans un sens ou dans l'autre, toutes les longueurs *commensurables à l'unité de longueur*[1] en marquant l'extrémité sur D. Mais l'on peut aussi reporter des longueurs « obtenues géométriquement » mais incommensurables à l'unité de longueur, comme la diagonale d'un carré de côté L, celle d'un pentagone régulier de même coté ou la longueur d'un cercle de diamètre L (en utilisant un fil !). On obtient ainsi de nouveaux points de D n'appartenant pas à Q, mais on peut penser qu'il y en a bien d'autres. Intuitivement D est un objet « continu, sans lacunes », tandis que Q est « discontinu ». Pour rendre compte des propriétés intuitives de la droite géométrique, il apparaît donc nécessaire de définir de nouveaux nombres, appelés *nombres réels irrationnels*.

Nous allons introduire un ensemble contenant $\mathbb{Q}$, mais plus « complet » noté $\mathbb{R}$ et appelé le *corps des nombres réels*. Il permettra de « bien représenter » une droite géométrique et nous verrons qu'il a bien d'autres avantages.

Toute construction de $\mathbb{R}$ est faite à partir de $\mathbb{Q}$, mais elle est beaucoup plus délicate que celle de $\mathbb{Q}$ à partir de $\mathbb{Z}$ ou celle de $\mathbb{Z}$ à partir de $\mathbb{N}$. Historiquement les premières constructions rigoureuses de $\mathbb{R}$ ont été obtenues à fin du XIXe siècle, par Richard Dedekind, Charles Meray, Georg Cantor... (*cf.* ci-dessous l'introduction de la partie 3).

[1] Rappelons que, selon les mathématiciens de l'antiquité, deux longueurs sont dites commensurables s'il existe une longueur permettant de les mesurer simultanément. Il est équivalent, en termes modernes, de dire que leur rapport est un nombre rationnel.

Dans ce module nous commencerons par donner une *présentation axiomatique* de $\mathbb{R}$ basée sur un résultat d'existence *admis* (comme nous l'avons fait plus haut pour $\mathbb{N}$). Cette présentation *suffit pour étudier l'analyse* (et pour poursuivre la lecture du livre et des suivants L2 et L3...).
Pour le lecteur souhaitant approfondir les choses, nous donnerons ensuite, à la fin de ce module, les grandes lignes d'une *construction* de $\mathbb{R}$ basée sur la notion de *nombre décimal*. Dans le livre L3, nous complétons cette construction et nous présentons les autres constructions classiques de $\mathbb{R}$.

1 Le corps des nombres réels

1.1 Bornes inférieures et supérieures

On va utiliser de façon essentielle la notion de *relation d'ordre* (*cf.* page 33 la section 5.3 du module I.1). Rappelons que $\mathbb{Q}$ possède une relation d'ordre total (pour laquelle les notions d'élément minimal et de minimum sont confondues). Dans cet ensemble ordonné, une partie non minorée n'a pas de minimum ; il en va de même pour l'ensemble vide. Mais même un ensemble non vide et minoré peut ne pas avoir de minimum.

Exercice 1.
L'ensemble $A := \{x \in \mathbb{Q} \mid 0 < x\}$ est minoré, mais n'a pas de minimum.
Solution. Par définition, le nombre 0 minore A. On raisonne par l'absurde. On suppose que a est le minimum de A. Alors $a > 0$ et, comme a est rationnel, il en est de même de $a/2$, d'où les inégalités contradictoires :

$$a \leqslant \frac{a}{2} < a.$$

D'une façon générale, tout intervalle ouvert $]a, b[:= \{x \in \mathbb{Q} \mid \ a < x < b\}$ de $\mathbb{Q}$ est à la fois minoré et majoré (par a et b respectivement), sans avoir de minimum ni de maximum. Ainsi, la notion d'élément minimum ou maximum, qui est très utile dans les ensembles d'entiers, l'est beaucoup moins dans les ensembles de rationnels ou dans d'autres ensembles ordonnés. Dans de tels cas, c'est plutôt la notion de *borne inférieure ou supérieure* qui est pertinente.

Définition 1. On dit qu'un sous-ensemble non vide A d'un ensemble ordonné *admet une borne inférieure* s'il est minoré et s'il existe un minorant de A plus grand que tous les autres ; on le note alors $\inf A$.
On définit de manière similaire une borne supérieure (éventuelle), qui est le plus petit des majorants ; on la note $\sup A$.

Donc $\inf A$ est *le plus grand des minorants de* A. Lorsqu'il existe, il est unique, puisque c'est le maximum d'un ensemble (*cf.* page 34).

Exercice 2.
Montrer que si A a un minimum, alors il a aussi une borne inférieure qui est égale à $\min A$ (et une propriété similaire pour l'élément maximum éventuel).

Solution. Soit $a := \min A$. Tout minorant m de A vérifie notamment l'inégalité $m \leqslant a$ puisque $a \in A$. Or a est lui-même un minorant de A puisque $a \leqslant b$ pour tout $b \in A$; majorant tous les autres minorants, c'est donc bien le plus grand d'entre eux.
Il en va de même pour le maximum.

Soient a et b deux rationnels tels que $a < b$. L'intervalle $]a, b[$ de $\mathbb{Q}$ admet exactement a et b comme bornes respectivement inférieure ou supérieure, comme le bon sens le suggère ; et il en va de même des intervalles fermés de l'un ou l'autre côté, ou des deux. Montrons par exemple que a est bien le plus petit minorant de $]a, b[$. C'est déjà un minorant, par définition. Soit $\alpha \in \mathbb{Q}$ tel que $\alpha > a$. Alors il existe $c \in]a, b[$ tel que $c < \alpha$: on peut prendre $c := (a + \alpha)/2$ si $\alpha \leqslant b$ ou $c := (a + b)/2$ sinon. Donc α n'est pas un minorant de $]a, b[$. Comme l'ordre sur $\mathbb{Q}$ est total, on en déduit que tout minorant de $]a, b[$, qui ne peut donc pas être strictement supérieur à a, est inférieur ou égal à a et donc que a est bien le plus grand des minorants de $]a, b[$.
Mais une partie non vide et majorée de $\mathbb{Q}$ n'admet pas nécessairement une borne supérieure comme le montre l'exemple ci-dessous. On rappelle la notation $\mathbb{Q}_+ := \{x \in \mathbb{Q} \mid x \geqslant 0\}$.

Exercice 3.

Montrer que $A := \{x \in \mathbb{Q}_+ \mid x^2 < 2\}$ est non vide et borné, mais ne possède pas de borne supérieure.

Solution. L'ensemble A est non vide car $0 \in A$.
Il est borné, par exemple par 10, car $|x| \geqslant 10$ implique $x^2 \geqslant 100 > 2$.
On note $A' := \{x \in \mathbb{Q}_+ \mid x^2 > 2\}$. On a $A \cap A' = \varnothing$ et $A \cup A' = \mathbb{Q}_+$ car il n'y a pas de rationnel dont le carré est égal à 2 (*cf.* la proposition 66 de la page 105).
La preuve que nous allons donner utilise des propriétés de la fonction $f : \mathbb{Q}_+ \to \mathbb{Q}_+$ définie par $f : x \mapsto \frac{2x+2}{x+2}$. Des calculs élémentaires donnent, en notant $y := f(x)$:

$$y - x = \frac{2 - x^2}{x + 2} \quad \text{et} \quad y^2 - 2 = 2\frac{x^2 - 2}{(x + 2)^2}.$$

Ainsi $x^2 - 2$ et $f(x)^2 - 2$ sont de *même signe* et $f(x) - x$ est du *signe opposé* à celui de $x^2 - 2$. Donc, si $x \in A$, alors $f(x) \in A$ et $f(x) > x$ et si $x \in A'$, alors $f(x) \in A'$ et $f(x) < x$.
On montre le résultat cherché en raisonnant par l'absurde. On suppose que A admet une borne supérieure dans $\mathbb{Q}$, et on la note a. Puisque $0 \in A$, on a $a \geqslant 0$. On note $b := f(a) = \frac{2a+2}{a+2}$. On a $a \in A$ ou $a \in A'$. Montrons que les deux cas sont impossibles et que l'on aboutit à une contradiction.

- Supposons $a \in A$. Alors $b \in A$ et $b - a > 0$, ce qui contredit le fait que a est un majorant de A.
- Supposons $a \in A'$. Alors $b \in A'$ et $b - a < 0$, donc b n'est pas un majorant de A puisque a est le plus petit des majorants. Il existe donc $c \in A$ tel que $b < c$ et l'on a $b^2 < c^2 < 2$, donc $b \in A$. Mais c'est impossible puisque $A \cap A' = \varnothing$.

L'application f utilisée ci-dessus s'appelle une homographie.
Nous l'étudierons plus loin (*cf.* l'exemple 1 de la page 536).

Cette absence de borne supérieure matérialise le fait qu'il manque, dans $\mathbb{Q}$, un élément dont le carré vaut 2. Intuitivement, $\mathbb{Q}$ a un « trou » à la position correspondante : les carrés des rationnels passent de valeurs strictement inférieures à 2 à des valeurs strictement supérieures « sans continuité ». Le fait qu'il existe un ensemble totalement ordonné de « nombres » dépourvu de ces « trous » (contenant $\mathbb{Q}$) est loin d'être évident : c'est tout l'objet de *la construction de l'ensemble des nombres réels*. Dans $\mathbb{R}$, le problème disparaîtra : l'ensemble $S := \{x \in \mathbb{R} \mid x^2 < 2\}$ aura une borne inférieure et une borne supérieure, notées respectivement $-\sqrt{2}$ et $\sqrt{2}$ et l'on aura $\sqrt{2}^2 = (-\sqrt{2})^2 = 2$.

Exercice 4.
Soient A et B deux parties d'un même ensemble ordonné vérifiant $A \subset B$.

1. Montrer que tout minorant de B est aussi un minorant de A ; que peut-on dire pour les majorants ?
2. Montrer que si A et B ont chacun un élément minimum, alors $\min A \geqslant \min B$; que peut-on dire pour les plus grands éléments ?
3. Montrer que si A et B ont chacun une borne inférieure, alors $\inf A \geqslant \inf B$; que peut-on dire pour les bornes supérieures ?

Solution.

1. Tout élément a de A appartient à B et vérifie donc $a \geqslant m$ pour tout minorant de B : par suite m minore A.
   De la même façon les majorants de B sont aussi majorants de A.
2. Si $\mu := \min A$ et $\nu := \min B$, alors $\mu \in A$ et $\nu \leqslant \mu$, soit $\min B \leqslant \min A$.
   De la même façon, $\max B \geqslant \max A$.
3. Dans ce cas, un peu plus général, $\inf B$ est un minorant de B, donc un minorant de A, et il est inférieur ou égal au plus grand des minorants de A, c'est-à-dire $\inf A$.
   De la même façon, $\sup B \geqslant \sup A$.

Exercice 5.
Soit A une partie d'un ensemble ordonné E admettant une borne inférieure $m = \inf A$. Montrer que pour tout $x \in E$, $x > m$, il existe un élément $a \in A$ tel que $a < x$. Écrire la propriété similaire pour la borne supérieure.
Solution. Par définition, m est le plus grand de tous les minorants de A. Par suite x ne peut être un minorant, c'est donc qu'il existe un a dans A tel que $x > a$.
De même, pour tout ensemble non vide A admettant une borne supérieure $M = \sup A$ et pour tout nombre $y < M$, il existe un élément $b \in A$ tel que $b > y$.

1.2 Le corps des nombres réels

1.2.1 Axiomatique de $\mathbb{R}$

Dans cette partie nous donnons une présentation *axiomatique* des nombres réels. Les propriétés de ces nombres sont déduites d'un « postulat d'existence » (*cf.* la définition 5 et le

postulat qui la suit). La démarche est analogue à celle présentée plus haut pour l'ensemble $\mathbb{N}$ des entiers (*cf.* page 63) et dans la ligne de celle utilisée par le mathématicien grec Euclide (vers 300 avant J.C.) pour mettre en place la géométrie à partir de cinq *postulats* admis.
En fait le postulat d'existence des nombres réels peut être *démontré* dans le cadre de la théorie des ensembles. C'est l'objet de la partie 3 ci-dessous où nous présentons les grandes lignes d'une « construction du corps des nombres réels ». Cette partie est assez abstraite et un peu technique et elle peut être omise en première lecture.

La présentation axiomatique des nombres réels est basée sur la notion de corps totalement ordonné que nous allons rappeler (*cf.* la section 5.3 de la page 33 du module I.1).

Définition 2. Un ensemble E est dit totalement ordonné si deux éléments $x,\ y \in E$ quelconques sont comparables.

Définition 3. Un *anneau totalement ordonné* est un anneau $\mathcal{A}$ muni d'une relation d'ordre total $\preceq$ compatible avec les opérations, c'est-à-dire vérifiant, pour tous a, b, c et d de $\mathcal{A}$:

$$(a \preceq b \quad \text{et} \quad c \preceq d) \Longrightarrow a + c \preceq b + d$$

$$(0 \preceq a \preceq b \quad \text{et} \quad 0 \preceq c \preceq d) \Longrightarrow a\,c \preceq b\,d.$$

Un *corps totalement ordonné* est un corps $\mathcal{R}$ dont l'anneau sous-jacent est totalement ordonné.

Un élément x d'un anneau totalement ordonné est dit positif (resp. négatif) si $0 \preceq x$ (resp. $x \preceq 0$). Il est dit strictement positif (resp. strictement négatif) s'il est positif (resp. négatif) et différent de 0.

Dans un anneau totalement ordonné, la multiplication par un nombre strictement négatif change le sens des inégalités. En particulier, on ne peut multiplier des inégalités terme à terme que si tous les réels intervenant dans ces inégalités sont positifs.

Proposition 1. Soit $\mathcal{A}$ un anneau totalement ordonné. Alors :

(i) pour tout $x \in \mathcal{A}$, x^2 est positif ;

(ii) 1 est strictement positif ;

(iii) pour tout $n \in \mathbb{N}^*$, $n.1 := 1 + \cdots + 1 > 0$; l'anneau $\mathcal{A}$ est de *caractéristique nulle*.

Démonstration.

(i) Si x est positif, en multipliant l'inégalité $0 \preceq x$ par x, on obtient $0 \preceq x^2$. Si x est négatif, on se ramène au cas précédent en utilisant $x^2 = (-x)^2$.

(ii) Résulte de (i), car $1 = 1^2$ et $1 \neq 0$.

(iii) En utilisant (ii), on montre par récurrence que $n.1 > 0$, donc $n.1 \neq 0$ et l'anneau $\mathcal{A}$ est de *caractéristique nulle* (*cf.* le théorème et définition 39 de la page 142). ∎

Corollaire 2.

(i) Il existe une et une seule relation d'ordre sur $\mathbb{Z}$ qui fasse de $\mathbb{Z}$ un anneau totalement ordonné.

(ii) Il existe une et une seule relation d'ordre sur $\mathbb{Q}$ qui fasse de $\mathbb{Q}$ un corps totalement ordonné.

(iii) Soit $(\mathcal{R}, \preceq)$ un corps totalement ordonné contenant $\mathbb{Q}$ comme sous-corps. Pour tous $x,\ y \in \mathbb{Q}$, $x \preceq y$ si, et seulement si, $x \leqslant y$. Autrement dit, la relation d'ordre $\preceq$ sur $\mathcal{R}$ équivaut sur $\mathbb{Q}$ à la relation d'ordre ordinaire $\leqslant$.

Démonstration. Cette preuve peut être omise en première lecture.

(i) Soit $\preceq$ une relation d'ordre sur $\mathbb{Z}$ qui en fait un anneau totalement ordonné. On note $\mathcal{Z}_+ = \{x \in \mathbb{Z} \mid 0 \preceq x\}$. D'après (iii) de la proposition 1, $\mathbb{N} \subset \mathcal{Z}_+$ donc $-\mathbb{N} \subset -\mathcal{Z}_+$ et $\mathbb{Z}_+ = \mathbb{N} = \mathcal{Z}_+$. Par suite les éléments positifs sont les mêmes pour $\preceq$ et pour la relation d'ordre usuelle. On en déduit que ces deux relations coïncident.

(ii) Soit $\preceq$ une relation d'ordre total sur $\mathbb{Q}$ qui en fait un corps totalement ordonné. Pour montrer que $\preceq$ coïncide avec la relation d'ordre ordinaire $\leqslant$, il suffit de montrer que les éléments *positifs* sont les mêmes pour les deux relations.

Soit $x \in \mathbb{Q}$. Montrons que $x \leqslant 0 \Rightarrow 0 \preceq x$. D'après (i), c'est vrai pour $x \in \mathbb{Z}$. Soit $x := 1/n$, avec $n \in \mathbb{N}^*$. On ne peut pas avoir $\frac{1}{n} \preceq 0$. Sinon, l'on aurait $1 = n\frac{1}{n} \preceq 0$, ce qui est impossible. Donc, puisque la relation d'ordre $\preceq$ est totale : $0 \preceq \frac{1}{n}$.

Soit $x \in \mathbb{Q}$ tel que $0 \leqslant x$. On a $x := m/n$, avec $m \in \mathbb{N}$ et $n \in \mathbb{N}^*$. De $0 \preceq m$ et $0 \preceq \frac{1}{n}$, l'on déduit par multiplication des inégalités $0 \preceq x$.

Soit inversement $x \in \mathbb{Q}$ tel que $0 \preceq x$. Si $x \neq 0$, on montre, en raisonnant par l'absurde, que $0 < x$. Supposons, en effet, que $x < 0$. Alors $0 < -x$, donc $0 \preceq -x$ et $x \preceq 0$. Puisque $0 \preceq x$, on a $x = 0$ et une contradiction. Ainsi, dans tous les cas : $0 \leqslant x$.

(iii) Cette assertion résulte immédiatement de (ii). ∎

Pour un corps totalement ordonné contenant $\mathbb{Q}$ comme sous-corps, on peut donc noter, sans inconvénient, la relation d'ordre $\leqslant$, ce que nous ferons dans la suite.

Définition 4. On dit qu'un ensemble ordonné possède la propriété de la borne supérieure (resp. inférieure) si toute partie non vide majorée (resp. minorée) admet une borne supérieure (resp. inférieure).

Définition 5 (Axiomes des nombres réels). On dit qu'un ensemble vérifie les axiomes des nombres réels s'il est muni d'une structure de corps totalement ordonné contenant $\mathbb{Q}$ comme sous-corps et possédant la propriété de la borne supérieure.

Postulat d'existence des nombres réels. On *admet l'existence* d'un ensemble vérifiant les axiomes des nombres réels.

Notons provisoirement $\mathcal{R}$ un tel ensemble. Par passage à l'opposé, on montre que toute partie non vide minorée de $\mathcal{R}$ admet une borne inférieure.

Remarques.

- On a ainsi « complété » le corps $\mathbb{Q}$ pour obtenir des propriétés dont nous avons mis en évidence l'absence sur le corps des rationnels.

- Le lecteur pourra trouver à la page 581 les grandes lignes d'une construction d'un tel corps.
- On peut démontrer l'« unicité » d'un tel corps $\mathcal{R}$ dans le sens suivant : si $\mathcal{R}_1$ et $\mathcal{R}_2$ sont deux corps totalement ordonnés, contenant $\mathbb{Q}$ comme sous-corps, et vérifiant la propriété de la borne supérieure, alors il existe un isomorphisme d'anneaux croissant φ de $\mathcal{R}_1$ sur $\mathcal{R}_2$ (on dit aussi que φ est un isomorphisme d'anneaux totalement ordonnés). Nous donnerons une démonstration dans une prochaine édition de L3). L'application φ est un isomorphisme de corps et induit l'identité sur $\mathbb{Q}$. On peut de plus montrer qu'un tel isomorphisme est nécessairement *unique* (*cf.* l'exercice IV.1.7).
  L'existence d'un tel isomorphisme unique montre que les propriétés de $\mathcal{R}_1$ et de $\mathcal{R}_2$ concernant leurs relations d'ordre respectives et leurs opérations d'anneaux sont identiques. On ne change donc rien en remplaçant l'un par l'autre et l'on peut donc en fixer un arbitrairement choisi :
- On notera désormais $\mathbb{R}$ un tel corps. On l'appellera *le corps des réels*.
- Dans certains ouvrages on donne une version *équivalente* des axiomes des nombres réels de la définition 5 ci-dessus en remplaçant la propriété de la borne supérieure par la propriété d'Archimède (*cf.* la proposition 10) *et* la propriété dite des segments emboîtés (*cf.* la proposition 5 et l'exercice IV.1.1). Cette version est plus proche de l'idée intuitive de « boucher les trous » de $\mathbb{Q}$.

Exemple : existence d'un irrationnel. Reprenons l'exercice 3 de la page 511. Le sous-ensemble $A := \{x \in \mathbb{Q}^+ \mid x^2 < 2\}$ de $\mathbb{R}$ est non vide et borné. Il admet donc, *dans* $\mathbb{R}$, une borne supérieure a. En utilisant les propriétés de la fonction $f : \mathbb{R}_+ \to \mathbb{R}_+$ définie par $f : x \mapsto \frac{2x+2}{x+2}$ et le réel $b := f(a)$, on montre en utilisant *mutatis mutandis* les arguments de l'exercice 3 de la page 511 que a ne peut pas vérifier $a^2 < 2$, ni $a^2 > 2$. Donc $a^2 = 2$. Ainsi a est un réel positif dont le carré est 2. On le note $\sqrt{2}$. Ce n'est pas un rationnel car nous avons vu qu'il n'existe pas de rationnel de carré deux. On a aussi $(-\sqrt{2})^2 = 2$ et $-\sqrt{2} \neq \sqrt{2}$. Les seuls réels de carré deux sont $\pm\sqrt{2}$; en effet, d'après le théorème 21 de la page 297, le polynôme réel $X^2 - 2$ admet *au plus deux racines*.
La caractérisation suivante de la borne supérieure dans $\mathbb{R}$ est très utile.

Proposition 3. Pour tout sous-ensemble majoré non-vide $A \subset \mathbb{R}$, $b \in \mathbb{R}$ est la borne supérieure de A si, et seulement si, b vérifie les deux conditions suivantes :

(i) b est un majorant de A ;

(ii) pour tout $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$, il existe un élément $x \in A$ tel que $b - \varepsilon < x \leqslant b$.

Démonstration. Supposons que $b \in \mathbb{R}$ soit la borne supérieure de A. Soit $\varepsilon > 0$. Puisque b est un majorant de A, s'il n'existait pas d'élément $x \in A$ tel que $b - \varepsilon < x \leqslant b$, $b - \varepsilon$ serait un majorant de A strictement plus petit que b, d'où une contradiction.
Inversement supposons que $b \in \mathbb{R}$ vérifie les conditions (i) et (ii). Supposons que la borne supérieure b' de A soit strictement plus petite que b. Soit $\varepsilon := \frac{b-b'}{2} > 0$. D'après (ii), il existe $x \in A$ tel que $b - \varepsilon < x \leqslant b$. On a $b' < x$, d'où une contradiction. ∎

Il existe une caractérisation similaire pour la borne inférieure. La preuve est analogue.

Proposition 4. Pour tout sous-ensemble minoré non-vide $A \subset \mathbb{R}$, $b \in \mathbb{R}$ est la borne inférieure de A si, et seulement si, b vérifie les deux conditions suivantes :

(i) b est un minorant de A ;

(ii) pour tout $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$, il existe un élément $x \in A$ tel que $b \leqslant x < b + \varepsilon$.

Notation. Lorsque a et b appartenant à $\mathbb{R}$ vérifient $a \leqslant b$, on note :

$$[a, b] := \{x \in \mathbb{R} \mid a \leqslant x \leqslant b\}.$$

Les ensembles $[a, b]$, avec $a \leqslant b$, sont ce que l'on appelle les *segments* de $\mathbb{R}$.

Proposition 5. Soit $(S_n)_{n\in\mathbb{N}}$ une suite de segments de $\mathbb{R}$ décroissante pour l'inclusion, c'est-à-dire vérifiant : $\forall (p, q) \in \mathbb{N}^2$, $p \leqslant q \Longrightarrow S_q \subset S_p$.
Alors l'intersection $\bigcap\limits_{n\in\mathbb{N}} S_n$ n'est pas vide.

On dit que les segments sont emboîtés et que $\mathbb{R}$ possède la propriété des segments emboîtés.

Démonstration. Pour tout $n \in \mathbb{N}$, il existe deux réels a_n et b_n, avec $a_n \leqslant b_n$, tels que $S_n = [a_n, b_n]$. Considérons l'ensemble $A := \{a_n \mid n \in \mathbb{N}\}$. C'est une partie non vide de $\mathbb{R}$ majorée par b_0 puisque l'inclusion $S_n \subset S_0$ se traduit par les inégalités $a_0 \leqslant a_n \leqslant b_n \leqslant b_0$. Elle admet donc une borne supérieure $\alpha \in \mathbb{R}$. De la même façon, l'ensemble $B := \{b_n \mid n \in \mathbb{N}\}$ admet une borne inférieure β.
Or, tout élément de B majore A et tout élément de A minore B puisque, pour tout $(p, q) \in \mathbb{N}^2$, on a, en notant $n := \max(p, q)$, $a_p \leqslant a_n \leqslant b_n \leqslant b_q$. On en déduit que α, le plus petit des majorants de A, est inférieur ou égal à tous les éléments de B. C'est donc un minorant de B et, par conséquent, il est inférieur ou égal à sa borne inférieure β. Le segment $[\alpha, \beta]$ est donc une partie non vide de $\mathbb{R}$.
Ainsi, on a $\forall n \in \mathbb{N}$, $a_n \leqslant \alpha \leqslant \beta \leqslant b_n$ ce qui prouve que le segment $[\alpha, \beta]$ est inclus dans tous les S_n donc dans leur intersection qui par suite est bien non vide. ■

La proposition ci-dessous dit que, contrairement à $\mathbb{Q}$, $\mathbb{R}$ *n'est pas dénombrable*. On la déduit de la proposition 5.

Proposition 6. L'ensemble $\mathbb{R}$ des nombres réels n'est pas dénombrable.

Démonstration.

(i) On prouve d'abord un résultat préliminaire. Soit $x = (x_n)_{n\in\mathbb{N}^*}$ une suite de $[0, 1]$ (c'est-à-dire une application $x : \mathbb{N}^* \to [0, 1]$ telle que $\forall n \in \mathbb{N}^*$, $x(n) = x_n$). Nous allons montrer qu'il existe une suite $\{I_n\}_{n\in\mathbb{N}^*}$ de segments emboîtés contenus dans $[0, 1]$ telle que, pour tout $n \in \mathbb{N}^*$, $x_n \notin I_n$.
On procède par récurrence. On écrit $[0, 1]$ comme réunion de trois segments de longueur $1/3$: $[0, 1] = [0, 1/3] \cup [1/3, 2/3] \cup [2/3, 1]$. L'intersection de ces trois segments est vide. Il existe donc au moins un de ces segments qui ne contient pas x_1 ; on le note I_1. On écrit ensuite I_1 comme réunion de trois segments de longueur $1/3^2$. Il existe au moins un de ces segments qui ne contient pas x_2 ; on le note I_2. On suppose que l'on a construit une suite finie de segments

emboîtés $\{I_1, \dots I_n\}$ telle que, pour tout $k \in [\![1, n]\!]$, $x_k \notin I_k$ et que I_k soit de longueur $1/3^k$. On écrit I_n comme réunion de trois segments de longueur $1/3^{n+1}$. Il existe au moins un de ces segments qui ne contient pas x_{n+1} ; on le note I_{n+1}.

(ii) Pour prouver que $\mathbb{R}$ n'est pas dénombrable, on raisonne par l'absurde. On suppose que $\mathbb{R}$ est dénombrable. Alors $[0, 1]$ est *a fortiori* dénombrable et il existe une application bijective $x : \mathbb{N}^* \to [0, 1]$. Notons $x = (x_n)_{n\in\mathbb{N}^*}$; $[0, 1]$ est l'image de x, c'est-à-dire $[0, 1] = \bigcup\limits_{n\in\mathbb{N}^*} \{x_n\}$. Appliquons à la suite x le résultat de (i) : on obtient une suite $\{I_n\}_{n\in\mathbb{N}^*}$ de segments emboîtés de $[0, 1]$ telle que que, pour tout $n \in \mathbb{N}^*$, $x_n \notin I_n$.
D'après la proposition 5, l'intersection $\bigcap\limits_{n\in\mathbb{N}^*} I_n$ n'est pas vide. Soit a un point de cette intersection. On a $a \in [0, 1]$, mais $x_n \neq a$ pour tout $n \in \mathbb{N}^*$, donc $a \notin \bigcup\limits_{n\in\mathbb{N}^*} \{x_n\} = [0, 1]$ et l'on a une contradiction. ∎

Nous donnerons plus loin une autre preuve de la proposition ci-dessus, basée sur l'écriture décimale des nombres réels : *cf.* la proposition 66 de la page 592 (*cf.* aussi l'exercice IV.1.58).

1.2.2 Valeur absolue

Définition 6. On définit la *valeur absolue* d'un réel x par :

$$|x| = \max(x, -x) = \begin{cases} x & \text{si } x \geqslant 0 \\ -x & \text{si } x < 0 \end{cases}$$

La valeur absolue prolonge celle qui a été définie sur $\mathbb{Q}$ et possède les mêmes propriétés :

Proposition 7. Étant donnés deux réels x et y quelconques, on a :

- $|x| \geqslant 0$ et $\big(|x| = 0 \iff x = 0\big)$,
- $|x\,y| = |x|\,|y|$,
- $|x + y| \leqslant |x| + |y|$. (*Inégalité triangulaire*)

Corollaire 8. Étant donnés deux réels x et y, on a :

$$\Big||x| - |y|\Big| \leqslant |x - y|.$$

Démonstration. La relation $x = (x - y) + y$ combinée à l'inégalité triangulaire entraîne $|x| \leqslant |x - y| + |y|$ et l'on a donc :

$$|x| - |y| \leqslant |x - y|\,. \qquad (*)$$

En permutant les rôles joués par x et y, on obtient $|y| - |x| \leqslant |x - y|$, ce qui avec $(*)$ donne le résultat. ∎

Exemple. Si u est un réel tel que $|u| \leqslant k < 1$, comme $1 - |u| \leqslant |1 + u| \leqslant 1 + |u|$, on a :

$$1 - k \leqslant |1 + u| \leqslant 1 + k.$$

Proposition 9. Soient $(x, y) \in \mathbb{R}^2$ et $h \in \mathbb{R}_+^*$. On a l'équivalence :

$$|y - x| \leqslant h \Longleftrightarrow x - h \leqslant y \leqslant x + h.$$

Démonstration. C'est immédiat en remarquant que :

$$|x - y| \leqslant h \Longleftrightarrow \max(x - y, y - x) \leqslant h \Longleftrightarrow (x - y \leqslant h \quad \text{et} \quad y - x \leqslant h).$$ ■

1.2.3 Propriété d'Archimède

Proposition 10. Pour tout réel x, et pour tout réel $y > 0$, il existe un entier $n \in \mathbb{N}$ tel que $n\,y > x$.

C'est une propriété très simple, partagée par $\mathbb{N}$, $\mathbb{Z}$ et $\mathbb{Q}$ (pour lesquels la démonstration est assez simple), mais que tous les groupes ordonnés n'ont pas. En hommage à Archimède qui l'avait utilisée de manière intensive pour sa « méthode d'exhaustion », on dit que « $\mathbb{R}$ est *archimédien* ». Plus généralement un corps ordonné possèdant cette propriété est dit archimédien.

Démonstration. Soient $x \in \mathbb{R}$ et $y \in \mathbb{R}_+^*$. Raisonnons par l'absurde en supposant qu'il n'existe pas d'entier n tel que $n\,y > x$. Alors l'ensemble $A := \{n\,y \mid n \in \mathbb{N}\}$ est une partie de $\mathbb{R}$ non vide (elle contient $0 = 0\,y$) et majorée par x. Elle possède donc une borne supérieure a.
On a donc $\forall n \in \mathbb{N},\ n\,y \leqslant a$, et, par suite : $\forall n \in \mathbb{N},\ (n+1)\,y \leqslant a$. On a ainsi, pour tout $n \in \mathbb{N}$, $n\,y \leqslant a - y$, ce qui montre que $a - y$ majore A. C'est la contradiction cherchée puisque $a - y < a$ et que a est le plus petit des majorants de A. ■

D'après la propriété d'Archimède, pour tout $\varepsilon > 0$, il existe $n \in \mathbb{N}^*$ tel que $n\varepsilon > 1$, c'est-à-dire $\frac{1}{n} < \varepsilon$. On en déduit une variante utile de la proposition 3.

Point Méthode

Pour tout sous-ensemble majoré non-vide $A \subset \mathbb{R}$, $b \in \mathbb{R}$ est la borne supérieure de A si, et seulement si, b vérifie les deux conditions suivantes :

(i) b est un majorant de A ;

(ii) pour tout $n \in \mathbb{N}^*$, il existe un élément $x \in A$ tel que $b - \frac{1}{n} < x \leqslant b$.

Il y a, *mutatis mutandis*, un critère analogue pour la borne inférieure.

Remarque. Le lecteur vérifiera facilement que le critère du point méthode ci-dessus reste valable, *mutatis mutandis*, pour les bornes supérieures dans $\mathbb{Q}$ ($A \subset \mathbb{Q}$, $b \in \mathbb{Q}$ et $x \in \mathbb{Q}$).

Proposition et définition 11 (Partie entière).
Pour tout $x \in \mathbb{R}$, il existe un unique entier $n \in \mathbb{Z}$ tel que $n \leqslant x < n + 1$.
Cet entier est appelé *partie entière* de x et se note $E(x)$, $[x]$ ou $\lfloor x \rfloor$.

Démonstration. L'unicité est immédiate, puisque s'il existait deux tels entiers n et m distincts, on aurait, par exemple, $n < m$, c'est-à-dire $n + 1 \leqslant m$ et l'on arriverait ainsi à la contradiction $x < n + 1 \leqslant m \leqslant x$.

Montrons l'existence. Soit $x \in \mathbb{R}_+$. En appliquant la proposition 10 à x et $y = 1$, on obtient un entier $m \in \mathbb{N}$ tel que $m > x$. L'ensemble $A := \{p \in \mathbb{N} \mid p > x\}$ est donc une partie non vide de $\mathbb{N}$, qui possède par conséquent un plus petit élément $p_1 > x \geqslant 0$. L'entier $n := p_1 - 1$ appartient donc à $\mathbb{N}$ et pas à A par définition de p_1, donc vérifie $n \leqslant x < p_1 = n + 1$.

Soit $x \in \mathbb{R}_-$. Le caractère archimédien de $\mathbb{R}$ nous fournit un entier $q \in \mathbb{N}$ tel que $q > -x$ et ce qui précède un entier naturel n_0 tel que $n_0 \leqslant x + q < n_0 + 1$. L'entier relatif $n := n_0 - q$ vérifie donc les inégalités $n \leqslant x < n + 1$ requises. ∎

Définition 7. Une partie A de $\mathbb{R}$ est *dense* dans $\mathbb{R}$ si, quels que soient les réels x et y vérifiant $x < y$, il existe au moins un élément r de A vérifiant $x < r < y$.

Proposition 12. L'ensemble $\mathbb{Q}$ est *dense* dans $\mathbb{R}$.

Démonstration. C'est une conséquence du caractère archimédien de $\mathbb{R}$.

- Soient $x, y \in \mathbb{R}$ tels que $x < y$. Notons $\varepsilon := y - x > 0$. Il existe un entier $q \in \mathbb{N}^*$ tel que $q\varepsilon > 1$, c'est-à-dire $1/q < \varepsilon$, par exemple $q := E\left(\dfrac{1}{\varepsilon}\right) + 1$.
- Parmi les multiples entiers de $1/q$, il y en a au moins un strictement supérieur à x. Notons p le plus petit. Vu le choix de q, il est alors assez évident que p/q est strictement inférieur à y. Prouvons-le rigoureusement. On a, par définition de p :

$$\frac{p}{q} \leqslant x < \frac{p+1}{q} \qquad \text{et donc} \qquad x < \frac{p+1}{q} \leqslant x + \frac{1}{q} < x + (y - x) = y$$

ce qui prouve que le rationnel $r = \dfrac{p+1}{q}$ vérifie $x < r < y$. ∎

Corollaire 13. Pour tous $x,\ y \in \mathbb{R}$ vérifiant $x < y$, il existe une infinité de nombres rationnels strictement compris entre x et y.

Démonstration. Le résultat est vrai si x et y sont rationnels. Supposons x et y quelconques. Il existe un rationnel r_1 tel que $x < r_1 < y$. Il existe un rationnel r_2 tel que $r_1 < r_2 < y$, donc une infinité de rationnels strictement compris entre r_1 et r_2 et *a fortiori* entre x et y. ∎

Exercice 6.

Montrer que $\mathbb{R} \setminus \mathbb{Q}$ est dense dans $\mathbb{R}$.

Solution. L'exemple de la page 515 nous a montré l'existence d'un irrationnel $a := \sqrt{2}$. Soit $(x, y) \in \mathbb{R}^2$ tel que $x < y$. Puisque $\mathbb{Q}$ est dense dans $\mathbb{R}$, il existe un rationnel r tel que $x + a < r < y + a$. On a alors $x < r - a < y$ et $r - a$ ne peut pas être rationnel sinon $a = r - (r - a)$ serait rationnel.

Remarque. Soit $x \in \mathbb{R}$. On a $x = \sup\{r \in \mathbb{Q} \mid r < x\}$. En effet x majore $A_x := \{r \in \mathbb{Q} \mid r < x\}$ et c'est le plus petit majorant. Sinon il existerait un majorant y tel que $y < x$, donc un rationnel r' tel que $y < r' < x$ et l'on aurait $r' \in A_x$ et une contradiction. Cette remarque est à la base d'une construction de $\mathbb{R}$ par le procédé des *sections commençantes*.

Proposition 14 (Propriété d'Archimède multiplicative).
Pour tout réel x et tout réel $y > 1$, il existe $n \in \mathbb{N}$ tel que $y^n > x$.

Démonstration. Posons $r := y - 1$. C'est un réel strictement positif par hypothèse. Or, par la formule du binôme de Newton :

$$\forall n \in \mathbb{N} \setminus \{0,1\}\,,\ y^n = (1+r)^n = 1 + n\,r + \sum_{p=2}^{n} \binom{n}{p} r^p \geqslant n\,r$$

et le caractère archimédien de $\mathbb{R}$ nous prouve l'existence d'un entier n, que l'on peut supposer supérieur à 2, tel que $n\,r > x$. On a alors $y^n > x$. ∎

1.3 Intervalles de $\mathbb{R}$

Définition 8. Les *intervalles* de $\mathbb{R}$ sont $\mathbb{R}$ et les parties de $\mathbb{R}$ de la forme :

$$\begin{array}{ll}
[a,b] = \left\{x \in \mathbb{R} \;\middle|\; a \leqslant x \leqslant b\right\} &]a,b[= \left\{x \in \mathbb{R} \;\middle|\; a < x < b\right\} \\
[a,b[= \left\{x \in \mathbb{R} \;\middle|\; a \leqslant x < b\right\} &]a,b] = \left\{x \in \mathbb{R} \;\middle|\; a < x \leqslant b\right\} \\
[a,+\infty[= \left\{x \in \mathbb{R} \;\middle|\; a \leqslant x\right\} &]a,+\infty[= \left\{x \in \mathbb{R} \;\middle|\; a < x\right\} \\
]-\infty,b] = \left\{x \in \mathbb{R} \;\middle|\; x \leqslant b\right\} &]-\infty,b[= \left\{x \in \mathbb{R} \;\middle|\; x < b\right\}
\end{array}$$

avec a et b réels.

Remarques.

- L'ensemble vide est un intervalle puisque, par exemple, $\varnothing =]0,0[$.
- Si $a \leqslant b$, l'intervalle $[a,b]$ est un *segment*.
- Pour $a < b$, l'intervalle $]a,b[$ est appelé *intervalle ouvert* d'extrémités a et b.
- Les intervalles $]a,+\infty[$, et $]-\infty,b[$ sont appelés *demi–droites ouvertes*.
- Les intervalles $[a,+\infty[$ et $]-\infty,b]$ sont appelés *demi–droites fermées*.
- Pour $a < b$, les intervalles $[a,b[$ et $]a,b]$ sont appelés *intervalles semi–ouverts* ou *semi–fermés*.

Définition 9. Si I est un intervalle de $\mathbb{R}$, on appelle *intérieur de* I l'intervalle *ouvert* ayant les mêmes bornes que I. Un point $a \in I$ est dit *intérieur à* I s'il appartient à l'intérieur de I.

En utilisant le corollaire 13, on vérifie que $a \in I$ est intérieur à I si, et seulement s'il existe $\varepsilon > 0$ tel que $]x-\varepsilon, x+\varepsilon[\subset I$.
Un intervalle est ouvert si, et seulement s'il est égal à son intérieur.

Définition 10. Les parties C de $\mathbb{R}$ telles que :

$$\forall x \in C\,,\ \forall y \in C\,,\ \text{tels que } x \leqslant y,\ [x,y] \subset C,$$

sont dites *convexes*.

Proposition 15. Les intervalles de $\mathbb{R}$ sont les parties *convexes*.

Démonstration. Les intervalles sont évidemment des parties convexes. Montrons la réciproque. Soit I une partie convexe de $\mathbb{R}$. Si I est vide, alors c'est un intervalle. Supposons donc I non vide et posons $a = \inf I$ et $b = \sup I$ avec comme convention $a = -\infty$ si I est non minoré et $b = +\infty$ si I est non majoré.
Montrons que les seules possibilités pour la partie I sont celles données dans le tableau ci–dessous :

	$b \in I$	$b \in \mathbb{R} \setminus I$	$b = +\infty$
$a \in I$	$I = [a,b]$	$I = [a,b[$	$I = [a,+\infty[$
$a \in \mathbb{R} \setminus I$	$I =]a,b]$	$I =]a,b[$	$I =]a,+\infty[$
$a = -\infty$	$I =]-\infty,b]$	$I =]-\infty,b[$	$I =]-\infty,+\infty[$

- Il est évident que dans chaque cas, l'ensemble I est inclus dans l'intervalle indiqué.
- Pour la réciproque, il suffit de montrer l'inclusion $]a,b[\subset I$ puisqu'alors on aura :

  $$I =]a,b[\,,\ I =]a,b]\,,\ I = [a,b[\quad \text{ou} \quad I = [a,b]$$

  en fonction de l'appartenance ou non de a et b à I.
  Soit $z \in]a,b[$. Comme $z < b$, le réel z n'est pas un majorant de I, et l'on peut trouver $y \in I$ tel que $z < y$. De même, on peut trouver $x \in I$ tel que $x < z$.
  On a ainsi $z \in [x,y]$ avec $x \in I$ et $y \in I$, ce qui entraîne $z \in I$. Donc $]a,b[\subset I$. ■

  Remarque. Il est clair, au vu de la démonstration précédente, que si $a < b$, les intervalles $[a,b]$, $[a,b[$, $]a,b]$ et $]a,b[$ admettent a pour borne inférieure et b pour borne supérieure.

Exercice 7.
Montrer que les intervalles de $\mathbb{R}$ contenant au moins deux points ne sont pas des ensembles dénombrables.
Solution. Montrons d'abord que le segment $I := [0,1]$ n'est pas dénombrable. On raisonne par l'absurde. On suppose I dénombrable. On en déduit, en utilisant les translations $x \mapsto x+n$, pour tout $n \in \mathbb{Z}$, que les segments $I_n := [n, n+1]$ sont dénombrables. On a $\mathbb{R} = \bigcup\limits_{n \in \mathbb{Z}} I_n$; par suite, $\mathbb{R}$ est une réunion dénombrable d'ensembles dénombrable et il est donc dénombrable. On aboutit à une contradiction.
Ainsi I n'est pas dénombrable. En utilisant des homothéties et des translations, on en déduit que si $a < b$, le segment $[a,b]$ n'est pas dénombrable. Si un intervalle J contient deux points distincts, il contient le segment qui les joint, celui-ci n'étant pas dénombrable, J ne l'est pas non plus.

---

2 Suites numériques

2.1 Généralités sur les suites

Les suites sont un outil très important en *analyse mathématique*. Elles permettent notamment de donner un sens précis à l'idée qu'une succession de nombres s'approche d'un nombre donné : idée qui remonte au moins au mathématicien grec Eudoxe[2]. Archimède avait (entre autres) utilisé cette idée pour approcher le nombre π en calculant le périmètre de polygones réguliers dont le nombre de côtés augmente. De nos jours les suites sont aussi utilisées dans de nombreuses applications, en particulier en relation avec des problèmes de *modélisation*.

Rappelons la définition des suites (voir page 19).

Définition 11. Un ensemble E étant donné, on appelle *suite de* E toute application de $\mathbb{N}$ à valeurs dans E. Lorsque $E = \mathbb{R}$, on parle de *suite réelle*, et si $E = \mathbb{C}$, de *suite complexe* ; ces deux cas constituent les *suites numériques*.

Si $u : \mathbb{N} \to E$ est une suite, on la note aussi $(u_n)_{n\in\mathbb{N}}$, où $u_n := u(n)$, ou (u_n). Par extension, on appelle encore suites les applications de $\mathbb{N}^*$ dans E ou, plus généralement, les applications de $\{n \in \mathbb{N} \mid n \geqslant n_0\}$ dans E. Si $(u_n)_{n\in\mathbb{N}}$ est une suite, on appelle suite *tronquée à l'ordre* n_0, la « suite » : $(u_n)_{n\geqslant n_0}$.

Si $u : \mathbb{N} \to E$ est une suite, $u(\mathbb{N}) := \{u_n\}_{n\in\mathbb{N}} \subset E$ est *l'image de la suite*. Cette image est au plus dénombrable ; elle est finie ou infinie.

Beaucoup des notions exposées ci-dessous seront communes aux deux cas des suites numériques (réelles ou complexes). On notera pour cela $\mathbb{K}$ le corps $\mathbb{R}$ des nombres réels ou le corps $\mathbb{C}$ des nombres complexes et $\mathbb{K}^* := \mathbb{K} \setminus \{0\}$.

Dans l'étude de nombreux problèmes « pratiques », en physique, chimie, biologie, épidémiologie, écologie, économie…, on s'intéresse à l'évolution au cours du temps d'une quantité u « continue » ou « discrète » (quantité physique : température, pression… ; population : nombre d'habitants d'une ville, d'individus d'une espèce, de bactéries… ; montant d'un compte en banque…). Pour cela, on considère la valeur u_n que prend u à des valeurs successives du temps t : $t_n := t_0 + n\Delta t$, où Δt est un intervalle *fixé* (« bien choisi » en fonction du problème). On obtient ainsi une suite (u_n). Dans de tels cas, on parle souvent de modélisation « en temps discret ». On s'intéresse souvent à l'évolution « à long terme » de u et l'on étudie pour cela « ce que devient u_n pour n grand ».

2.1.1 Définitions et opérations

L'ensemble des suites de $\mathbb{K}$ est noté $\mathbb{K}^{\mathbb{N}}$.

Rappels et premières propriétés

- Une suite $(u_n) \in \mathbb{K}^{\mathbb{N}}$ est *constante* s'il existe $k \in \mathbb{K}$ tel que $\forall n \in \mathbb{N},\ u_n = k$. Par une récurrence immédiate, cela revient au même de dire : $\forall n \in \mathbb{N},\ u_{n+1} = u_n$.
- Une propriété est vraie *à partir d'un certain rang* s'il existe $n_0 \in \mathbb{N}$ tel que la propriété soit vraie pour tous les termes u_n de la suite vérifiant $n \geqslant n_0$, autrement dit pour la suite tronquée $(u_n)_{n\geqslant n_0}$.

[2] On attribue à Eudoxe de Cnide -408 -355 l'invention de *la méthode d'exhaustion* pour calculer des aires. Cette méthode a été reprise par Archimède -287 -212.

- Une suite *stationnaire* est une suite constante à partir d'un certain rang. Une suite (u_n) est stationnaire s'il existe un entier n_0 tel que la suite *tronquée* $(u_n)_{n\geqslant n_0}$ soit constante.
- On dit qu'une suite nulle à partir d'un certain rang est *à support fini*.
- Une suite (u_n) est *périodique* s'il existe un entier $\tau \in \mathbb{N}^*$ tel que $\forall n \in \mathbb{N}\,,\ u_{n+\tau} = u_n$. Un tel entier τ est appelé *une période* de (u_n) et l'on dit aussi que (u_n) est τ-périodique. Dans ce cas, 2τ, 3τ, ..., sont aussi des périodes de la suite.
  Comme toute partie non vide de $\mathbb{N}^*$ admet un plus petit élément, une suite périodique admet une *plus petite période*, évidemment unique.
  Montrons que, si τ est *la plus petite période*, l'ensemble des périodes est l'ensemble des multiples non nuls de τ. Soit τ' une période ; en utilisant la division euclidienne $\tau' = q\tau + r$, $0 \leqslant r < \tau$, on obtient, pour tout $n \in \mathbb{N}$:
  $$u_n = u_{n+\tau'} = u_{n+q\tau+r} = u_{n+r+q\tau} = u_{n+r},$$
  donc r est nul ou est une période strictement inférieure à τ ce qui est impossible. Par suite $\tau' = q\tau$.

Opérations sur les suites

L'ensemble $\mathbb{K}^{\mathbb{N}}$ est muni des opérations usuelles d'addition et de multiplication terme à terme :
$$(u_n) + (v_n) = (u_n + v_n) \qquad \text{et} \qquad (u_n) \times (v_n) = (u_n\, v_n).$$
On sait alors (voir l'exemple 3 de la page 134) que $(\mathbb{K}^{\mathbb{N}}, +, \times)$ est un anneau.
De même, avec la multiplication externe :
$$\forall (u_n) \in \mathbb{K}^{\mathbb{N}}\,,\ \forall \lambda \in \mathbb{K}\,,\ \lambda.(u_n) = (\lambda\, u_n);$$
$(\mathbb{K}^{\mathbb{N}}, +, .)$ est un $\mathbb{K}$-espace vectoriel.

2.1.2 Suites extraites

Nous avons introduit ci-dessus, la notion de suite tronquée : si (u_n) est une suite numérique, une *suite tronquée* de (u_n) est une suite de la forme $(u_n)_{n\geqslant n_0}$, avec $n_0 \in \mathbb{N}$. Cela revient à retirer les premiers termes de la suite. Mais l'on peut aussi retirer, par exemple, tous les termes de rang impair, ou tous ceux d'indice premier, etc. Tant qu'il reste une infinité de termes, nous avons encore, en renumérotant les termes, une suite, qui est appelée *suite extraite*, ou encore *sous-suite*.
La première dénomination indique l'idée de *sélectionner* certains termes : on choisit un entier n_0, et l'on garde le terme x_{n_0} ; on choisit alors un $n_1 > n_0$, et l'on garde x_{n_1} ; etc. Si ce processus est infini, on obtient une nouvelle suite $(x_{n_k})_{k\in\mathbb{N}}$, dont les termes sont une partie des termes initiaux. Plus formellement :

Définition 12. Une *suite extraite*, ou *sous-suite*, d'une suite (u_n) est une suite de la forme $(v_k := u_{n_k})_{k\in\mathbb{N}}$, où $(n_k)_{k\in\mathbb{N}}$ est une suite *strictement croissante* d'entiers naturels, c'est-à-dire une application strictement croissante de $\mathbb{N}$ dans $\mathbb{N}$: $\varphi : k \mapsto n_k$.

On note parfois $(u_{\varphi(k)})_{k\in\mathbb{N}}$ la suite extraite. On vérifie que la connaissance d'une application strictement croissante $\varphi : \mathbb{N} \mapsto \mathbb{N}$ est équivalente à celle de son image, c'est-à-dire d'un sous-ensemble *infini* A de $\mathbb{N}$. Si φ est donnée, $A := Im\ \varphi$ est infini. Inversement étant donné $A \subset \mathbb{N}$ infini, on *définit* $\varphi(k)$ comme le $k^{ème}$ élément de A pour l'ordre induit par celui de $\mathbb{N}$.

Exemple. Étant donnée une suite (u_n), on utilise couramment ses deux sous-suites $(u_{2p})_{p\in\mathbb{N}}$ et $(u_{2p+1})_{p\in\mathbb{N}}$ (dites suites des termes de rang pair et des termes de rang impair) .
La suite (u_n) est 2-périodique si, et seulement si, ces deux sous-suites sont constantes.

On vérifie facilement la généralisation suivante.

Proposition 16. Soit (u_n) une suite et $\tau \in \mathbb{N}^*$, $\tau \geqslant 2$. Les deux conditions suivantes sont équivalentes.

(i) Les τ sous-suites v_q de (u_n) définies par $(v_{q,p} := u_{q+p\tau})_{p\in\mathbb{N}}$, $q \in [\![0, \tau-1]\!]$ sont constantes.

(ii) La suite (u_n) est périodique, de période τ.

Exercice 8.
Soit $u : \mathbb{N} \to \mathbb{K}$ une suite d'image finie. Montrer qu'il existe une suite extraite constante.

Solution. On note $\{a_1, \ldots, a_m\}$ l'image de u. On a $\mathbb{N} = \bigcup_{i=1}^{m} u^{-1}(a_i)$. Il existe donc $i_0 \in [\![1, m]\!]$ tel que l'ensemble $u^{-1}(a_{i_0})$ soit *infini*. Par suite, il existe une suite extraite constante égale à a_{i_0}.

Exercice 9.
Soit $u : \mathbb{N} \to \mathbb{R}$ une suite décroissante d'image finie. Montrer qu'elle est constante à partir d'un certain rang.
Solution. Notons d'abord que si une suite réelle v est *strictement* décroissante, elle prend une infinité de valeurs.
Pour montrer que u est constante à partir d'un certain rang, on raisonne par l'absurde. Si ce n'était pas le cas, il existerait une suite v strictement décroissante extraite de u. Celle-ci serait d'image finie (puisque contenue dans celle de u) et l'on aurait une contradiction.
On peut donner une autre solution en utilisant l'exercice précédent. Il existe une suite extraite $v = (v_h := u_{n_h})_{h\in\mathbb{N}}$ constante et égale à $a \in \mathbb{R}$. Pour tout $n \in \mathbb{N}$, tel que $n \geqslant n_0$ il existe $h \in \mathbb{N}$ tel que $n_h \leqslant n < n_{h+1}$. On a $a = u_{n_h} \geqslant u_n \geqslant u_{n_{h+1}} = a$, donc $u_n = a$.

2.1.3 Suites bornées

Définition 13. Une suite réelle $(u_n) \in \mathbb{R}^{\mathbb{N}}$ est *majorée* (respectivement *minorée*) si $\{u_n \mid n \in \mathbb{N}\}$ est une partie majorée (respectivement minorée) de $\mathbb{R}$, c'est-à-dire s'il existe $M \in \mathbb{R}$ tel que $\forall n \in \mathbb{N},\ u_n \leqslant M$ (respectivement $\forall n \in \mathbb{N},\ u_n \geqslant M$).

Définition 14. On dit qu'une suite $(u_n) \in \mathbb{K}^{\mathbb{N}}$ est *bornée* si la suite réelle $(|u_n|)$ est majorée. Plus précisément, si $M \in \mathbb{R}_+$, on dit que la suite (u_n) est *bornée par* M si $\forall n \in \mathbb{N},\ |u_n| \leqslant M$.

Remarque. Les propriétés suivantes sont immédiates :

- Une suite réelle est bornée si, et seulement si, elle est majorée et minorée.
- Une suite complexe est bornée si, et seulement si, sa partie réelle et sa parie imaginaire sont bornées.

Exercice 10.
Une suite est bornée si, et seulement si, elle est bornée à partir d'un certain rang (c'est-à-dire si, et seulement si, l'une de ses suites tronquées est bornée).
On a un énoncé similaire pour « minorée » et « majorée ».
Solution. Il est évident que si une suite est bornée, ses suites tronquées le sont aussi.
Inversement, supposons (u_n) bornée à partir d'un certain rang N, c'est-à-dire qu'il existe $\mu \geqslant 0$ tel que pour tout $n \geqslant N$, $|u_n| \leqslant \mu$. Posons $\tilde{\mu} := \max(\mu, |u_0|, |u_1|, \ldots, |u_{N-1}|)$; c'est le plus grand d'un nombre fini de réels positifs, donc c'est un réel positif. De plus on a $|u_n| \leqslant \tilde{\mu}$ pour tout $n \in \mathbb{N}$ comme on le voit en distinguant les cas $n < N$ et $n \geqslant N$. Donc la suite (u_n) est bornée.

Proposition 17. L'ensemble des suites bornées de $\mathbb{K}$ est un sous–espace vectoriel et un sous-anneau de $\mathbb{K}^{\mathbb{N}}$.

Démonstration. Les suites constantes $(0)_{n\in\mathbb{N}}$ et $(1)_{n\in\mathbb{N}}$ sont évidemment bornées et si u et v sont bornées respectivement par M_1 et M_2, alors

- les suites $u \pm v$ sont bornées par $M_1 + M_2$,
- la suite $u \times v$ est bornée par $M_1\, M_2$,
- pour tout scalaire $\lambda \in \mathbb{K}$, la suite $\lambda.u$ est bornée par $|\lambda|\, M_1$. ■

2.1.4 Suites définies par une récurrence

Définitions

Soit X une partie de $\mathbb{K}$ et $f : X \to X$. Le théorème 2 de la page 64 montre que, pour tout $a \in X$, il existe une unique suite (u_n) vérifiant :

$$u_0 := a \qquad \text{et} \qquad \forall n \in \mathbb{N},\ u_{n+1} := f(u_n).$$

On dit que (u_n) est une suite définie par une récurrence (ou suite récurrente) et que $u_0 = a$ est la *valeur initiale* ou la *condition initiale* de la suite. On a $u_{n+1} = f^{\circ n}(u_0)$, où l'on a noté $f^{\circ n} = f \circ \ldots \circ f$ (n compositions). On dit que l'application $f^{\circ n}$ est la $n^{\text{ème}}$ *itérée* de l'application f. La suite récurrente (u_n) est obtenue en appliquant à u_0 les itérées successives de f. L'image d'une suite récurrente définie par $f : X \to X$ est appelée une *orbite* de f ; c'est un sous-ensemble (fini ou infini) de X.
En pratique, la donnée est souvent une application $f : Y \to \mathbb{K}$, où $Y \subset \mathbb{K}$ et il faut *trouver* $X \subset Y$ tel que $f(X) \subset X$; on dit que X est un sous-ensemble *invariant par* f. Ce problème peut être difficile.

Exemple. Soit $f : Y :=]-\infty, 2] \to \mathbb{R}$ définie par $f : x \mapsto f(x) := \sqrt{2-x}$. On a $f([-2,2]) \subset [0, \sqrt{2}]$. En effet, si $|x| \leqslant 2$, on a $f^2(x) = 2 - x \leqslant 4$, donc $0 \leqslant f(x) \leqslant 2$. On peut donc choisir $X := [-2, 2]$ ou $X = [0, 2]$; on a dans les deux cas $f(X) \subset X$.

On peut considérer le cas, plus général, de suites récurrentes définies par une partie $X \subset \mathbb{K}^2$ et une application $f : X \to X$. Il existe une unique suite (dite récurrente d'ordre 2), définie par les *deux* conditions initiales $u_0,\ u_1 \in X$ et, pour tout $n \in \mathbb{N}$, $u_{n+2} = f(u_n, u_{n+1})$. Nous n'étudierons pas de façon générale de telles suites mais nous en donnerons quelques exemples. Le cas, dit *linéaire*, où $f(x, y) = ay + bx$ est traité dans le module IV.6 (*cf.* page 787).
On peut aussi considérer le cas de suites définies de la façon suivante.
Soient $X \subset \mathbb{K}$ et $f : X \times \mathbb{N} \to X$. Pour tout $a \in X$, il existe une unique suite (u_n) vérifiant :

$$u_0 = a \qquad \text{et} \qquad \forall n \in \mathbb{N},\ u_{n+1} = f(u_n, n).$$

On dit que l'on a une récurrence *non autonome*. Un exemple classique est la suite $(n!)$ définie par :

$$f(x, n) := (n+1)\,x \quad \text{et} \quad u_0 := 1.$$

Les suites récurrentes sont un sujet important en mathématiques, dans les sciences expérimentales et en modélisation, et ceci pour diverses raisons. La plus ancienne est la *résolution numérique* des équations de la forme $F(x) = x$; le premier exemple historique est le calcul de $\sqrt{2}$ par Héron d'Alexandrie vers l'an 0 (*cf.* l'exemple de la page 558). Cette question sera traitée en détail dans le deuxième volume de cette série (*cf.* L2, méthode de Babylone). Dans les problèmes d'origine « pratique » la fonction f définissant la suite traduit souvent la loi d'évolution d'un phénomène. Récemment[3] la théorie des suites récurrentes s'est très fortement développée (avec, en particulier, l'étude de suites définies par l'itération d'un *polynôme du second degré*) en relation avec ce que l'on appelle « la théorie du chaos » et celle des fractales[4] et aussi avec des problèmes de modélisation.[5] Par exemple, les suites, dites *logistiques*, définies par une fonction polynôme de la forme $f(x) := \mu x(1-x)$, que nous étudierons plus loin (*cf.* la page 570), apparaissent en dynamique des populations. La notion de chaos et son étude systématique dépassent largement le niveau de notre livre ; nous nous limiterons essentiellement à des exemples « significatifs » mais néanmoins abordables avec des moyens élémentaires.

Points fixes et cycles

Soient X, Y deux ensembles tels que $X \subset Y$ et $f : X \to Y$. On dit que $\ell \in X$ est un *point fixe* de f si $f(\ell) = \ell$. Si ℓ est un point fixe de f, $u_0 := \ell$ et f définissent une suite récurrente u et celle-ci est *constante*.
Si $X \subset \mathbb{K}$ et $Y \subset \mathbb{K}$ et si f est continue, nous verrons plus loin que la détermination des points fixes est *essentielle* pour l'étude de la *convergence* des suites récurrentes (*cf.* 2.2.6).
Soient X un ensemble et $f : X \to X$. Les images des suites récurrentes *périodiques* non constantes définies par f sont appelées des *cycles* de f. Plus précisément on a la définition suivante.

[3] Depuis les années 70 et grâce à la montée en puissance des ordinateurs
[4] Introduite par le mathématicien franco-américain Benoît Mandelbrot en 1975.
[5] L'itération d'une application, d'une part, et les équations différentielles, d'autre part, sont des exemples de ce que l'on appelle en mathématiques et en sciences expérimentales des *systèmes dynamiques*.

Définition 15. Soient X un ensemble et $f : X \to X$. On dit qu'un sous-ensemble fini A de X est un *cycle* de f s'il a au moins deux éléments et s'il existe une suite récurrente *périodique* définie par f dont l'image est A. Si $\operatorname{card} A = \tau$, on dit que A est un τ-cycle (ou cycle de longueur τ) de f.

Nous verrons plus loin que la donnée d'un τ-cycle de f est équivalente à celle de τ-points ordonnés *deux à deux distincts* $u_0, u_1, \ldots, u_{\tau-1}$ de X tels que :

$$u_1 = f(u_0),\ u_2 = f(u_1),\ \ldots\ , u_{\tau-1} = f(u_{\tau-2})\,, u_0 = f(u_{\tau-1}).$$

Cette dernière donnée est considérée à permutation circulaire près, ce qui justifie la terminologie.

Un sous-ensemble fini A de X, tel que $\operatorname{card} A \geqslant 2$, peut être l'image d'une suite récurrente définie par f sans être un cycle de f : nous en donnerons un exemple plus loin (*cf.* l'exemple 1 de la page 529). Par contre un tel ensemble contient toujours un point fixe ou un cycle (*cf.* l'exercice 11).

Nous allons maintenant donner quelques propriétés des suites récurrentes *périodiques*. Nous n'utiliserons pas tout de suite ces résultats. La lecture de la fin de ce paragraphe peut donc être provisoirement omise par le lecteur qui s'y reportera au moment où ce sera nécessaire (*cf.* 2.2.6). On peut ainsi poursuivre la lecture à partir de 2.1.5.

On ne confondra pas la notion de fonction périodique f et celle de suite périodique définie par f.

Proposition 18. Soient X un ensemble, $f : X \to X$, $u_0 \in X$ et $\tau \in \mathbb{N}$ tel que $\tau \geqslant 2$. On note u la suite récurrente définie par f et u_0.

(i) La suite récurrente u est τ-périodique si, et seulement si, $u_\tau = f^{\circ\tau}(u_0) = u_0$, autrement dit si, et seulement si, u_0 est un *point fixe* de $f^{\circ\tau}$.

(ii) On suppose que la suite u est périodique. Les deux conditions suivantes sont équivalentes :

a) Le cardinal de l'image de u est τ.

b) La plus petite période de la suite u est τ.

Si u_0 est un *point fixe* de $f^{\circ\tau}$, on dit aussi que u_0 est un point τ-périodique de f. Si de plus il n'existe pas d'entier $\tau' < \tau$ tel que $f^{\circ\tau}(u_0) = u_0$, on dit que u_0 est un point de plus petite période (ou période primitive) τ de f. Si la suite $u = (u_n)$ est périodique à partir d'un certain rang, on dit que u_0 est un point *pré-périodique* de f.

Démonstration.

(i) La condition $u_\tau = u_0$ est évidemment une condition nécessaire de τ-périodicité. Inversement, supposons $f^{\circ\tau}(u_0) = u_0$. Pour tout $n \in \mathbb{N}$, on a :

$$f^{\circ n} \circ f^{\circ\tau}(u_0) = f^{\circ(\tau+n)}(u_0) = u_{n+\tau} = f^{\circ n}(u_0) = u_n.$$

(ii) Montrons b) $\Rightarrow$ a). D'après (i), pour tout $i \in [\![1, \tau - 1]\!]$, $u_0 \neq u_i$. Nous allons en déduire que les τ points $u_0, u_1, \ldots, u_{\tau-1}$ sont *deux à deux distincts*. On raisonne par l'absurde. Supposons qu'il existe $i, j \in [\![0, \tau - 1]\!]$, tels que $i < j$ et $u_i = u_j$. Alors :

$$u_0 = u_\tau = f^{\circ\tau-i}(u_i) = f^{\circ\tau-i}(u_j) = u_{\tau+j-i} = u_{j-i}.$$

Ainsi $u_0 = u_{j-i}$, avec $0 < j - i \leqslant \tau - 1$ et l'on a une contradiction.

L'image de u est $\{u_0, u_1, \ldots, u_{\tau-1}\}$. D'après ce qui précède son cardinal est τ.

Montrons a) $\Rightarrow$ b). Par hypothèse u est périodique. On note τ_1 sa plus petite période. D'après l'implication b) $\Rightarrow$ a), on a $\tau_1 = \operatorname{card}(\operatorname{Im} u) = \tau$. ■

D'après la proposition 18 ci-dessus, un τ-cycle est l'image d'une suite récurrente de plus petite période τ. Au lieu de cycle (resp. τ-cycle) de f on dit aussi *orbite périodique* (resp. *orbite de période primitive* τ) de f.

Soit A un τ-cycle de f. Par définition, il existe $u_0 \in A$ tel que si (u_n) est la suite définie par u_0 et f, l'on ait $\{u_0, u_1, \ldots, u_{\tau-1}\} = A$ et $f(u_{\tau-1}) = u_0$. Par suite $f(A) = A$ et f opère sur A par permutation circulaire. On note souvent un cycle $(u_0, u_1, \ldots, u_{\tau-1})$ pour rappeler « l'ordre » des points. On peut remplacer u_0 par $v_0 := u_q$, $q \in [\![0, \tau - 1]\!]$, et l'on obtient des propriétés similaires en remplaçant la suite (u_n) par la suite (v_n). Ainsi $(u_0, u_1, \ldots, u_{\tau-1})$, $(u_1, u_2, \ldots, u_\tau)$, $(u_2, u_3, \ldots, u_{\tau+1}) \ldots$ représentent le même cycle. Pour tout $q \in [\![0, \tau - 1]\!]$, on a $f^{\circ\tau}(u_q) = u_{q+\tau} = u_q$, donc u_q est un point fixe de $f^{\circ\tau}$.
On notera que l'on peut avoir un sous-ensemble fini $A \subset X$ tel que $f(A) = A$ mais qui n'est pas un cycle de f (par exemple un ensemble formé de deux points fixes).

Exercice 11.
Soient $f : X \to X$ et $A \subset X$ l'image d'une suite récurrente $u = (u_n)$ définie par f. On supppose que A est *fini*. Montrer que la suite u est périodique à partir d'un certain rang et que A contient un point fixe ou un cycle.
Solution. Les u_n ne peuvent pas être deux à deux distincts, sinon A serait infini. Il existe donc $p, q \in \mathbb{N}$ tels que $p < q$ et $u_p = u_q$. Notons $v = (v_n)$ la suite définie par f et $v_0 := u_p$. On a $v_{q-p} = v_0$, donc, d'après l'assertion (i) de la proposition 18, la suite v est périodique de période $q - p$ et son image B est un point fixe ou un cycle. On a $B \subset A$. La suite u est périodique à partir du rang p.

En utilisant la proposition 18, on vérifie facilement les remarques suivantes.

Remarques.

- Deux cycles de f sont d'intersection vide ou égaux.
- Les points d'un τ-cycle de f sont des points fixes de $f^{\circ\tau}$ et ce ne sont pas des points fixes de $f^{\circ\tau'}$ pour $\tau' < \tau$.
- Si τ' divise τ, les points fixes de $f^{\circ\tau'}$ sont des points fixes de $f^{\circ\tau}$.
- Si a est un point fixe de $f^{\circ\tau}$ qui n'est pas un point fixe de f, il appartient à un τ'-cycle de f, τ' divisant τ.

- Supposons que a est un point fixe de $f^{\circ\tau}$, alors $f(a)$ est aussi un point fixe de $f^{\circ\tau}$. En effet $f^{\circ\tau} \circ f(a) = f \circ f^{\circ\tau}(a) = f(a)$. Par suite, si l'on note F l'ensemble des points fixes de $f^{\circ\tau}$ on a $f(F) \subset F$. On vérifie, en considérant l'orbite de chaque point, que l'on peut écrire F comme la réunion (disjointe) des cycles de f dont l'ordre τ' divise τ et des points fixes de f.

Nous verrons que les points fixes et les cycles de f jouent un rôle très important dans l'étude des suites récurrentes définies par f. Pour les suites issues de problèmes de modélisation, les points fixes correspondent à des phénomènes stationnaires (équilibres...) et les cycles à des phénomènes périodiques.

Exemples.

1. Soit $I := [0,1]$. On cherche $f : I \to I$ continue, affine par morceaux et admettant $(0, 1/2, 1)$ pour 3 cycle. On doit avoir $f(0) = 1/2$, $f(1/2) = 1$ et $f(1) = 0$. Le plus simple est de chercher f affine sur $[0, 1/2]$ et $[1/2, 1]$. Il y a une solution (unique) :

   $$f : x \mapsto x + 1/2 \quad \text{si} \quad x \in [0, 1/2] \quad \text{et} \quad f : x \mapsto -2x + 2 \ \text{ si } \ x \in \,]1/2, 1].$$

   On vérifie que f admet un unique point fixe $x := 2/3$. Par ailleurs, l'image de f est I. Elle induit donc une application de I dans I, toujours notée f.

   Soit $u_0 := 13/24$. Notons $u = (u_n)$ la suite définie par f et u_0. On a $u_1 = 11/12$, $u_2 = 1/6$, $u_3 = 2/3$ et la suite est constante à partir de $n = 4$. L'image de f est $A := \{\frac{13}{24}, \frac{11}{12}, \frac{1}{6}, \frac{2}{3}\}$. Elle est finie ($\operatorname{card} A = 4$) mais ce n'est pas un cycle.

2. On définit $g : \mathbb{R} \to \mathbb{R}$ par $g : x \mapsto f(x) := -\frac{3}{2}x^2 - \frac{1}{2}x + 1$. On a $g(0) = 1$, $g \circ g(0) = g(1) = -1$ et $g \circ g \circ g(0) = g(-1) = 0$. Par suite $(0, 1, -1)$ est un 3 cycle de g.

Exemple. *L'application tente.* On définit une application continue, affine par morceaux : $T : [0,1] \to [0,1]$, dite *application tente*, par :

$$x \mapsto T(x) = 2x \ \text{ si } \ x \in [0, 1/2] \quad \text{et} \quad x \mapsto T(x) = 2(1 - x) \ \text{ si } \ x \in \,]1/2, 1].$$

L'étude des *cycles* de T se ramène à celle des *points fixes* de $T^{\circ n}$, $n \in \mathbb{N}^*$. Pour cette étude, on utilise les graphes des fonctions $T^{\circ n}$ et leurs intersections avec la première bissectrice (*cf.* la figure IV.1.1 de la page suivante). Le graphe de T est « en forme de tente », ce qui justifie la terminologie. Intuitivement T consiste à étirer un segment élastique jusqu'à deux fois sa longueur puis à le replier sur lui même en fixant le milieu[6]. La fonction $T^{\circ n}$ est continue, affine par morceaux, et l'on vérifie facilement par récurrence que son graphe est obtenu en « juxtaposant 2^{n-1} tentes » dont la « base » a une longueur de 2^{-n} (*cf.* la figure IV.1.1).

- Les points fixes de T sont 0 et $2/3$.

[6]Penser à un boulanger – traditionnel – pétrissant la pate à pain : il existe une variante en deux dimensions de T appelée transformation du boulanger.

- On a :

  $T \circ T(x) = 4x$, si $0 \leqslant x \leqslant 1/4$, $T \circ T(x) = 2(1-2x)$, si $1/4 \leqslant x \leqslant 1/2$,

  $T \circ T(x) = 2(-1+x)$, si $1/2 \leqslant x \leqslant 3/4$, $T \circ T(x) = 4(1-x)$, si $3/4 \leqslant x \leqslant 1$.

  Les points fixes de $T \circ T$ sont les $4 = 2^2$ points 0, $\frac{2}{5}$, $\frac{2}{3}$, $\frac{4}{5}$.
  Par suite T admet un unique 2-*cycle* : $(\frac{2}{5}, \frac{4}{5})$.

- En explicitant $T^{\circ 3}$, on vérifie que ses points fixes sont les $8 = 2^3$ points 0, $\frac{2}{9}$, $\frac{2}{7}$, $\frac{4}{9}$, $\frac{4}{7}$, $\frac{6}{9}$, $\frac{6}{7}$, $\frac{8}{9}$. Les deux points 0 et $\frac{6}{9} = \frac{2}{3}$ sont les points fixes de T et les 6 points restants sont nécessairement la réunion de deux trois-cycles disjoints (on utilise les remarques ci-dessus). On vérifie facilement que T admet les deux 3-*cycles* : $(\frac{2}{9}, \frac{4}{9}, \frac{8}{9})$ et $(\frac{2}{7}, \frac{4}{7}, \frac{6}{7})$.

- La première bissectrice coupe chaque tente du graphe de $T^{\circ n}$ en deux points. On en déduit facilement que $T^{\circ n}$ admet 2^n points fixes.

On montrera plus loin que, pour tout $n \in \mathbb{N}$, l'application T admet un n-cycle (ce n'est pas du tout évident…; *cf.* l'exemple 4 de la page 577). Pour tout nombre premier p, on peut calculer le nombre de p-cycles de T ; par exemple T possède 7710 17-cycles (*cf.* l'exercice IV.1.28).

FIGURE IV.1.1 – Application tente

2.1.5 Suites récurrentes classiques

Nous allons maintenant présenter les exemples « classiques » de suites récurrentes : les suites arithmétiques et géométriques, puis les suites arithmético-géométriques et homographiques. Les deux premiers exemples ont déjà été rencontrés par le lecteur dans l'enseignement secondaire. Ces suites sont toutes des cas particuliers de récurrence homographique, c'est-à-dire du cas où $f(x) := \frac{ax+b}{cx+d}$, avec $ad - bc \neq 0$. Si $c = 0$ et $a = d = 1$ on obtient une suite arithmétique. Si $b = c = 0$ on obtient une suite géométrique. Si $c = 0$ et $ab \neq 0$, on obtient une suite arithmético-géométrique. Si $c \neq 0$, on obtient une suite homographique ne faisant pas partie des familles précédentes. Les cas arithmétiques, géométriques et arithmético-géométriques sont simples. Le cas homographique est plus délicat ; certains résultats seront traités en exercice.

Suites arithmétiques

Définition 16. Soit $r \in \mathbb{K}$. On appelle *suite arithmétique de raison* r toute suite $(u_n) \in \mathbb{K}^{\mathbb{N}}$ vérifiant $\forall n \in \mathbb{N}\,,\ u_{n+1} = u_n + r$.

Une telle suite est définie par la fonction $f : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & x+r \end{array}$ (translation) et la condition initiale $u_0 \in \mathbb{R}$. Notons que f est *bijective* et que son inverse $f^{-1} : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & x-r \end{array}$ définit une suite arithmétique de raison $-r$.

Ce type de suite très simple se rencontre dans de nombreuses applications pratiques : un compte bancaire sur lequel on verse une somme constante à intervalles constants, sans produire d'intérêts, a un encours qui est une suite arithmétique (tant que, du moins, l'on ne retire pas d'argent !). On parle donc aussi de *progression arithmétique*. La suite $(n)_{n\in\mathbb{N}}$ des nombres naturels est l'exemple le plus simple de suite arithmétique ($u_0 := 0,\ r := 1$).

On prouve facilement, par récurrence, le résultat suivant.

Proposition 19. Si (u_n) est une suite arithmétique de premier terme a et de raison r, alors $u_n = a + nr$ pour tout $n \in \mathbb{N}$.

Dans un certain nombre d'applications, on a besoin de connaître la somme des $n+1$ premiers termes de la suite : $\sum\limits_{k=0}^{n} u_k = u_0 + u_1 + \cdots + u_n$.

Elle est donnée par la proposition suivante.

Proposition 20. Soit (u_n) une suite arithmétique de premier terme a et de raison r. Alors :

$$\sum_{k=0}^{n} u_k = (n+1)a + r\frac{n(n+1)}{2}.$$

Démonstration. On sait que $u_k = a + kr$. Donc :

$$\sum_{k=0}^{n} u_k = \sum_{k=0}^{n}(a+kr) = (n+1)a + r\sum_{k=0}^{n} k.$$

On est ramené au calcul de $S_n := \sum\limits_{k=0}^{n} k$. On peut montrer facilement par récurrence que $S_n := \frac{n(n+1)}{2}$, mais on peut aussi prouver ce résultat en écrivant les entiers deux fois dans deux sens différents :

$$\begin{aligned} 2\sum_{k=0}^{n} k =& 1 + 2 + \cdots + (n-1) + n \\ &+ n + (n-1) + \cdots + 2 + 1. \end{aligned}$$

On peut alors additionner par colonne, chaque terme de la première ligne avec celui qui correspond dans la seconde : on obtient toujours $n+1$, et il y a n termes au total, d'où une somme égale à $n(n+1)$.

On dit que le mathématicien allemand C.F. Gauß avait, très jeune (vers 1780), stupéfié son maître d'école en inventant presque instantanément cette dernière méthode pour calculer la somme des entiers de 1 à 100. ∎

Remarque. La somme des n premiers entiers est $S_1(n) := 1+2+\cdots+n = \frac{n(n+1)}{2}$. Plus généralement, il existe des formules, dues à Blaise Pascal, Johann Faulhaber et Jakob Bernoulli[7], permettant de calculer, pour $k \in \mathbb{N}$, la somme $S_k(n) := 1^k + 2^k + \cdots + n^k$ des puissances $k^{\text{èmes}}$ des n premiers entiers sous forme de polynôme en n (*cf.* les exercices IV.1.24, IV.1.25 et IV.1.61). La formule de Pascal permet de calculer les $S_k(n)$ par récurrence sur k. Celle de Faulhaber et Bernoulli permet de calculer $S_k(n)$ sans récurrence en utilisant des nombres rationnels dits de Bernoulli B_n (*cf.* l'exercice 11 pour leur définition).

Exercice 12.

Montrer que si (u_n) est une suite arithmétique, alors pour tout $(p,q) \in \mathbb{N}^2$ tel que $p \leqslant q$, on a :

$$\Sigma := \sum_{k=p}^{q} u_k = (q-p+1)\,\frac{u_p + u_q}{2}.$$

Solution. On a $\Sigma = (q-p+1)a + \frac{(q-p+1)(p+q)}{2} = (q-p+1)\frac{u_p+u_q}{2}$

Exercice 13.

Montrer qu'une suite arithmétique ne peut être bornée que si sa raison est nulle.

Solution. Une telle suite s'écrit $u_n := a + nr$. Supposons $|u_n| \leqslant A$ pour tout n.
Alors $nr \leqslant A - a$ et $n(-r) \leqslant A + a$. Si $r \neq 0$, alors $r > 0$ ou $-r > 0$ et ceci contredirait le fait que $\mathbb{R}$ est archimédien.

Suites géométriques

Définition 17. Soit $r \in \mathbb{K}$. On appelle *suite géométrique de raison* r toute suite $(u_n) \in \mathbb{K}^{\mathbb{N}}$ vérifiant $\forall n \in \mathbb{N},\ u_{n+1} = r\,u_n$.

Cette fois chaque terme est déduit du précédent par multiplication par un nombre constant r.
Les suites constantes sont des suites géométriques (de raison 1).
Une suite géométrique est définie par la fonction $f : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & rx \end{array}$ (homothétie) et la condition initiale $u_0 \in \mathbb{R}$. Si $r \neq 0$, f est *bijective* et son inverse $f^{-1} : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & r^{-1}x \end{array}$ définit une suite géométrique de raison r^{-1}.
Un livret d'épargne sur lequel on n'effectue pas de nouveaux versements, mais qui produit des intérêts annuels, au taux de 2% par exemple, a un montant qui suit une progression géométrique de raison $1{,}02$.

[7] *Potestatum Numericarum Summa* 1665, *Academia Algebrae* 1631 et *Ars Conjectandi* 1713.

On vérifie immédiatement le résultat suivant.

Proposition 21. Si (u_n) est une suite géométrique de premier terme a et de raison r, alors $\forall n \in \mathbb{N}\,,\ u_n = a\,r^n$.

Proposition 22.
Si le premier terme d'une suite géométrique est nul tous les termes sont nuls. Une suite géométrique non nulle ne peut être bornée que si sa raison r vérifie $|r| \leqslant 1$.

Démonstration. Si $|r| \leqslant 1$ la suite est évidemment bornée. Supposons inversement qu'une suite géométrique soit bornée. Autrement dit qu'il existe $\mu \geqslant 0$ tel que $|ar^n| \leqslant \mu$ pour tout n. Comme $|ar^n| = |a||r|^n$ et $a \neq 0$ puisque nous avons supposé la suite non nulle, on a donc $|r|^n \leqslant \mu/|a|$.
La propriété d'Archimède multiplicative (proposition 14 de la page 520) nous dit alors que $|r| \leqslant 1$, sinon il existerait un entier n tel que $|r|^n > \mu/|a|$. ■

Si $a \neq 0$, la suite géométrique (ar^n) est τ-périodique si, et seulement si, $r^\tau = 1$ (*i.e.* r est une racine $\tau^{\text{ème}}$ de l'unité). Si $\mathbb{K} := \mathbb{R}$, et si τ est la plus petite période, les seules possibilités sont $\tau = 1$ et $\tau = 2$ ($r = 1$ et $r = -1$). Si $\mathbb{K} := \mathbb{C}$, la plus petite période τ peut prendre toutes les valeurs possibles : il suffit de prendre pour r une racine $\tau^{\text{ème}}$ primitive de l'unité, (*cf.* le module II.5, définition 13 de la page 260).
Comme pour les suites arithmétiques, on sait calculer la somme de termes consécutifs d'une suite géométrique :

Proposition 23. Soit (u_n) une suite géométrique de raison $r \neq 1$. Alors, pour tout (p, q) tel que $p \leqslant q$, on a :

$$\sum_{k=p}^{q} u_k = \frac{u_p - u_{q+1}}{1 - r}\cdot$$

Il faut que r soit différent de 1, sinon le second membre ne veut rien dire. Dans le cas $r := 1$, la suite est constante et la somme vaut $(q - p + 1)\,a$, où a est la valeur de cette constante.

Démonstration. Soit $S := u_p + u_{p+1} + \cdots + u_q$. On a $r\,S = u_{p+1} + u_{p+2} + \cdots + u_{q+1}$ puisque $\forall k \in \mathbb{N}\,,\ u_{k+1} = r\,u_k$. Donc $S - r\,S = u_p - u_{q+1}$, ce qui donne le résultat lorsque $r \neq 1$. ■

En particulier, pour $r \neq 1$ et $n \in \mathbb{N}^*$, on a $\displaystyle\sum_{k=0}^{n} r^k = \frac{1 - r^{n+1}}{1 - r}\cdot$

Exercice 14.

Soient x et y deux réels et $n \in \mathbb{N}$. Calculer $C_n := \sum\limits_{k=0}^{n} \cos(x + ky)$.

Solution. On introduit $S_n := \sum_{k=0}^{n} \sin(x+ky)$ de telle sorte que :

$$U_n := C_n + i\,S_n = \sum_{k=0}^{n} e^{i(x+ky)}.$$

On a $e^{i(x+ky)} = e^{ix}e^{iky} = e^{ix}(e^{iy})^k$, donc U_n est la somme de n termes consécutifs d'une suite géométrique de raison $r := e^{iy}$.

- Si $y \equiv 0\ [2\pi]$, alors $C_n = (n+1)\,\cos x$ et $S_n = (n+1)\,\sin x$.
- Sinon, on a $r \neq 1$ et donc :

$$U_n = \frac{e^{ix} - e^{i\left(x+(n+1)y\right)}}{1-e^{iy}} = \frac{e^{i\left(x+\frac{n+1}{2}y\right)}\left(e^{-i\frac{n+1}{2}y} - e^{i\frac{n+1}{2}y}\right)}{e^{i\frac{y}{2}}\left(e^{-i\frac{y}{2}} - e^{i\frac{y}{2}}\right)}$$

$$= e^{i\left(x+\frac{n}{2}y\right)}\,\frac{-2i\,\sin\frac{n+1}{2}y}{-2i\,\sin\frac{y}{2}} = e^{i\left(x+\frac{n}{2}y\right)}\,\frac{\sin\frac{n+1}{2}y}{\sin\frac{y}{2}}.$$

En prenant les parties réelle et imaginaire, on obtient :

$$C_n = \cos\left(x+\frac{n}{2}y\right)\,\frac{\sin\frac{n+1}{2}y}{\sin\frac{y}{2}} \qquad \text{et} \qquad S_n = \sin\left(x+\frac{n}{2}y\right)\,\frac{\sin\frac{n+1}{2}y}{\sin\frac{y}{2}}.$$

Il y a de nombreux liens entre suites arithmétiques et géométriques ; les exercices suivants en donnent deux exemples.

Exercice 15.

Soit (u_n) une suite géométrique de premier terme a et de raison r, avec $a \in \mathbb{R}_+^*$ et $r \in \mathbb{R}_+^*$. Montrer que $u_n > 0$ pour tout $n \in \mathbb{N}$, et que la suite $(\ln u_n)$ est une suite arithmétique de raison $\ln r$ et de premier terme $\ln a$.

Solution. La relation $u_n > 0$ est vraie si $n := 0$ car $a > 0$. Si elle est vraie pour n, alors $u_{n+1} = ru_n > 0$, ce qui prouve la relation pour tout n par récurrence.

Il existe donc $v_n := \ln u_n$ et la relation $u_{n+1} = r\,u_n$ donne $v_{n+1} = v_n + \ln r$, ce qui montre que la suite (v_n) est arithmétique de raison $\ln r$ et de premier terme $\ln a$.

Exercice 16.

Soit (u_n) une suite géométrique de premier terme a et de raison r, avec $a \in \mathbb{C}^*$ et $r \in \mathbb{C}^*$. Montrer que $(|u_n|)$ est aussi une suite géométrique, et que des arguments bien choisis des u_n forment une suite arithmétique ; préciser les premiers termes et les raisons de ces deux suites.

Solution. Le module de u_n vaut $|u_n| = |ar^n| = |a|\,|r^n| = |a|\,|r|^n$. C'est bien la forme d'une suite géométrique de premier terme $|a|$ et de raison $|r|$.

Posons $a = |a|\,\mathrm{e}^{\mathrm{i}\alpha}$ et $r = |r|\,\mathrm{e}^{\mathrm{i}\beta}$. Il vient $u_n = |ar^n|\,\mathrm{e}^{\mathrm{i}(\alpha+n\beta)}$. La suite $w_n := \alpha + n\beta$ est formée de nombres dont chacun est un argument du terme u_n de même indice. Elle est arithmétique, de premier terme α, l'un des arguments de a, et de raison β, l'un des arguments de r.

Suites arithmético-géométriques

Une suite *arithmético-géométrique* est une suite qui vérifie une relation de récurrence de la forme $u_{n+1} = a\,u_n + b$, avec $(a, b) \in \mathbb{K}^2$.
Une telle suite est définie par la fonction $f : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & ax+b \end{array}$ (transformation affine) et la condition initiale $u_0 \in \mathbb{R}$. Si $a \neq 0$, f est *bijective* et son inverse $f^{-1} : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ t & \longmapsto & a^{-1}t - a^{-1}b \end{array}$ définit une suite arithmético-géométrique.

Cas particuliers. Ce sont :

- les suites arithmétiques de raison b lorsque $a = 1$ (en particulier constantes si $b = 0$),
- les suites géométriques de raison a lorsque $b = 0$.

Si $a \neq 1$, l'étude de la suite *se ramène à celle d'une suite géométrique*.
Supposons $a \neq 1$ et considérons une suite (u_n) vérifiant $\forall n \in \mathbb{N},\ u_{n+1} = a\,u_n + b$. Il existe un unique élément $\alpha \in \mathbb{K}$ tel que $a\,\alpha + b = \alpha$: $\alpha := b/(1-a)$. La suite (v_n) définie par $v_n := u_n - \alpha$ vérifie alors, pour tout $n \in \mathbb{N}$:

$$v_{n+1} = u_{n+1} - \alpha = a\,u_n + b - (a\,\alpha + b) = a\,v_n.$$

C'est une suite géométrique de raison a, ce qui permet de calculer v_n, puis u_n :

$$\forall n \in \mathbb{N},\ v_n = v_0\,a^n \qquad \text{et donc} \qquad u_n = \alpha + a^n\,(u_0 - \alpha).$$

Il est utile pour la suite de formaliser un peu ce que nous venons de faire dans ce cas très simple. Nous avons utilisé un changement de coordonnées : $h : x \mapsto y := x - \alpha$, d'inverse $h^{-1} : y \mapsto x := y + \alpha$. On a $v_n = h(u_n)$, $u_n = h^{-1}(v_n)$, donc :

$$v_{n+1} = h(u_{n+1}) = h \circ f(u_n) = h \circ f \circ h^{-1}(v_n).$$

En posant $g := h \circ f \circ h^{-1}$, on a $g(y) = ay$. On a ainsi, par changement de coordonnées, remplacé l'étude de la récurrence définie par f par celle de la récurrence définie par g. Nous utiliserons une démarche similaire dans le pararaphe suivant pour l'étude des suites homographiques.

Suites homographiques

Une suite *homographique* est une suite qui vérifie une relation de récurrence de la forme $u_{n+1} = \frac{a\,u_n+b}{c\,u_n+d}$, avec $a, b, c, d \in \mathbb{K}$ et $(c, d) \neq (0, 0)$.
Lorsque $c = 0$ (et donc $d \neq 0$), il s'agit simplement d'une suite arithmético-géométrique. On peut donc supposer $c \neq 0$. Si $ad - bc = 0$, la suite est stationnaire (à partir de $n = 1$). Nous supposerons toujours dans la suite $ad - bc \neq 0$.
Contrairement aux exemples précédents où les suites étaient définies par une relation $u_{n+1} = f(u_n)$, avec $f : \mathbb{K} \to \mathbb{K}$, ce qui assurait la définition de la suite, on n'est pas assuré de l'existence : il faut, pour cela, trouver une partie X de $\mathbb{K}$, stable par l'homographie $f : x \mapsto \frac{a\,x+b}{c\,x+d}$ et contenant le premier terme. Ce n'est pas toujours facile et l'on se contente souvent de vérifier que pour un « bon » choix de u_0, la suite est bien définie.

La fonction f n'est pas définie en $x := -d/c$. Il faut donc supposer $u_0 \neq -d/c$ et plus généralement qu'il n'existe pas de $n \in \mathbb{N}$ tel que $u_n = -d/c$.
On vérifie que f définit une bijection de $\mathbb{R} \setminus \{-d/c\}$ sur $\mathbb{R} \setminus \{a/c\}$ dont l'inverse est l'homographie $\varphi : y \mapsto \frac{-dy+b}{cy-a}$. Si $a + d = 0$, la situation est simple : $-d/c = a/c$ et f définit une *bijection* de $X := \mathbb{R} \setminus \{-d/c\}$ sur lui même ; ainsi la suite homographique est définie pour tout $u_0 \neq -d/c$ (et $\varphi = f$). Le cas général est plus difficile. On peut décrire précisément l'ensemble des points u_0 pour lesquels la suite est définie mais cette description est un peu délicate. Elle peut être omise en première lecture.
Pour que la suite (u_n) soit définie, il faut et il suffit qu'il n'existe pas d'entier $m \in \mathbb{N}^*$ tel que $u_0 = \varphi^{\circ m}(-d/c)$. Notons Z l'ensemble des points ω de $\mathbb{K}$ tels qu'il existe $m \in \mathbb{N}$ tel que $\omega = \varphi^{\circ m}(-d/c)$. C'est l'image d'une suite finie ou infinie, donc Z est fini ou infini dénombrable. En particulier, s'il existe m tel que $\varphi^{\circ m}(-d/c)$ soit défini et égal à a/c, alors $\varphi^{\circ m+1}(-d/c)$ n'est pas défini et la suite est finie : $\operatorname{card} Z = m + 1$. Soit $X := \mathbb{K} \setminus Z$; la suite (u_n) est définie si, et seulement si, $u_0 \in X$. On a $f(X) \subset X$: la partie X est stable par f. Elle n'est pas vide et elle est même infinie non dénombrable, puisque Z est dénombrable et que $\mathbb{K}$ ne l'est pas (*cf.* la proposition 6 de la page 516). La partie X de $\mathbb{K}$ peut être « compliquée » ; par exemple, dans le cas réel, il est possible qu'elle ne contienne aucun intervalle ouvert non vide (*cf.* l'exercice 36 de la page 576).

Exemple 1. Dans la solution de l'exemple 3 de la page 511, nous avons considéré, pour un rationnel positif a donné, le rationnel $b := \dfrac{2a+2}{a+2}\cdot$ En réitérant l'opération, on obtient une suite homographique (u_n) vérifiant $\forall n \in \mathbb{N},\ u_{n+1} = f(u_n)$, avec $f : x \mapsto \frac{2x+2}{x+2}\cdot$
La fonction f est évidemment définie sur $\mathbb{R}_+$ et à valeurs dans $\mathbb{R}_+$, ce qui permet d'assurer, pour tout $a \in \mathbb{R}_+$, l'existence d'une suite (u_n) vérifiant :

$$u_0 = a \qquad \text{et} \qquad \forall n \in \mathbb{N},\ u_{n+1} = f(u_n).$$

Pour une telle suite, on peut exprimer u_n en fonction de n de la façon suivante. Comme pour les suites arithmético-géométriques, on commence par chercher les points fixes de f, c'est-à-dire les éléments $x \in \mathbb{K}$ tels que $f(x) = x$. Un calcul élémentaire aboutit à $f(x) - x = \dfrac{2 - x^2}{x+2}$, ce qui donne les deux points fixes $\alpha := \sqrt{2}$ et $-\sqrt{2} = -\alpha$.
Si $u_0 = \pm\alpha$, la suite est constante. Supposons donc $u_0 \neq \pm\alpha$. Soit $x \in \mathbb{R}_+$. On a alors $f(x) - \alpha = f(x) - f(\alpha) = \dfrac{2(x-\alpha)}{(2+\alpha)\,(x+2)}$ et $f(x) + \alpha = \dfrac{2(x+\alpha)}{(2-\alpha)\,(x+2)}\cdot$
Il vient donc $\dfrac{f(x)-\alpha}{f(x)+\alpha} = \dfrac{2-\alpha}{2+\alpha}\,\dfrac{x-\alpha}{x+\alpha}$, pour $x \neq -\alpha$, ce qui donne :

$$\forall n \in \mathbb{N},\ \frac{u_{n+1}-\alpha}{u_{n+1}+\alpha} = k\,\frac{u_n-\alpha}{u_n+\alpha} \qquad \text{avec} \qquad k := \frac{2-\alpha}{2+\alpha} = \frac{\sqrt{2}-1}{\sqrt{2}+1}\cdot$$

La suite (v_n) définie par $v_n := \dfrac{u_n-\alpha}{u_n+\alpha}$ est géométrique de raison k, ce qui donne v_n en fonction de n :

$$\forall n \in \mathbb{N},\ v_n = v_0\,k^n = \frac{a-\sqrt{2}}{a+\sqrt{2}}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^n.$$

On peut obtenir une expression de u_n en fonction de n : la relation $v_n = \dfrac{u_n - \alpha}{u_n + \alpha}$ donne $v_n \neq 1$ et $u_n = \alpha\,\frac{1+v_n}{1-v_n} = \sqrt{2}\,\frac{1+v_n}{1-v_n}$.

L'inverse de f est l'application, également homographique, $\varphi : y \mapsto \frac{2y-2}{-y+2}$. Les valeurs interdites de u_0 sont $\gamma_0 := -2$ (pour la définition de u_1), $\gamma_1 := \varphi(-2) = -3/2$ (pour la définition de u_2), $\gamma_2 := \varphi(-3/2) = -10/7\ldots$ On obtient une suite homographique dont il faudrait étudier la définition. Pour montrer que cette suite est bien définie et la calculer on procède « indirectement ». Les valeurs interdites de u_0 correspondent aux valeurs de $v_0 = \frac{u_0-\sqrt{2}}{u_0+\sqrt{2}}$ telles que $k^q v_0 = 1$ pour un $q \in \mathbb{N}^*$, c'est-à-dire $v_0 \in \{k^{-q} \mid q \in \mathbb{N}^*\}$.

On obtient $\gamma_p = \sqrt{2}\,\dfrac{1+k^{-(p+1)}}{1-k^{-(p+1)}} = \sqrt{2}\,\dfrac{k^{p+1}+1}{k^{p+1}-1}\cdot$ On vérifie que les valeurs interdites forment un ensemble *infini*. On a, pour tout $p \in \mathbb{N}$, $\gamma_p = \varphi^{\circ p}(-2)$; on en déduit que les γ_p sont *rationnels*.

Exemple 2. Considérons la suite définie (?) par son premier terme u_0 et la relation de récurrence $\forall n \in \mathbb{N}\,,\ u_{n+1} = \dfrac{2u_n - 1}{u_n}\cdot$

- Tout d'abord, étudions pour quelles valeurs de u_0, la suite est bien définie. Il est évident qu'il faut $u_0 \neq 0$ pour que l'on puisse calculer $u_1 = (2u_0 - 1)/u_0$. Mais il ne faut pas non plus que u_1 soit nul, ce qui impose $u_0 \neq 1/2$ pour la définition de u_2.
  Les valeurs interdites de u_0 sont donc $\alpha_0 := 0$ (pour la définition de u_1), $\alpha_1 := 1/2$ (pour la définition de u_1). Ensuite, il ne faut pas que u_1 soit égal à α_1, ce qui exclut α_2 tel que $\frac{2\alpha_2-1}{\alpha_2} = \frac{1}{2}\cdots$ Ces valeurs constituent ainsi une suite (α_p) telle que $\frac{2\alpha_{p+1}-1}{\alpha_{p+1}} = \alpha_p$, ce qui donne $\alpha_{p+1} = \frac{1}{2-\alpha_p}\cdot$
  Cela conduit à une autre suite homographique (α_p) dont il faudrait étudier la définition. Mais ici c'est simple, car l'application $x \mapsto 1/(2-x)$ laisse stable l'intervalle $[0,1]$ qui contient le premier terme α_0. En calculant les premiers termes : $0,\ 1/2,\ 2/3,\ 3/4,\ldots$ on devine que $\alpha_p = p/(p+1)$, ce qu'il est aisé de démontrer par récurrence.
- Cherchons les points fixes de $f : x \mapsto (2x-1)/x$. Cela se ramène à l'équation $x^2 - 2x + 1 = 0$, dont la seule racine (double) est 1. Si $u_0 = 1$, la suite est constante.
  Soit $x \neq 1$. On a $f(x) - 1 = \dfrac{x-1}{x}$ ce qui donne $\dfrac{1}{f(x)-1} = \dfrac{x}{x-1} = \dfrac{1}{x-1} + 1$.
  Si $u_0 \neq 1$, on a donc $\forall n \in \mathbb{N}\,,\ \dfrac{1}{u_{n+1}-1} = \dfrac{1}{u_n - 1} + 1$, ce qui donne, pour tout $n \in \mathbb{N}$:
  $$\frac{1}{u_n - 1} = \frac{1}{u_0 - 1} + n \qquad \text{soit} \qquad u_n = 1 + \frac{u_0 - 1}{1 + n(u_0 - 1)}\cdot$$

Nous allons voir que les méthodes employées dans les deux exemples précédents pour exprimer $u_n = f(u_{n-1})$ $(n \in \mathbb{N}^*)$ en fonction de n et u_0 se généralisent à toutes les suites homographiques du type $u_{n+1} = f(u_n)$ avec $f : x \mapsto \frac{ax+b}{cx+d}$, $c \neq 0$ et $ad - bc \neq 0$.

L'équation $f(x) = x$ est équivalente à l'équation du second degré $cx^2 + (d-a)x - b = 0$.

On distingue le cas où cette équation a deux racines *distinctes* α et β et celui où elle possède une *racine double* α.

Cas des racines distinctes. Si $u_0 := \alpha$ ou $u_0 = \beta$, la suite est constante.
Supposons $u_0 \neq \beta$. On procède alors comme dans l'exemple 1 ci-dessus en « remplaçant, par changement homographique h de coordonnées » l'homographie f par l'application $h \circ f \circ h^{-1} = g : y \mapsto ky$, c'est-à-dire en se ramenant à une suite géométrique.
On a :

$$u_{n+1} - \alpha = \frac{au_n + b}{cu_n + d} - \alpha = \frac{au_n + b}{cu_n + d} - \frac{a\alpha + b}{c\alpha + d} = \frac{ad - bc}{cu_n + d}\,\frac{u_n - \alpha}{c\alpha + d}. \tag{1}$$

De même $u_{n+1} - \beta = \dfrac{ad - bc}{cu_n + d}\,\dfrac{u_n - \beta}{c\beta + d}$, d'où : $\dfrac{u_{n+1} - \alpha}{u_{n+1} - \beta} = \dfrac{c\beta + d}{c\alpha + d}\,\dfrac{u_n - \alpha}{u_n - \beta}$.
Puisque $ad - bc \neq 0$, on a (le lecteur le vérifiera) $c\alpha + d \neq 0$ et $c\beta + d \neq 0$ et :

$$\frac{u_{n+1} - \alpha}{u_{n+1} - \beta} = k\,\frac{u_n - \alpha}{u_n - \beta}, \quad \text{avec} \quad k := \frac{c\beta + d}{c\alpha + d} \in \mathbb{K}^*. \tag{2}$$

Puisque $c \neq 0$ et $\alpha \neq \beta$, on a $k \neq 1$. Si l'on pose $v_n := h(u_n) := \frac{u_n - \alpha}{u_n - \beta}$, (2) s'écrit $v_{n+1} = kv_n$. La suite (v_n) est donc *géométrique* et $v_n = k^n v_0$. Inversement, on peut écrire u_n en fonction de v_n : $u_n = h^{-1}(v_n) := \frac{\beta v_n - \alpha}{v_n - 1}$. Puisque v_n est défini pour tout $n \in \mathbb{N}$, on peut calculer u_n si $v_n \neq 1$. Ainsi, la suite $(u_n)_{n\in\mathbb{N}}$ est bien définie si, et seulement s'il n'existe pas de $m \in \mathbb{N}$ tel que $k^m v_0 = 1$, c'est-à-dire si $v_0 \notin \{k^{-m} \mid m \in \mathbb{N}\}$. On vérifie que $h(-d/c) = k^{-1}$ et $h^{-1}(k^{-1}) = -d/c$, c'est-à-dire que $u_0 = -d/c$ si, et seulement si, $v_0 = k^{-1}$.
Si $\mathbb{K} = \mathbb{R}$ et si α et β ne sont pas réels, ce sont des nombres complexes conjugués et $\bar{k} = k^{-1}$, donc $1 = k\bar{k} = |k|^2$: k est de module 1.
On notera que k dépend de la façon de nommer les racines. Si l'on échange les rôles de α et β, k est changé en $1/k$.
Soit $\tau \in \mathbb{N}$, $\tau \geqslant 2$. Le nombre $k \in \mathbb{C}$ est une racine $\tau^{ème}$ primitive de l'unité si, et seulement si, toutes les suites homographiques définies par f et $u_0 \neq \alpha$ et β sont périodiques de plus petite période τ. On peut montrer que, pour tout $\tau \in \mathbb{N}$, $\tau \geqslant 2$, il existe une homographie *réelle* de plus petite période τ (*cf.* l'exercice IV.1.53 de la page 610). *Cas d'une racine double.* Si $u_0 := \alpha$, la suite est constante. Supposons $u_0 \neq \alpha$. On procède alors comme dans l'exemple 2 ci-dessus en « remplaçant, par changement homographique h de coordonnées » l'homographie f par $h \circ f \circ h^{-1} = g : y \mapsto y + r$, c'est-à-dire en se ramenant à une suite arithmétique.
On a $\alpha = \frac{a-d}{2c}$, $ad - bc = \frac{(a+d)^2}{4}$ et $c\alpha + d = \frac{a+d}{2}$. La relation (1) reste vraie. On a donc, comme dans l'exemple 2 ci-dessus :

$$\begin{aligned}\frac{1}{u_{n+1} - \alpha} &= \frac{1}{u_n - \alpha}\,\frac{(c\alpha + d)(cu_n + d)}{ad - bc} = \frac{1}{u_n - \alpha}\,\frac{2(cu_n + d)}{a + d}\\ &= \frac{1}{u_n - \alpha}\,\frac{2\,(c(u_n - \alpha) + c\alpha + d)}{a + d} = \frac{1}{u_n - \alpha} + \frac{2c}{a + d}.\end{aligned}$$

Notons $r := \frac{2c}{a+d}$. Si l'on pose $v_n := h(u_n) := \frac{1}{u_n - \alpha}$, on obtient $v_n = v_0 + nr$. La suite

(v_n) est *arithmétique*. Inversement, on peut calculer u_n en fonction de v_n :

$$u_n = h^{-1}(v_n) = \frac{1}{v_n} + \alpha \qquad \text{si } v_n \neq 0.$$

Ainsi, la suite $(u_n)_{n\in\mathbb{N}}$ est bien définie si, et seulement si, $u_0 = \alpha + \frac{1}{v_0}$, avec $v_0 \notin \{-nr\}_{n\in\mathbb{N}}$. On vérifie que $u_0 = -d/c$ si, et seulement si, $v_0 = -r$.

Point Méthode

Étude des suites homographiques du type : $u_{n+1} = f(u_n)$ *avec* $f : x \mapsto \frac{a\,x+b}{c\,x+d}$.
On suppose $ad - bc \neq 0$ et $c \neq 0$.

- On cherche *les points fixes* de f, ce qui se ramène à la résolution d'une équation du second degré.
- Si cette équation possède *deux racines distinctes* α et β, la suite $(v_n) = \left(\frac{u_n-\alpha}{u_n-\beta}\right)$ est, quand elle est définie, *géométrique*, de raison $k := \frac{c\beta+d}{c\alpha+d}$.
- Si elle possède *une racine double* α, la suite $(v_n) = \left(\frac{1}{u_n-\alpha}\right)$ est, quand elle est définie, *arithmétique*, de raison $r := \frac{2c}{a+d}$.
- Si $\mathbb{K} = \mathbb{R}$ et si l'équation n'a pas de racine réelle, on considère les deux racines complexes (conjuguées) et l'on est dans le premier cas.
- *Si la suite* (u_n) *est définie*, on peut, dans les deux cas, exprimer u_n en fonction de v_n, ce qui permet de calculer u_n en fonction de n et de u_0. On a $u_n = \frac{\beta v_n-\alpha}{v_n-1}$ ou $u_n = \frac{1}{v_n} + \alpha$.

2.2 Convergence d'une suite

Pour des raisons théoriques ou dans les applications des suites, l'une des questions les plus naturelles que l'on est amené à se poser sur une suite donnée, est de savoir « où se trouvent les termes lorsque l'on avance suffisamment loin dans les indices » (plus précisèment « quand on considère la suite tronquée à un ordre arbitrairement grand »), c'est ce que l'on appelle le *comportement asymptotique* de la suite. Il est très important, en particulier, de savoir reconnaître les *suites convergentes* et ce sera un point essentiel de notre activité sur les suites dans le reste de ce module.
Mais si, pendant des siècles (voire des millénaires...), le problème central dans l'étude des suites a été celui de leur convergence éventuelle, on peut aussi se poser d'autres questions. Est-ce que la suite est périodique ? Est-ce que la suite est périodique à partir d'un certain rang ? Est-ce que l'ensemble des valeurs de la suite a des propriétés de densité ? Est-ce que la suite est *sensible au choix de la condition initiale* ? (c'est-à-dire que son comportement asymptotique peut beaucoup changer si l'on change très peu le terme initial). En relation avec la montée en puissance des ordinateurs, de telles questions ont fait leur apparition dans l'étude de divers sujets en mathématiques ou dans les sciences expérimentales dans la seconde moitié du XXe siècle. (Par exemple, comme nous l'avons dit plus haut, en relation

avec la théorie du *chaos déterministe*.) L'étude de ces problèmes peut être très difficile et nous nous limiterons à quelques exemples « significatifs » que l'on peut traiter avec des techniques élémentaires.

2.2.1 Suites positives tendant vers 0

Définition 18. On dit qu'une suite de réels positifs *tend vers zéro* si pour tout nombre réel $\varepsilon > 0$, les termes de la suite sont inférieurs ou égaux à ε à partir d'un certain rang.

Exemple. La suite $\left(\frac{1}{n+1}\right)$ tend vers 0.
Soit en effet un réel $\varepsilon > 0$ fixé. Le caractère archimédien de $\mathbb{R}$ nous donne l'existence d'un entier N tel que $N\varepsilon \geqslant 1$ et pour $n \geqslant N$, on a donc $1/(n+1) \leqslant \varepsilon$.

Pour qu'une suite tende vers 0, il n'est pas nécessaire qu'elle soit décroissante, comme on le voit ici :

Exercice 17.
Démontrer que la suite de terme général $u_n := (1 + (-1)^n)/n$ pour $n \geqslant 1$ est positive ou nulle et tend vers zéro, mais n'est pas décroissante.
Solution. On a $u_n = 0$ si n est impair, et $u_n = 2/n$ si n est pair non nul. Chacun de ces termes est positif ou nul. Comme dans l'exemple ci-dessus, la limite est immédiate puisque l'on dispose, pour tout n, de la majoration $u_n \leqslant 2/n$.
Elle n'est pas décroissante, puisque $u_{2k+2} - u_{2k+1} = u_{2k+2} = 1/(k+1) > 0$.

La définition précédente peut se réécrire plus formellement :

$$\forall \varepsilon > 0\,,\ \exists N \in \mathbb{N}\,,\ \forall n \geqslant N\,,\ u_n \leqslant \varepsilon. \tag{3}$$

Il est important de comprendre que le nombre N ici (le « certain rang ») dépend de la valeur ε considérée : on peut, si nécessaire, souligner ce point en écrivant $N(\varepsilon)$ ou N_ε au lieu de N.
On vérifie facilement que (3) est équivalente à :

$$\forall \varepsilon > 0\,,\ \exists N \in \mathbb{N}\,,\ \forall n \geqslant N\,,\ u_n < \varepsilon. \tag{4}$$

Le lecteur pourra se convaincre que l'on peut aussi changer « $n \geqslant N$ » en « $n > N$ » dans (3) ou (4) sans en changer le sens. Ce qui donne quatre façons d'écrire la même propriété : on écrit et on utilise l'une ou l'autre suivant les préférences de chacun.

Proposition 24.
Soit $r \in [0, 1[$. Toute suite géométrique positive de raison r tend vers zéro.

Démonstration. Si l'on note a le premier terme d'une telle suite (u_n), on a $u_n = a\,r^n$ par la proposition 21 de la page 533. Le résultat est évident si $r = 0$ ou $a = 0$; dans la suite, on suppose $r > 0$ et $a > 0$. Posons $t := 1/r$, de sorte que $t > 1$ puisque $0 < r < 1$.
Soit $\varepsilon > 0$. La proposition 14 de la page 520 (propriété d'Archimède multiplicative) nous donne l'existence d'un entier N tel que $t^N \geqslant a/\varepsilon$ et donc aussi $t^n \geqslant a/\varepsilon$ pour tout $n \geqslant N$. Pour tout $n \geqslant N$, on a donc $a\,r^n \leqslant \varepsilon$ c'est-à-dire $u_n \leqslant \varepsilon$.
Donc on a bien montré que $\forall \varepsilon > 0,\ \exists N \in \mathbb{N}\,,\ \forall n \in \mathbb{N}\,,\ u_n \leqslant \varepsilon$. ∎

2.2.2 Limites des suites

Considérons maintenant des suites numériques, donc à valeurs dans $\mathbb{K} = \mathbb{R}$ ou $\mathbb{K} = \mathbb{C}$. Pour $x \in \mathbb{K}$, la notation $|x|$ doit donc être comprise comme le module si $\mathbb{K} = \mathbb{C}$ ou la valeur absolue si $\mathbb{K} = \mathbb{R}$. Nous allons nous ramener à des suites réelles positives par l'intermédiaire de $|u_n|$.

Définition 19. On dit qu'une suite numérique (u_n) *admet* $\ell \in \mathbb{K}$ *comme limite* si la suite de terme général $|u_n - \ell|$ tend vers zéro, autrement dit si :

$$\forall \varepsilon > 0\,,\ \exists N \in \mathbb{N}\,,\ \forall n \geqslant N\,,\ |u_n - \ell| \leqslant \varepsilon. \tag{5}$$

Pour les raisons indiquées précédemment, on peut remplacer dans cette formule l'une et/ou l'autre des deux dernières inégalités larges par des inégalités strictes.

En revanche, on ne peut pas remplacer $\forall \varepsilon > 0$ par $\forall \varepsilon \geqslant 0$. Sinon, en prenant $\varepsilon = 0$, on trouverait un rang N à partir duquel on aurait $|u_n - \ell| \leqslant 0$, soit $u_n = \ell$. Les seules suites admettant ℓ pour limite seraient les suites égales à ℓ à partir d'un certain rang !

Si $\mathbb{K} = \mathbb{R}$, une façon équivalente de formuler la notion de limite d'une suite est de dire que la suite (u_n) *admet* $\ell \in \mathbb{R}$ *comme limite* si, et seulement si, tout intervalle *ouvert* de centre ℓ contient *tous les termes de la suite à partir d'un certain rang*. Si $\mathbb{K} = \mathbb{C}$, on a une formulation analogue en remplaçant les intervalles ouverts de centre ℓ par les disques ouverts de centre ℓ.

Proposition 25 (Unicité de la limite).
Soit (u_n) une suite numérique admettant ℓ_1 et ℓ_2 comme limites, alors $\ell_1 = \ell_2$.

C'est une propriété très simple, mais fondamentale, de la notion de limite. Elle nous permet de parler de *la* limite d'une suite : celle-ci n'existe pas nécessairement, mais si elle existe, il n'y en a pas d'autre possible.

Démonstration. Supposons que (u_n) admette ℓ_1 et ℓ_2 comme limites, avec $\ell_1 \neq \ell_2$. Alors d'après, (5), nous avons, pour tout $\varepsilon > 0$:

$$\exists N_1 \in \mathbb{N}\,,\ \forall n \geqslant N_1\,,\ |u_n - \ell_1| \leqslant \varepsilon, \tag{$*$}$$

$$\exists N_2 \in \mathbb{N}\,,\ \forall n \geqslant N_2\,,\ |u_n - \ell_2| \leqslant \varepsilon. \tag{$**$}$$

Nous avons ici écrit N_1, N_2 pour souligner que le « certain rang » pourrait *a priori* dépendre de la limite considérée. On peut choisir $\varepsilon := |\ell_1 - \ell_2|/3$; on a $\varepsilon > 0$ car nous avons supposé que $\ell_1 \neq \ell_2$.

Donc il existe N_1, N_2, tels que $(*)$ et $(**)$. Choisissons un nombre entier n supérieur à la fois à N_1 et N_2, par exemple $n := \max(N_1, N_2)$. Pour cette valeur de n, nous avons à la fois $|u_n - \ell_1| \leqslant \varepsilon$ et $|u_n - \ell_2| \leqslant \varepsilon$, donc par l'inégalité triangulaire :

$$3\varepsilon = \big|\ell_1 - \ell_2\big| = |(u_n - \ell_2) - (u_n - \ell_1)| \leqslant |u_n - \ell_1| + |u_n - \ell_2| \leqslant 2\varepsilon.$$

Le nombre réel ε vérifie donc à la fois $\varepsilon > 0$ par hypothèse et $3\varepsilon \leqslant 2\varepsilon$ par cette inégalité, ce qui est absurde. ∎

Définition 20. On dit qu'une suite (u_n) est *convergente* si elle admet une limite et dans ce cas celle-ci est notée $\lim u_n$. Dans le cas contraire, on dit que la suite est *divergente*.

Si ℓ est la limite, on dit aussi que la suite converge vers ℓ. La limite est aussi notée $\lim\limits_{n\to+\infty} u_n$ par analogie avec la notation similaire sur les fonctions. Cette notation est plus « lourde », mais utile lorsque le terme général contient plusieurs lettres différentes (considérer par exemple $\lim\limits_{n\to+\infty} \frac{n-m}{n+m} = 1$ et $\lim\limits_{m\to+\infty} \frac{n-m}{n+m} = -1$).

On peut écrire formellement la propriété de convergence :

$$(u_n) \text{ est convergente} \iff \exists \ell \in \mathbb{K}\,,\ \forall \varepsilon > 0\,,\ \exists N \in \mathbb{N}\,,\ \forall n \geqslant N\,,\ |u_n - \ell| \leqslant \varepsilon. \tag{6}$$

On peut facilement déduire de la formule précédente l'expression formelle de la propriété de divergence :

$$(u_n) \text{ est divergente} \iff \forall \ell \in \mathbb{K}\,,\ \exists \varepsilon > 0\,,\ \forall N \in \mathbb{N}\,,\ \exists n \geqslant N\,,\ |u_n - \ell| > \varepsilon. \tag{7}$$

Exercice 18.

Utiliser (7) pour montrer que la suite de terme général $(-1)^n$ est divergente.

Solution. Pour cela, considérons $\ell \in \mathbb{R}$ arbitraire, et remarquons que l'un des deux nombres $|1-\ell|$ ou $|-1-\ell|$ doit être strictement supérieure à $1/2$ (sinon l'inégalité triangulaire donnerait $2 = |(1-\ell)-(-1-\ell)| \leqslant 1/2+1/2 = 1$). Par conséquent, il existe bien un $\varepsilon > 0$, à savoir $\varepsilon = 1/2$, tel que pour tout N on peut trouver $n \geqslant N$ avec $|u_n - \ell| > \varepsilon$ (prendre $n := N$ ou bien $n := N+1$).

Exercice 19.

Soit (u_n) une suite réelle convergente. On suppose que, pour tout $n \in \mathbb{N}$, $u_n \in \mathbb{Z}$. Montrer que (u_n) est stationnaire et que sa limite appartient à $\mathbb{Z}$.

Solution. Soit $\ell := \lim u_n$. On applique la définition de la limite avec $\varepsilon := 1/2$. Alors, il existe $n_0 \in \mathbb{N}^*$ tel que, pour tout $n \geqslant n_0$, l'on ait $|u_n - \ell| < 1/2$. Alors $|u_n - u_{n_0}| \leqslant |u_n - \ell| + |\ell - u_{n_0}| < 1$. Mais un entier de module strictement inférieur à 1 est nul, donc, pour tout $n \geqslant n_0$, $u_n = u_{n_0}$ et la suite est constante à partir de n_0. On a alors $\ell = u_{n_0} \in \mathbb{Z}$.

Exercice 20 (Moyenne de Cesaro).

Soit (u_n) une suite de $\mathbb{K}$. On lui associe la suite (v_n) de K, définie par $v_0 := u_0$ et, pour tout $n \in \mathbb{N}^*$, $v_n := \dfrac{u_1 + u_2 + \cdots + u_n}{n}$. Elle est appelée *moyenne de Cesaro* de (u_n). Montrer que, si (u_n) converge vers ℓ, alors (v_n) converge vers ℓ.

Solution. Supposons que (u_n) converge vers ℓ. Soit $\varepsilon > 0$, il existe $p \in \mathbb{N}^*$ tel que, pour tout $n \geqslant p$, $|u_n - \ell| < \varepsilon/2$. On a, pour tout $n \geqslant p$:

$$\begin{aligned} v_n - \ell &= \frac{u_1 + u_2 + \cdots + u_n}{n} - \ell = \frac{u_1 + u_2 + \cdots + u_n - n\ell}{n} \\ &= \frac{(u_1 - \ell) + (u_2 - \ell) + \cdots + (u_p - \ell)}{n} + \frac{(u_{p+1} - \ell) + (u_2 - \ell) + \cdots + (u_n - \ell)}{n}. \end{aligned}$$

On en déduit :

$$|v_n - \ell| \leqslant \frac{1}{n}\sum_{k=1}^{p}|u_k - \ell| + \frac{1}{n}\sum_{k=p+1}^{n}|u_k - \ell|.$$

Notons $C := \sum_{k=1}^{p}|u_k - \ell|$, on a : $|v_n - \ell| \leqslant \frac{C}{n} + \frac{n-p}{n}\frac{\varepsilon}{2} \leqslant \frac{C}{n} + \frac{\varepsilon}{2}$. Par ailleurs $\lim_{n\to+\infty}\frac{C}{n} = 0$.
Il existe $q \in \mathbb{N}^*$, tel que, pour tout $n \geqslant q$, $\frac{C}{n} < \frac{\varepsilon}{2}$.
En posant $N := \max\{p, q\}$, on obtient : pour tout $n \geqslant N$, $|v_n - \ell| < \varepsilon$.

La réciproque du résultat ci-dessus est en général fausse. La suite (v_n) peut converger sans que la suite (u_n) converge. Voici un exemple. La suite $(u_n := (-1)^n)$ diverge ; par ailleurs $v_n = 0$ si $n \in \mathbb{N}^*$ est pair et $v_n = -\frac{1}{n}$ si n est impair, donc $\lim_{n\to+\infty} v_n = 0$.
Nous terminons par la définition de la notion de *limite infinie*. Nous reviendrons sur cette question dans la partie 2.5.

Définition 21. On dit qu'une suite (u_n) de $\mathbb{K}$ *tend vers l'infini* si la suite $(1/u_n)$ est définie à partir d'un certain rang, et a pour limite zéro. Si de plus la suite est réelle et positive (resp. négative) à partir d'un certain rang, on dit que la suite tend vers $+\infty$ (resp. $-\infty$).

Une suite tendant vers l'infini est divergente, mais une suite peut être divergente sans tendre vers l'infini : *cf.* par exemple l'exercice 18 de la page ci-contre.

Exercice 21.
Soit (u_n) une suite réelle. On suppose qu'elle ne tend pas vers l'infini. Montrer qu'il existe une suite extraite de (u_n) qui est *bornée*.
Solution. S'il existe une suite extraite constante et égale à 0 le problème est résolu. Sinon il existe $n_0 \in \mathbb{N}$ tel que, pour tout $n \geqslant n_0$, l'on ait $u_n \neq 0$. Notons $(v_n := 1/u_n)_{n\geqslant n_0}$. Si (u_n) ne tend pas vers l'infini, (v_n) ne tend pas vers 0. Alors, il existe $\varepsilon_0 > 0$ tel que pour tout $N \in \mathbb{N}$, il existe $n \geqslant N$ vérifiant $|v_n| > \varepsilon_0$. Ceci permet de construire par récurrence une suite $(v_{n(k)})$, extraite de (v_n), telle que, pour tout $k \in \mathbb{N}$, $v_{n(k)} \geqslant \varepsilon_0$: connaissant $n(k)$, on trouve $n(k+1) \geqslant n(k)+1$ tel que $|v_{n(k+1)}| \geqslant \varepsilon_0$ (on utilise $N := n(k)+1$). La suite $(u_{n(k)})$ extraite de (u_n) est bornée par $1/\varepsilon_0$.

2.2.3 Premières propriétés des limites

Rappelons que la propriété de tendre vers zéro est une propriété du type « à partir d'un certain rang », donc la convergence aussi. On traduit cela en parlant du « caractère asymptotique » des notions de convergence et de limite. Ainsi :

Proposition 26. Une suite (u_n) converge si, et seulement si, ses suites tronquées convergent et, dans ce cas, toutes ont la même limite.

En d'autres termes on ne modifie pas la limite d'une suite en supprimant les premiers termes de celle-ci. On peut aussi dire que pour tout entier p, on a $\lim_{n\to+\infty} u_{n+p} = \lim u_n$.

Théorème 27. Toute suite convergente est bornée.

Par conséquent toute suite non bornée est divergente. En revanche, il existe des suites bornées qui ne sont pas convergentes, comme la suite $\big((-1)^n\big)$ vue ci-avant.

Démonstration. Soit (u_n) une suite convergente, de sorte qu'elle vérifie (6). Choisissons une valeur quelconque de ε, disons $\varepsilon = 1$. Il existe donc un N, tel que pour tout $n \geqslant N$, $|u_n - \ell| \leqslant 1$. Cette dernière inégalité implique $|u_n| \leqslant 1 + |\ell|$. Autrement dit, la suite est bornée à partir d'un certain rang (par le nombre $1+|\ell|$) ; mais nous avons vu à l'exercice 10 de la page 525 que cela implique qu'elle est bornée. ■

Le résultat suivant découle immédiatement de la définition d'une suite convergeant vers 0.

Proposition 28. Soient (u_n) une suite numérique et ℓ un nombre. S'il existe une suite (r_n) tendant vers 0 telle que $\forall n \in \mathbb{N}\,,\ |u_n - \ell| \leqslant r_n$, alors $\lim u_n = \ell$.

C'est un moyen pratique pour montrer qu'une suite (u_n) converge vers un réel ℓ donné. On majore $|u_n - \ell|$ pour le rendre plus petit que le $n^{\text{ème}}$ terme d'une suite connue qui tend vers 0.

Exemple. Si (u_n) est une suite qui converge vers ℓ, alors $\big(|u_n|\big)$ converge vers $|\ell|$.
En effet, l'inégalité triangulaire donne : $|u_n| \leqslant |u_n - \ell| + |\ell|$ et $|\ell| \leqslant |\ell - u_n| + |u_n|$, d'où :

$$\big||u_n| - |\ell|\big| \leqslant |u_n - \ell| \underset{n\to+\infty}{\longrightarrow} 0.$$

Le résultat suivant permet de caractériser les bornes supérieures et inférieures d'un sous-ensemble de $\mathbb{R}$ (quand elles existent) à l'aide de la notion de suite convergente. On dit que l'on a une *caractérisation séquentielle* de ces bornes.

Proposition 29. Soient $A \subset \mathbb{R}$ un sous-ensemble non vide et $a \in \mathbb{R}$:

- $a = \sup A$ si, et seulement si, a est un majorant de A et s'il existe une suite (x_n) de A qui converge vers a ;
- $a = \inf A$ si, et seulement si, a est un minorant de A et s'il existe une suite (x_n) de A qui converge vers a.

Démonstration. Il suffit de traiter le cas de la borne supérieure. Supposons $a = \sup A$. d'après le point méthode de la page 518, pour tout $n \in \mathbb{N}^*$, il existe $x_n \in A$, tel que $a - \frac{1}{n} < x_n \leqslant a$. On a $|x_n - a| = a - x_n < \frac{1}{n}$ et la suite $(1/n)$ tend vers 0, donc, d'après la proposition 28 ci-dessus, la suite (x_n) de A converge vers a.

Inversement, supposons que a est un majorant de A et qu'il existe une suite (x_n) de A qui converge vers a. Pour tout $\varepsilon > 0$, il existe $n \in \mathbb{N}$ tel que $a - \varepsilon < x_n < a + \varepsilon$. Mais, puisque $x_n \leqslant a$, on a $a - \varepsilon < x_n \leqslant a$. On en déduit que a est la borne supérieure de A en utilisant la proposition 3 de la page 515. ■

Lemme 30. Soit (u_n) une suite de $\mathbb{K}$ convergeant vers $\ell \in \mathbb{K}$.

- Si $\mathbb{K} := \mathbb{R}$ et $\ell > 0$, alors il existe un réel $\alpha > 0$ tel que $u_n \geqslant \alpha$ à partir d'un certain rang ; en particulier, on a $u_n > 0$ à partir d'un certain rang.
- Si $\ell \neq 0$, alors il existe un réel $\alpha > 0$ tel que $|u_n| \geqslant \alpha$ à partir d'un certain rang ; en particulier, on a $u_n \neq 0$ à partir d'un certain rang.

Démonstration. Le deuxième cas se déduit du premier et de l'exemple précédent en considérant $|u_n|$. Supposons donc $\ell > 0$ et posons $\alpha := \ell/2$. C'est un réel strictement positif donc, par définition, il existe $N \in \mathbb{N}$ tel que $|u_n - \ell| \leqslant \alpha$ pour tout $n \geqslant N$. Soit $n \geqslant N$. on a :

$$u_n = \ell + (u_n - \ell) \geqslant \ell - |u_n - \ell| \geqslant \ell - \alpha = \alpha$$

ce qui prouve le résultat. ∎

Proposition 31. Si une suite (u_n) converge vers ℓ, alors toutes ses sous-suites convergent également vers ℓ.

Démonstration. Soit donc (n_k) une suite strictement croissante d'entiers. Par une récurrence immédiate, on a $\forall k \in \mathbb{N},\ n_k \geqslant k$. Supposons que (u_n) converge vers ℓ. Soit $\varepsilon > 0$; il existe un rang N à partir duquel on a $|u_n - \ell| \leqslant \varepsilon$. Pour tout $k \geqslant N$, on a alors $n_k \geqslant k \geqslant N$ et donc $|u_{n_k} - \ell| \leqslant \varepsilon$. ∎

Cette proposition est très utile, par contraposée, pour montrer qu'une suite diverge.

Exemple. La suite $(u_n) = \big((-1)^n\big)$ diverge puisque sa sous-suite (u_{2n}) est constante égale à 1 et converge donc évidemment vers 1, alors que la sous-suite (u_{2n+1}) est constante et converge vers -1.
Comparer cette démonstration à celle de l'exercice 18 de la page 542 !

Comme on le voit dans l'exercice suivant, l'utilisation des sous-suites (u_{2n}) et (u_{2n+1}) est utile et donne une sorte de réciproque à la proposition 31.

Exercice 22.
Soit (u_n) une suite numérique dont les deux sous-suites (u_{2n}) et (u_{2n+1}) convergent vers la même limite ℓ. Alors (u_n) converge également vers ℓ.
Solution. Soit $\varepsilon > 0$. Il existe deux entiers N_1 et N_2 tels que $|u_{2n} - \ell| \leqslant \varepsilon$ pour tout $n \geqslant N_1$ et $|u_{2n+1} - \ell| \leqslant \varepsilon$ pour tout $n \geqslant N_2$.
Soit $n \in \mathbb{N}$ tel que $n \geqslant N := 2N_1 + 2N_2 + 1$.

- Si $n = 2p$ est pair, alors $2p \geqslant 2N_1$ et donc $p \geqslant N_1$, ce qui donne $|u_n - \ell| \leqslant \varepsilon$.
- De même, si $n = 2p + 1$ est impair, on a $p \geqslant N_2$ et donc $|u_n - \ell| \leqslant \varepsilon$.

On a donc $\forall n \geqslant N,\ |u_n - \ell| \leqslant \varepsilon$, ce qui termine la preuve.

2.2.4 Opérations sur les limites

Dans la pratique, plutôt que d'utiliser la définition, il est plus facile de déterminer la limite d'une suite en se ramenant à des suites connues et en effectuant des additions, multiplications, etc.

Proposition 32. Soient (u_n) une suite *bornée* et (v_n) une suite de limite nulle. Alors la suite $(u_n v_n)$ converge vers zéro.

Remarquer que l'on ne suppose pas (u_n) convergente. Par exemple, il résulte immédiatement de cette proposition que la suite de terme général $\frac{\sin(n!+e^n)}{n}$ converge vers zéro.

Démonstration. On suppose qu'il existe $\mu \in \mathbb{R}_+$ tel que, pour tout $n \in \mathbb{N}$, $|u_n| \leqslant \mu$. On a $|u_n v_n| \leqslant \mu |v_n|$. La suite $\mu|v_n|$ converge vers 0, donc, d'après la proposition 28, la suite $(u_n v_n)$ converge vers 0. ■

Théorème 33. Soient (u_n) et (v_n) deux suites numériques convergentes. Alors (u_n+v_n), (u_n-v_n) et $(u_n v_n)$ sont des suites convergentes ; c'est aussi le cas de $(1/v_n)$ et (u_n/v_n) si $\lim v_n \neq 0$. On a de plus :

$$\lim(u_n+v_n) = \lim u_n + \lim v_n, \qquad \lim(u_n v_n) = \lim u_n \lim v_n,$$

$$\text{si } \lim v_n \neq 0, \text{ alors } \lim 1/v_n = 1/\lim v_n \qquad \lim \frac{u_n}{v_n} = \frac{\lim u_n}{\lim v_n}.$$

Démonstration.

Par hypothèse, les suites (u_n) et (v_n) sont convergentes, appelons a et b leurs limites.

Commençons par le cas de la somme. On a :

$$\forall \varepsilon > 0, \exists N_1 \in \mathbb{N}, \forall n \geqslant N_1, \quad |u_n - a| \leqslant \varepsilon/2,$$
$$\forall \varepsilon > 0, \exists N_2 \in \mathbb{N}, \forall n \geqslant N_2, \quad |v_n - b| \leqslant \varepsilon/2.$$

Posons $N := \max(N_1, N_2)$; $\forall n \geqslant N$, on a :

$$|u_n + v_n - a - b| \leqslant |u_n - a| + |v_n - b| \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Ainsi $\lim(u_n + v_n) = a + b$.

Passons au cas du produit. Rappelons d'abord qu'une suite convergente est bornée (théorème 27 de la page 544). On a :

$$u_n v_n - ab = u_n(v_n - b) + b(u_n - a)$$

La suite u_n est bornée et la suite $(v_n - b)$ tend vers 0, donc, d'après la proposition 32, la suite $(u_n(v_n - b))$ converge vers 0. De même, la suite $(b(u_n - a))$ converge vers 0. D'après le résultat sur la somme, la suite $(u_n(v_n - b) + b(u_n - a))$ converge vers 0 et la suite $u_n v_n$ converge vers ab.

Pour démontrer le résultat pour le quotient, il suffit, en utilisant le résultat pour le produit, de prouver que $(1/v_n)$ tend vers $1/b$ ($b \neq 0$).

Rappelons que si (v_n) a une limite non nulle, alors v_n est non nul au delà d'un certain rang (*cf.* le lemme 30 de la page 545). Pour majorer $\left|\frac{1}{v_n} - \frac{1}{b}\right| = \left|\frac{b-v_n}{v_n\, b}\right|$, on va majorer $|v_n - b|$ et minorer $|v_n|$. D'après le résultat du lemme 30 : il existe un réel $\alpha > 0$ et un rang n_0 à partir duquel on a $|v_n| \geqslant \alpha$. Pour $n \geqslant n_0$, on a donc $\left|\frac{1}{v_n} - \frac{1}{b}\right| \leqslant \frac{|v_n - b|}{b\,\alpha}$, et comme le majorant obtenu tend vers 0, on a bien montré que $(1/v_n)$ tend vers $1/b$. ■

Soient (u_n) une suite numérique convergente et $a \in \mathbb{K}$ donné. Alors :

$$\lim(a + u_n) = a + \lim u_n, \quad \lim a\, u_n = a \lim u_n, \quad \text{et} \quad \text{si } a \neq 0, \quad \lim \frac{u_n}{a} = \frac{\lim u_n}{a}.$$

De plus, si (u_n) a une limite non nulle, alors $\lim a/u_n = a/\lim u_n$.

Corollaire 34. L'ensemble $\mathcal{C}_{\mathbb{K}}$ des suites convergentes à valeurs dans $\mathbb{K}$ est un sous–espace vectoriel et un sous-anneau de $\mathbb{K}^{\mathbb{N}}$.
L'application $(u_n) \mapsto \lim u_n$ est linéaire ; elle est aussi un homomorphisme d'anneaux de $\mathcal{C}_{\mathbb{K}}$ dans $\mathbb{K}$.

Dans le théorème 33 de la page précédente, on a supposé que les deux suites convergent. Si l'une au moins diverge, voire les deux, la situation peut être bien plus complexe.
Certains cas sont cependant assez faciles à traiter. Par exemple :

Exercice 23.
Soient (u_n) une suite divergente et (v_n) une suite convergente.
Montrer que $(u_n + v_n)$ et $(u_n - v_n)$ sont divergentes.
Solution. Soit $w_n := u_n + v_n$, d'où $u_n = w_n - v_n$; si (w_n) était convergente comme (v_n), leur différence (u_n) le serait aussi ce qui n'est pas.
Pour la différence, il suffit de changer v_n en $-v_n$.
Les suites (n) et $(-n)$ sont divergentes ; leur somme est nulle, donc convergente.
Les suites (n) et (n) sont divergentes ; leur différence est nulle, donc convergente.

Il y a des exemples de suites divergentes dont la somme (ou la différence) est convergente : les suites (n) et $(-n)$ sont divergentes ; leur somme est nulle, donc convergente ; les suites (n) et (n) sont divergentes ; leur différence est nulle, donc convergente.

Définition 22. On dit qu'une partie A de $\mathbb{K}$ est *dense* dans une partie $X \subset \mathbb{K}$ si $A \subset X$ et si tout élément de X est limite d'une suite d'éléments de A.

Lorsque $X = \mathbb{K} = \mathbb{R}$, on a déjà défini la notion de densité (*cf.* la définition 7 de la page 519). L'exercice suivant montre qu'il s'agit en fait de la même notion.

Exercice 24.

1. Soit A une partie de $\mathbb{R}$. Montrer que les deux propriétés suivantes sont équivalentes :
   (i) $\forall (x,y) \in \mathbb{R}^2\,,\ x < y \Rightarrow \exists a \in A\,,\ x < a < y\,;$
   (ii) $\forall x \in \mathbb{R}\,,\ \exists (a_n) \in A^{\mathbb{N}}\,,\ x = \lim a_n\,.$
2. Montrer que tout réel est limite d'une suite de rationnels.

Solution.

1. • Supposons (i). Soit $x \in \mathbb{R}$. Pour tout $\varepsilon > 0$, il existe par hypothèse un élément a de A vérifiant $x - \varepsilon < a < x + \varepsilon$. En prenant $\varepsilon = 1/n$, on construit ainsi une suite (a_n) d'éléments de A telle que $\forall n \in \mathbb{N}^*\,,\ |x - a_n| \leqslant 1/n$ et cette suite converge donc vers x.
   • Supposons (ii). Soit $(x,y) \in \mathbb{R}^2$ tel que $x < y$. Posons $z := (x+y)/2$ et considérons une suite (a_n) d'éléments de A convergeant vers z. Pour $\varepsilon := z - x = (y-x)/2$, il existe un rang N à partir duquel on a $|a_n - z| < \varepsilon$, soit $x < a_n < y$. L'élément a_N de A convient donc.
2. On a vu que $A := \mathbb{Q}$ vérifiait (i), il vérifie donc (ii).

Exercice 25.

Montrer que si A est dense dans $\mathbb{R}$, alors $B := \left\{a + ib \;\middle|\; (a,b) \in A^2\right\}$ est dense dans $\mathbb{C}$.

Solution. Soit $z = x + iy$ un nombre complexe quelconque. Il existe deux suites (a_n) et (b_n) d'éléments de A telles que $\lim a_n = x$ et $\lim b_n = y$. Alors la suite $u_n := a_n + ib_n$ converge, par somme et produit par i, vers $x + iy = z$. Donc B est dense dans $\mathbb{C}$.

Exercice 26.

Soit $A \subset \mathbb{R}$ une partie *dénombrable*. Montrer que $\mathbb{R} \setminus A$ est *dense* dans $\mathbb{R}$.

Solution. On raisonne par l'absurde. On suppose que $\mathbb{R} \setminus A$ n'est pas dense. Alors, il existe $x, y \in \mathbb{R}$, $x < y$, tels que $(\mathbb{R} \setminus A) \cap]x,y[= \varnothing$ et l'on a $]x,y[\subset A$. L'ensemble $]x,y[$ est non-dénombrable (*cf.* l'exercice 7 de la page 521), donc, *a fortiori*, A est aussi non dénombrable et l'on a une contradiction.

Le lecteur trouvera à l'exercice IV.1.11 une description des sous-groupes additifs de $\mathbb{R}$: ils sont soit denses dans $\mathbb{R}$ (comme par exemple $\mathbb{Q}$), soit de la forme $a\,\mathbb{Z}$, avec $a \geqslant 0$.

Proposition 35. Soient $V \subset \mathbb{K}$ et $f : V \to \mathbb{K}$ une application continue.

(i) Si (u_n) est une suite *convergente* de V, alors $f(u_n)$ est une suite *convergente* de $\mathbb{K}$. Si $\lim u_n = \ell$, alors $f(\ell) = \lim f(u_n)$.

(ii) Si $A \subset B \subset V$ et si A est dense dans B, alors $f(A)$ est dense dans $f(B)$.

Démonstration. L'assertion (i) est évidente. Montrons (ii). Soit $c \in f(B)$. Il existe $b \in B$ tel que $c = f(b)$. Puisque A est dense dans B, il existe une suite (u_n) de A qui converge vers b. La suite $f(u_n)$ est une suite de $f(A)$ et, d'après (i), elle converge vers $f(b) = c$. Par suite $f(A)$ est dense dans $f(B)$. ∎

Caractérisation des suites complexes convergentes

Proposition 36. Soient (u_n) une suite complexe et $\ell \in \mathbb{C}$. La suite (u_n) converge vers ℓ si, et seulement si, les deux suites réelles $(\operatorname{Re} u_n)$ et $(\operatorname{Im} u_n)$ convergent respectivement vers $\operatorname{Re} \ell$ et $\operatorname{Im} \ell$.

Démonstration.

- Si $\lim u_n = \ell$, les inégalités $|\operatorname{Re} u_n - \operatorname{Re} \ell| \leqslant |u_n - \ell|$ et $|\operatorname{Im} u_n - \operatorname{Im} \ell| \leqslant |u_n - \ell|$ montrent que $\lim \operatorname{Re} u_n = \operatorname{Re} \ell$ et $\lim \operatorname{Im} u_n = \operatorname{Im} \ell$.
- La réciproque provient des opérations sur les limites, puisque $u_n = \operatorname{Re} u_n + i \operatorname{Im} u_n$. ■

Corollaire 37. Si (u_n) converge vers $\ell \in \mathbb{C}$, alors $\lim |u_n| = |\ell|$ et $\lim \overline{u}_n = \overline{\ell}$.

2.2.5 Limites de suites « classiques »

Suites arithmétiques

Comme on l'a vu dans l'exercice 13 de la page 532, une suite arithmétique de raison non nulle est non bornée. Elle ne peut donc pas converger. Les seules suites arithmétiques convergentes sont donc les suites constantes.
Plus précisément, soit (u_n) une suite arithmétique de raison $r \neq 0$.
On a $|u_n| = |u_0 + nr| \geqslant n|r| - |u_0|$, donc, la suite $1/u_n$ est définie à partir d'un certain rang et elle tend vers 0. Par suite la suite (u_n) tend vers l'infini.

Suites géométriques

Proposition 38. Une suite géométrique (u_n) de raison $r \in \mathbb{C}$:

- est constante donc converge vers u_0 si $r = 1$,
- converge vers 0 si $|r| < 1$ ou $u_0 = 0$,
- diverge dans tous les autres cas ; si $|r| > 1$, elle tend vers l'infini.

Autrement dit, une suite géométrique *non nulle* de raison r converge si, et seulement si :

$$|r| < 1 \quad \text{ou} \quad r = 1.$$

Démonstration. Soit (u_n) une suite géométrique. D'après la proposition 21 de la page 533, on a $\forall n \in \mathbb{N},\ u_n = u_0 r^n$. Les cas $r = 1$ ou $u_0 = 0$ étant évidents, nous supposerons $r \neq 1$ et $u_0 \neq 0$.

- La proposition 22 de la page 533, montre qu'une suite géométrique ne peut converger que si sa raison r vérifie $|r| \leqslant 1$. Donc si $|r| > 1$, la suite diverge. Plus précisément, elle tend vers l'infini.
- Supposons $|r| < 1$. Il suffit donc de montrer que (r^n) tend vers 0, le résultat s'en déduit facilement.
  Le résultat est évident lorsque $r = 0$, on peut donc supposer $r \neq 0$.
  Posons $t := 1/|r|$: on a $t > 1$ puisque $|r| < 1$.

Soit $\varepsilon > 0$. D'après la proposition 14 de la page 520 (propriété d'Archimède multiplicative), il existe un entier N tel que $t^N \geqslant 1/\varepsilon$. Ainsi, pour tout $n \geqslant N$, $t^n \geqslant 1/\varepsilon$ et $r^n \leqslant \varepsilon$, d'où le résultat.

- Reste le cas où $|r| = 1$ (et toujours $r \neq 1$). La suite garde alors un module constant $|u_0| \neq 0$, donc ne peut pas converger vers 0. Supposons qu'elle converge ; sa limite est donc un nombre complexe $\ell \neq 0$. Par passage à la limite dans l'égalité $u_{n+1} = r\,u_n$, on obtient $\ell = r\,\ell$, ce qui donne $r = 1$, puisque $\ell \neq 0$, et entraîne une contradiction (1 est l'unique point fixe de $f : x \mapsto rx$). Donc la suite diverge. ■

Exercice 27.

En 1202, dans son *Liber abaci*, le mathématicien italien Leonardo Bonacci, dit Leonardo da Pisa ou Fibonacci (1125–1250), a introduit la célèbre suite qui porte son nom. Elle est définie par la récurrence linéaire d'ordre 2 dite « de Fibonacci » : $F_{n+2} = F_n + F_{n+1}$ et les conditions initiales $F_0 = 0$ et $F_1 = 1$.

1. Montrer que la suite de Fibonacci (F_n) tend vers $+\infty$.
2. Montrer qu'il existe des suites géométriques (w_n) vérifiant la récurrence linéaire de Fibonacci $w_{n+2} = w_{n+1} + w_n$.
3. Montrer que la suite (u_n) définie par la récurrence linéaire de Fibonacci et les conditions initiales $u_0 := 1$ et $u_1 := \frac{1+\sqrt{5}}{2}$ est une suite géométrique tendant vers $+\infty$.
4. Montrer que la suite (v_n) définie par la récurrence linéaire de Fibonacci et les conditions initiales $v_0 := 1$ et $v_1 := \frac{1-\sqrt{5}}{2}$ est une suite géométrique tendant vers 0. Essayer de vérifier *numériquement* le résultat avec une calculette programmable ou un ordinateur et constater que « cela ne marche pas très bien ». Quelle peut être la raison de ce phénomène ?
5. Donner une formule pour F_n et en déduire une formule pour F_{n+1}/F_n ($n \in \mathbb{N}^*$). Montrer que la suite $(F_{n+1}/F_n)_{n\in\mathbb{N}^*}$ converge vers le *nombre d'or*[8] : $\frac{1+\sqrt{5}}{2}$.

Solution.

1. On a $F_2 = 1$, $F_3 = 2$, $F_4 = 3$, $F_5 = 5$, $F_6 = 8$. On prouve par récurrence que, pour tout $n \geqslant 1$, $F_n \geqslant 1$, puis que, pour tout $n \geqslant 5$, $F_n \geqslant n$, d'où le résultat.
2. Cherchons une suite géométrique vérifiant la récurrence linéaire de Fibonacci (on pourra comparer à l'exercice II.1.16 de la page 108.) Par linéarité, il suffit de considérer les suites de la forme $(w_n := \rho^n)$, où $\rho \in \mathbb{R}$. On doit avoir $w_{n+2} = w_{n+1} + w_n$, c'est-à-dire $\rho^{n+2} - \rho^{n+1} - \rho^n = 0$, autrement dit $\rho^n(\rho^2 - \rho - 1) = 0$. En excluant la valeur triviale $\rho := 0$, on obtient l'équation du second degré $\rho^2 - \rho - 1 = 0$, dont les solutions sont $\rho_1 := \frac{1+\sqrt{5}}{2}$ (le nombre d'or) et $\rho_2 := \frac{1-\sqrt{5}}{2}$.

[8] Selon les conceptions de l'antiquité grecque, le nombre d'or est le rapport entre deux longueurs a et b telles que $\frac{a+b}{a} = \frac{a}{b}$. Il est associé à ce qu'Euclide appelle, dans ses *Éléments*, *découpage d'un segment en extrême et moyenne raison*. Il a reçu son nom à la Renaissance. On l'appelle aussi à la même époque *divine proportion* ou *section dorée*. On le note souvent Φ en hommage au sculpteur grec Phidias (490–430 avant J.C.) qui l'aurait utilisé dans certaines de ses statues et pour la décoration du Parthénon de l'acropole d'Athènes.

3. Comme nous venons de le voir, la suite géométrique $(t_n := (\rho_1^n))_{n\in\mathbb{N}}$ vérifie la récurrence de Fibonacci. On a $t_0 = 1 = u_0$ et $t_1 = \frac{1+\sqrt{5}}{2} = u_1$. On en déduit par récurrence $t_n = u_n$ pour tout $n \in \mathbb{N}$. Ainsi (u_n) est une suite géométrique de raison ρ_1. On a $\sqrt{5} > 2$, donc $|\rho_1| > 3/2 > 1$ et la suite (u_n) tend vers $+\infty$.
4. On procède comme dans 1 en remplaçant ρ_1 par $\rho_2 := \frac{1-\sqrt{5}}{2}$.
   On prouve ainsi que v_n est une suite géométrique de raison $\rho_2 := \frac{1-\sqrt{5}}{2}$.
   On a $\sqrt{5} < 5/2$, donc $|\rho_2| < 3/4 < 1$ et, par conséquent, la suite (u_n) tend vers 0.
   Si l'on programme numériquement (c'est-à-dire *en flottants*) le calcul de la suite récurrente, on observe qu'au début la suite a bien l'air de tendre vers 0 mais si l'on pousse plus loin les calculs, elle a l'air de diverger fortement. Ceci est évidemment dû à des problèmes d'arrondi. On peut voir que le point essentiel est l'utilisation, pour le calcul numérique, d'une *valeur arrondie* de $\frac{1-\sqrt{5}}{2}$ au lieu de la *valeur exacte* dès le premier pas. Une explication du phénomène fait l'objet de l'exercice IV.1.44 de la page 607.
5. Pour tous $\lambda_1, \lambda_2 \in \mathbb{R}$, la suite $(\lambda_1\rho_1^n + \lambda_2\rho_2^n)$ vérifie la récurrence de Fibonacci (qui est linéaire). Le système linéaire : $\lambda_1 + \lambda_2 = 0$, $\lambda_1\rho_1 + \lambda_2\rho_2 = 1$, admet l'unique solution : $\lambda_1 = 1/\sqrt{5}$, $\lambda_2 = -1/\sqrt{5}$. On en déduit, que, pour tout $n \in \mathbb{N}$, $F_n = \frac{1}{\sqrt{5}}(\rho_1^n - \rho_2^n)$ (*cf.* l'exercice II.1.16 de la page 108 pour une autre preuve de cette formule). Par suite, pour tout $n \in \mathbb{N}^*$:
   $$F_{n+1}/F_n = \frac{\rho_1^{n+1} - \rho_2^{n+1}}{\rho_1^n - \rho_2^n} = \rho_1 \frac{1 - (\rho_2/\rho_1)^{n+1}}{1 - (\rho_2/\rho_1)^n}.$$
   On a $|\frac{\rho_2}{\rho_1}| < \frac{2}{3} < 1$, donc la suite géométrique $\left((\frac{\rho_2}{\rho_1})^n\right)$ tend vers 0 et $\lim \frac{F_{n+1}}{F_n} = \rho_1$.

Suites arithmético-géométriques

Soit (u_n) une suite vérifiant une relation de récurrence de la forme $u_{n+1} = a\,u_n + b$, avec $(a, b) \in \mathbb{K}^2$, $a \neq 1$, $b \neq 0$.
Si $|a| < 1$, et donc en particulier $a \neq 1$, on a vu qu'il existait un unique élément $\ell \in \mathbb{K}$ tel que $\ell = a\,\ell + b$ (l'unique point fixe de $f : x \mapsto ax + b$)) et que la suite (v_n) définie par $v_n = u_n - \ell$ était géométrique de raison a. Elle converge donc vers 0, ce qui montre que (u_n) converge vers ℓ.
Si $|a| \geqslant 1$ et $a \neq 1$, on montre en se ramenant à une suite géométrique que la suite *diverge*.
Si $a = 1$, la suite est arithmétique. Si $b \neq 0$, la suite tend vers l'infini, donc elle diverge.
Si $a = 1$ et $b = 0$, la suite est constante, donc elle converge

Suites homographiques

En utilisant le point méthode de la page 539, on peut ramener l'étude de la convergence des suites homographiques (quand la condition initiale permet de les définir) à celles de la convergence de suites arithmétiques ou géométriques. En utilisant les résultats que nous

venons d'obtenir pour celles-ci, nous allons montrer que les suites homographiques « bien définies » convergent toujours et convergent vers l'un des points fixes.

Soit $f\ :\ x \mapsto \frac{a\,x+b}{c\,x+d}$, avec $a, b, c, d \in \mathbb{K}$, $c \neq 0$ et $ad - bc \neq 0$.

1. On suppose que les points fixes α et β de f sont *distincts*. On rappelle (*cf.* (2) de la page 538) que :

$$\frac{f(x)-\alpha}{f(x)-\beta} = k\,\frac{x-\alpha}{x-\beta}, \quad \text{avec} \quad k := \frac{c\beta+d}{c\alpha+d} \in \mathbb{K}^* \setminus \{1\}$$

   - On suppose que $|k| \neq 1$. Dans ce cas, quitte à échanger les noms de α et β, on peut supposer que $|k| < 1$. Nous allons montrer que toutes les suites récurrentes bien définies convergent. Plus précisément, si $u_0 = \beta$, la suite est constante et converge donc vers β et si $u_0 \neq \beta$, et si la suite est bien définie, elle converge vers α.

     Si $u_0 \neq \beta$, on vérifie par récurrence que, pour tout $n \in \mathbb{N}$, $u_n \neq \beta$ (en utilisant $k \neq 0$). La suite $(v_n) = \left(\frac{u_n-\alpha}{u_n-\beta}\right)$ est donc bien définie et elle est *géométrique*, de raison k, donc si $|k| < 1$, $\lim\limits_{n\to+\infty} v_n = 0$. On a (*cf.* le point méthode de la page 539) : $u_n = \frac{\beta v_n - \alpha}{v_n - 1}$, donc, si la suite (u_n) est bien définie, c'est-à-dire si, pour tout $n \in \mathbb{N}$, $v_n \neq 1$, $\lim\limits_{n\to+\infty} u_n = \alpha$.

     On peut décrire précisément l'ensemble $X \subset \mathbb{K}$ tel que, pour tout $u_0 \in X$, la suite soit définie et vérifier que, si $\mathbb{K} = \mathbb{R}$ (resp. $\mathbb{K} = \mathbb{C}$), X contient un intervalle (resp. un disque) ouvert de centre α.

     On a (*cf.* page 538) $u_0 \in X$ si, et seulement si, $v_0 \notin Y := \{k^{-n} \mid n \in \mathbb{N}\}$. On a $1 \notin Y$ et si l'on note Z l'image de Y par l'application $v \mapsto u = h(v) := \frac{\beta v - \alpha}{v-1}$, on obtient $X = \mathbb{K} \setminus Z$. L'ensemble Y est l'image de la suite géométrique (k^{-n}), qui *tend vers l'infini*. Par suite, Z est l'image de la suite $w_n := \left(h(k^{-n}) = \frac{\beta - \alpha k^n}{1-k^n}\right)_{n\in\mathbb{N}^*}$, de premier terme $h(k^{-1}) = -d/c$, qui converge vers β.

     On a, pour tout $n \in \mathbb{N}$, $v_n \neq 0$, donc $w_n \neq \alpha$. Alors, la suite $(w_n - \alpha)$ ne prend jamais la valeur 0 et converge vers $\beta - \alpha \neq 0$. On en déduit que $\rho := \inf\limits_{n\in\mathbb{N}^*} |w_n - \alpha| > 0$.

     Alors, si $\mathbb{K} := \mathbb{R}$, la suite (u_n) est bien définie pour tout $u_0 \in\ ]\alpha - \rho, \alpha + \rho[$, et, si $\mathbb{K} := \mathbb{C}$, la suite est bien définie pour tout $u_0 \in \{z \in \mathbb{C} \mid |z - \alpha| < \rho\}$.

     Si l'homographie est *réelle*, c'est-à-dire si $a, b, c, d \in \mathbb{R}$, les points α et β sont tous deux réels ou complexes conjugués. Dans le second cas, on a $|k| = 1$ et ce cas a été exclu dans l'étude ci-dessus.

   - On suppose $|k| = 1$. On définit la suite (v_n) comme ci-dessus. C'est une suite géométrique de raison k. Puisque $|k| = 1$ et $k \neq 1$, elle est divergente. On en déduit que v_n est divergente.

2. On suppose que f admet un seul point fixe α. Nous allons montrer que la suite récurrente (u_n), quand elle est bien définie, converge toujours vers α.

   Notons $r := \frac{2c}{a+d}$; on a $r \neq 0$. Si $u_0 = \alpha$, la suite et constante et converge donc vers α. Supposons $u_0 \neq \alpha$ et notons $v_0 := \frac{1}{u_0 - \alpha} \neq 0$, puis, pour tout $n \in \mathbb{N}$,

$v_n := v_0 + nr$. Nous avons vu que la suite (u_n) est bien définie si, et seulement si, la suite arithmétique (v_n) ne prend jamais la valeur 0 et que, dans ce cas, $u_n = \frac{1}{v_n} + \alpha$. Puisque $r \neq 0$, la suite arithmétique (v_n) tend vers l'infini et, par suite, la suite (u_n) tend vers α.

On peut décrire précisément l'ensemble $X \subset \mathbb{K}$ tel que, pour tout $u_0 \in X$, la suite soit définie et vérifier que, si $\mathbb{K} = \mathbb{R}$ (resp. $\mathbb{K} = \mathbb{C}$), X contient un intervalle (resp. un disque) ouvert de centre α.

La suite $(u_n)_{n\in\mathbb{N}}$ est bien définie si, et seulement si, $u_0 = \alpha + \frac{1}{v_0}$, avec $v_0 \notin \{-nr\}_{n\in\mathbb{N}^*}$. Notons Z l'image de la suite $(w_n := \alpha - \frac{1}{nr})_{n\in\mathbb{N}^*}$. On a donc $X = \mathbb{K} \setminus Z$. On termine comme dans le cas 1.

Exercice 28.
Soit $f : \mathbb{R}^* \to \mathbb{R}$ la fonction homographique définie par $x \in \mathbb{R}^* \mapsto \frac{x+1}{x}$.

1 Montrer que, pour tout $u_0 \in \mathbb{R}_+^*$, f définit une suite récurrente (u_n). Montrer que (u_n) est convergente. Quelle est sa limite ?

2 Retrouver un résultat de l'exercice 27 en montrant que la suite $(F_{n+1}/F_n)_{n\in\mathbb{N}^*}$ (où les F_n sont les nombres de Fibonacci) tend vers le nombre d'or.

Solution.

1 On a $f(\mathbb{R}_+^*) \subset \mathbb{R}_+^*$, donc si $u_0 > 0$ la suite homographique (u_n) est bien définie. Elle converge vers l'un des deux points fixes $\frac{1\pm\sqrt{5}}{2}$. Elle est positive donc elle converge vers $\frac{1+\sqrt{5}}{2}$.

2 On a, pour tout $n \in \mathbb{N}^*$, $F_{n+2} = F_n + F_{n+1}$, donc $\frac{F_{n+2}}{F_{n+1}} = 1 + \frac{F_n}{F_{n+1}}$. Par suite $f(F_{n+1}/F_n) = F_{n+2}/F_{n+1}$ et le résultat se déduit de 1.

2.2.6 Suites récurrentes et points fixes d'applications

Nous venons de voir que, dans le cas des suites géométriques, arithmético-géométriques et homographiques, la limite d'une suite récurrente définie par f, quand elle existe, est toujours un *point fixe* de f. Le résultat suivant est une généralisation, il est *essentiel* pour l'étude des suites récurrentes. Inversement, il permet, dans certains cas, de montrer *l'existence* d'un point fixe à l'aide d'une suite récurrente.

Proposition 39. Soient $I \subset \mathbb{R}$ un intervalle et $f : I \to I$ une application *continue*. Soit $u_0 \in I$, on note (u_n) la suite récurrente de I définie par u_0 et f. On suppose que cette suite tend vers ℓ quand $n \to +\infty$ et que $\ell \in I$. Alors $f(\ell) = \ell$: ℓ est un *point fixe* de f.

Démonstration. L'image par f de la suite $(u_n)_{n\in\mathbb{N}}$ est la suite $(f(u_n) = u_{n+1})_{n\in\mathbb{N}}$. La limite de cette dernière est $f(\ell)$, mais, d'après la proposition 26 de la page 543, c'est aussi ℓ, donc $f(\ell) = \ell$. ∎

Point Méthode

Si I est un intervalle réel et $f : I \to I$ une application *continue*, les limites possibles des suites récurrentes définies par f et $u_0 \in \mathbb{R}$ sont des *points fixes* de f. Pour étudier la convergence éventuelle de ces suites, il faut donc commencer par déterminer les points fixes de f. On peut s'aider d'une méthode graphique : on cherche les points d'intersection du graphe de f et de la droite $\{y = x\}$.
S'il n'y a pas de point fixe, toutes les suites récurrentes définies par f *divergent*.

Exercice 29.

1. Soient I un segment de $\mathbb{R}$ et $f : I \to I$ une application continue.
   Montrer que f admet un point fixe.
2. Soient I un segment de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une application continue.
   On suppose que $I \subset f(I)$. Montrer que f admet un point fixe.

Solution.

1. Soit $I = [a, b]$. On applique le théorème des valeurs intermédiaires à la fonction continue $g : x \in I \mapsto f(x) - x \in \mathbb{R}$. On a $g(a) \geqslant 0$ et $g(b) \leqslant 0$. Il existe donc $\ell \in I$, tel que $g(\ell) = 0$, c'est-à-dire $f(\ell) = \ell$.
2. On garde les notations définies ci-dessus. Il existe $c, d \in I$ tels que $f(c) = a$ et $f(d) = b$. On a $g(c) = f(c) - c = a - c \leqslant 0$ et $g(d) = f(d) - d = b - c \geqslant 0$, d'où le résultat en appliquant le théorème des valeurs intermédiaires.

Exemple. Soient I un segment et $f : I \to I$ une application continue *décroissante*. Alors f admet un *unique* point fixe.
D'après l'exercice précédent, il existe un point fixe ℓ. On raisonne par l'absurde. Soit ℓ_1 un autre point fixe. On peut supposer $\ell_1 < \ell$; on a $\ell_1 = f(\ell_1) \geqslant f(\ell) = \ell > \ell_1$, d'où une contradiction.

Corollaire 40. Soient $I \subset \mathbb{R}$ un intervalle et $f : I \to I$ une application *continue*. Soit $u_0 \in I$, on note (u_n) la suite récurrente de I définie par u_0 et f. On suppose que les suites $(u_{2p})_{p\in\mathbb{N}}$ et $(u_{2p+1})_{p\in\mathbb{N}}$ convergent respectivement vers ℓ_1 et ℓ_2 et que $\ell_1,\ \ell_2 \in I$. Alors ℓ_1 et ℓ_2 sont des points fixes de $f \circ f$, $f(\ell_1) = \ell_2$ et $f(\ell_2) = \ell_1$. Si $\ell_1 \neq \ell_2$, (ℓ_1, ℓ_2) est un 2-cycle de f. Si $\ell_1 = \ell_2$, ℓ_1 est un point fixe et (u_n) converge vers ℓ_1.

Il existe un résultat similaire quand, pour un entier $\tau > 2$ donné, les τ suites extraites $(u_{p\tau+q})_{p\in\mathbb{N}^*}$, $q \in [\![0, \tau - 1]\!]$ convergent (*cf.* l'exercice IV.1.27 de la page 603).

Démonstration. La suite (u_{2p}) et la suite (u_{2p+1}) sont des suites récurrentes définies par $f \circ f$ et les conditions initiales respectives u_0 et u_1. L'application $f \circ f$ est continue, donc $f \circ f(\ell_1) = \ell_1$ et $f \circ f(\ell_2) = \ell_2$.
L'image par f de la suite $(u_{2p})_{p\in\mathbb{N}}$ est la suite $\left(f(u_n) = u_{2p+1}\right)_{p\in\mathbb{N}}$. La limite de cette dernière est $f(\ell_1)$, mais, d'après la proposition 26, c'est aussi ℓ_2, donc $f(\ell_1) = \ell_2$. On montre de façon analogue que $f(\ell_2) = \ell_1$. Par suite $f \circ f(\ell_1) = \ell_2$ et $f \circ f(\ell_2) = \ell_1$. ■

Exercice 30.
Soient I un segment de $\mathbb{R}$ et $f : I \to I$ une application continue.
On suppose que f admet un 2-cycle. Montrer que f admet un point fixe.
Solution. Par hypothèse, il existe $a,\ b \in I,\ a < b$, tels que $f(a) = b$ et $f(b) = a$. Soit $g : I \to \mathbb{R}$, définie par $g : x \in I \mapsto f(x) - x$. On a $g(a) = b - a > 0$ et $g(b) = a - b < 0$. D'après le théorème des valeurs intermédiaires, il existe $c \in]a, b[$ tel que $g(c) = 0$. Alors $f(c) = c$: c est un point fixe.

2.3 Cas des suites réelles

2.3.1 Limites et inégalités

Dans le cas des suites réelles, on peut utiliser la relation d'ordre sur $\mathbb{R}$. Le résultat suivant montre la compatibilité avec la notion de limite.

Proposition 41.

(i) Si (u_n) est une suite réelle convergente telle que $u_n \geqslant 0$ à partir d'un certain rang, alors $\lim u_n \geqslant 0$.

(i) Si (u_n) et (v_n) sont deux suites réelles convergentes telles que $u_n \leqslant v_n$ à partir d'un certain rang, alors $\lim u_n \leqslant \lim v_n$.

Démonstration. Par différence, la deuxième assertion résulte de la première. Montrons (i) par l'absurde, en supposant que (u_n), à valeurs positives ou nulles à partir d'un rang N_1, converge vers $\ell < 0$. En prenant $\varepsilon := -\ell/2$, qui est bien strictement positif par hypothèse, il existe un rang N_2 à partir duquel on a $|u_n - \ell| \leqslant \varepsilon$, ce qui donne $u_n \leqslant \ell + \varepsilon = \ell/2 < 0$. En particulier, $u_N < 0$, avec $N := N_1 + N_2 \geqslant N_1$, ce qui est contraire à l'hypothèse. ■

Corollaire 42. Soient $a,\ b \in \mathbb{R}$ et (u_n) une suite réelle convergente de limite ℓ. Si $u_n \geqslant a$ (resp. $u_n \leqslant b$) à partir d'un certain rang, $\ell \geqslant a$ (resp. $\ell \leqslant b$). Si $u_n \in [a, b]$ à partir d'un certain rang, $\ell \in [a, b]$.

Une suite à termes strictement positifs n'a pas toujours une limite strictement positive : nous avons déjà vu par exemple que la suite $(1/10^n)$ tend vers zéro. Retenir que *les inégalités strictes ne passent pas à la limite*.

Le théorème suivant est souvent utilisé pour montrer des convergences de suites.

Théorème 43 (dit « des gendarmes »).
Soient (a_n) et (b_n) deux suites réelles convergentes admettant la même limite ℓ. Si (u_n) est une suite telle que $a_n \leqslant u_n \leqslant b_n$, à partir d'un certain rang, alors (u_n) est convergente et admet ℓ pour limite.

Le nom du théorème évoque une allégorie sous-jacente : la suite (u_n) est encadrée par (a_n) et (b_n) comme un voleur entre deux gendarmes. Si les deux gendarmes vont au poste (la limite), le voleur est contraint d'y aller aussi.

On notera que le point intéressant dans ce théorème n'est pas tellement que $\lim u_n = \ell$ (c'est une conséquence immédiate de la proposition 41 de la page précédente), mais le fait que la suite (u_n) converge, qui n'est pas une hypothèse, mais une conclusion.

Démonstration. Par hypothèse, il existe $N_0 \in \mathbb{N}$ tel que pour tout $n \geqslant N_0$, on a $a_n \leqslant u_n \leqslant b_n$. Par ailleurs, la convergence de (a_n) et (b_n) vers ℓ implique que pour tout $\varepsilon > 0$, il existe N_1 et N_2 tels que $n \geqslant N_1$ implique $|a_n - \ell| \leqslant \varepsilon$ et $n \geqslant N_2$ implique $|b_n - \ell| \leqslant \varepsilon$.
Si nous posons $N := \max(N_0, N_1, N_2)$, nous avons pour tout $n \geqslant N$:

$$\ell - \varepsilon \leqslant a_n \leqslant u_n \leqslant b_n \leqslant \ell + \varepsilon,$$

ce qui implique que $|u_n - \ell| \leqslant \varepsilon$. Nous avons bien montré la convergence de u_n vers ℓ. ■

2.3.2 Suites monotones

Définition 23. On dit qu'une suite réelle (u_n) est *croissante*, si $u_n \leqslant u_{n+1}$ pour tout n ; qu'elle est *décroissante* si $u_n \geqslant u_{n+1}$ pour tout n.
Dans les deux cas, la suite est dite *monotone*.

En général, une suite n'est pas monotone (considérer le cas de la suite $\big((-1)^n\big)$ par exemple). Une suite à la fois croissante et décroissante est constante.
On définit de manière similaire les suites *strictement croissantes* ou *strictement décroissantes*, avec des inégalités strictes. Toutes ces propriétés peuvent éventuellement n'être vérifiées qu'à partir d'un certain rang, ce qui n'a pas d'importance pour les propriétés de convergence.

Théorème 44. Une suite réelle croissante converge si, et seulement si, elle est majorée. De même une suite réelle décroissante converge si, et seulement si, elle est minorée.

Démonstration. La preuve du théorème est basée sur le fait que $\mathbb{R}$ possède la propriété de la borne supérieure (*cf.* la définition 5 de la page 514).

Soit une suite réelle (u_n) croissante, l'autre cas se déduisant en considérant $(-u_n)$.

Supposons d'abord (u_n) convergente : alors elle est bornée d'après le théorème 27 de la page 544, donc elle est majorée.

Inversement, supposons la suite majorée, et montrons qu'elle converge. Considérons l'image de la suite $A := \{u_0, u_1, u_2, \dots\}$. Il est non vide (il contient au moins u_0) et il est majoré puisque la suite est majorée, il admet donc une borne supérieure que nous notons ℓ.

Soit $\varepsilon > 0$. D'après la proposition 3 de la page 515, il existe (au moins) un entier $N \in \mathbb{N}$ tel que $\ell - \varepsilon < u_N \leqslant \ell$. La suite est croissante : donc pour tout $n \geqslant N$, on a $u_n \geqslant u_N > \ell - \varepsilon$. Puisque $u_n \in A$, on a aussi $u_n \leqslant \ell$, donc pour tout $n \geqslant N$, $\ell - \varepsilon < u_n \leqslant \ell$ et $|u_n - \ell| \leqslant \varepsilon$. La suite (u_n) converge donc vers ℓ. ■

Notons que ce théorème ne donne pas la « valeur » de la limite (dans la preuve celle-ci apparaît comme borne supérieure de l'ensemble des valeurs de la suite, ce qui n'est pas

« constructif »). Par exemple, on verra dans le module IV.4 que la suite (évidemment croissante) définie par : $\forall n \in \mathbb{N},\ u_n = \sum_{p=1}^{n} \frac{1}{p^3}$ est majorée, mais on ne sait pas exprimer sa limite à l'aide des constantes et des fonctions usuelles. On introduit donc pour cela une « fonction spéciale » (la fonction ζ)[9] telle que $\lim u_n = \zeta(3)$.

Remarque. Le théorème précédent nous donne la position de la limite par rapport aux termes de la suite. On a par exemple :

$$\zeta(3) \geqslant u_5 = 1 + \frac{1}{8} + \frac{1}{27} + \frac{1}{64} + \frac{1}{125} \geqslant 1{,}18.$$

Inversement, si l'on connaît la limite, on peut souvent en déduire des renseignements sur les termes de la suite : par exemple, une suite décroissante convergeant vers 0 est positive.

Exemple. Soit f une fonction *monotone* d'un intervalle I de $\mathbb{R}$ dans lui-même. On considère la suite récurrente (u_n) définie par $u_0 \in I$ et f.

1. On suppose f *croissante*. En utilisant $u_{n+1} - u_n = f(u_n) - f(u_{n-1})$, on vérifie par récurrence que, pour tout $n \in \mathbb{N}$, $u_{n+1} - u_n$ et $u_1 - u_0$ sont de même signe (ou tous les deux nuls), donc la suite (u_n) est monotone. Elle est croissante si $u_1 \geqslant u_0$ et décroissante si $u_1 \leqslant u_0$.

   Si l'intervalle I est borné, la suite récurrente (u_n) est bornée, elle admet donc une limite $\ell \in \mathbb{R}$. Si I est un segment, $\ell \in I$. Si, de plus, f est *continue*, alors ℓ est un point fixe de ℓ.
2. On suppose f *décroissante*. Alors la suite (u_n) n'est pas en général monotone. Par exemple, si $I := \mathbb{R}$, $f : x \mapsto -x$ et $u_0 := 1$, $u_n = (-1)^n$ n'est pas monotone. Par contre les suites extraites respectivement d'indices pairs et impairs : $(u_{2p} = 1)_{p\in\mathbb{N}}$ et $(u_{2p+1} = -1)_{p\in\mathbb{N}}$ sont constantes, donc monotones. Ceci est en fait général. La fonction $f \circ f$ est *croissante*, donc la suite $(v_p := u_{2p})_{p\in\mathbb{N}}$, qui est la suite récurrente définie par $f \circ f$ et $v_0 := u_0$ est monotone et la suite $(w_p := u_{2p+1})_{p\in\mathbb{N}}$, qui est la suite récurrente définie par $f \circ f$ et $w_0 := u_1$ est également monotone (l'une des deux suites croit et l'autre décroit).

   Si I est un *segment* et si f est continue, alors les suites (v_p) et (w_p) convergent respectivement vers des points fixes ℓ_0 et ℓ_1 de $f \circ f$. S'ils sont distincts, (ℓ_0, ℓ_1) est un 2-cycle (les suites récurrentes définies par f et $u_0 = \ell_0$ ou $u_0 = \ell_1$ sont périodiques de plus petite période 2) et la suite (u_n) « oscille » entre des valeurs qui se rapprochent de plus en plus de ℓ_0 et ℓ_1. Si $\ell_0 = \ell_1$, (u_n) converge vers ℓ_0.

Exemple. Soit $f : x \in \mathbb{R} \mapsto -\frac{x^3}{2} + \frac{3}{2}x \in \mathbb{R}$. On a $f'(x) = \frac{3}{2}(1 - x^2)$. La fonction est donc croissante sur le segment $I := [-1, 1]$. On a $f(\pm 1) = \pm 1$, donc I est invariant par f. On note encore $f : I \to I$; f a 3 points fixes 0, ± 1. On peut alors étudier les suites récurrentes définies par f et $u_0 \in I$.

[9] *cf.* la fonction zêta dans le deuxième volume L2 de cette série

1. Si $u_0 := 0$, la suite est constante et égale à 0.
2. Si $0 < u_0 \leqslant 1$, $u_1 - u_0 = f(u_0) - u_0 = \frac{u_0(1-u_0^2)}{2} \geqslant 0$, donc la suite (u_n) est croissante. Elle converge nécessairement vers 1.
3. Si $-1 \leqslant u_0 < 0$, la suite (u_n) est décroissante. Elle converge nécessairement vers -1.

Exercice 31.

Soit $a \in \mathbb{R}$. On définit la suite (u_n) par :

$$u_0 := a \qquad \text{et} \qquad \forall n \in \mathbb{N},\ u_{n+1} = \sin(u_n).$$

1. Montrer que $[-\pi/2, \pi/2]$ est stable par $\sin$ et en déduire que $(u_n)_{n\geqslant 1}$ est monotone.
2. Montrer que (u_n) converge. Quelle est sa limite ?

Solution.

1. L'image de la fonction $\sin$ est contenue dans $[-1, 1]$ et $[-1, 1] \subset [-\pi/2, \pi/2]$, donc $I := [-\pi/2, \pi/2]$ est stable par $\sin$. La fonction $\sin$ est croissante sur I, donc la suite (u_n) est monotone.
2. La suite converge d'après le théorème 44. La continuité de la fonction $\sin$ permet de montrer que la limite ℓ de (u_n) vérifie $\ell = \sin \ell$. Les variations de $x \mapsto \sin x - x$ permettent de conclure que $\ell = 0$.

Exemple : suite de Héron, calcul numérique de $\sqrt{2}$.

Le problème « résolu » par Héron d'Alexandrie est le calcul de valeurs approchées de la longueur du côté d'un carré de surface égale à 2 (que nous appelons $\sqrt{2}$). Sa méthode, vraisemblablement héritée des babyloniens[10] est la suivante. On part d'un *rectangle* de surface 2 et de côtés *rationnels* ℓ et L, $\ell < L$. Celui choisi par Héron est « le plus simple », c'est-à-dire $\ell = 1$, $L = 2$. Ensuite on remplace le rectangle initial par un nouveau rectangle, de même surface, dont le *grand* côté L_1 est encore rationnel, mais est *plus petit* que L. On choisit *la moyenne arithmétique* des deux côtés initiaux $L_1 := \frac{\ell+L}{2}$, c'est-à-dire $L_1 = \frac{1}{2}\left(L + \frac{2}{L}\right)$. On itère ensuite le processus. On obtient :

$$L_1 = \frac{3}{2}, \quad L_2 = \frac{17}{12}, \quad L_3 = \frac{577}{408}, \quad L_4 = \frac{665857}{470832}, \quad L_5 = \frac{886731088897}{627013566048},$$

soit, en écriture décimale :

$$L_1 = 1,5,\ L_2 = 1,\underline{41}6666\ldots,\ L_3 = 1,\underline{4142}156\ldots,$$

$$L_4 = 1,\underline{414213562374}5,\ L_5 = 1,\underline{41421356237309504880168}96\ldots.$$

[10] Un résultat, pour un problème équivalent, en écriture cunéiforme des chiffres (schématisée ici par : $\top$ (1) et $\langle$ (10)) figure sur une tablette babylonienne en argile (YBC7289, Yale University) datée vers 1700 avant notre ère. Il y a, en transposition dans notre notation décimale, 5 *décimales exactes*. Le système utilisé dans la tablette est la numération en base 60, avec des « chiffres en base 10 », et le résultat « est écrit » : $\top\ \langle\langle\top\top\top\top\ \langle\langle\langle\langle\top\ \langle$, que l'on traduit par $1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3}$.

Valeurs que l'on peut comparer à :

$$\sqrt{2} = 1{,}4142135623730950488016887\ldots$$

On constate que la méthode est « exceptionnellement efficace » : le nombre de décimales exactes (soulignées) double « en gros » à chaque étape $(0,\ 2,\ 5,\ 11,\ 23)$.
Il est vraisemblable que les mathématiciens babyloniens étaient allés jusqu'à l'étape 4.
Nous allons analyser la méthode de Héron avec nos outils et expliquer son efficacité.
On note $I :=]0, +\infty[$ et l'on définit $f : I \to I$ par $f : x \mapsto \frac{1}{2}\left(x + \frac{2}{x}\right)$. On a $1 < \sqrt{2} < 2$.
On considère la suite récurrente (x_n) définie par $x_0 := 2$ et f. L'unique solution de $f(x) = x$ dans I est $\sqrt{2}$. Si $\sqrt{2} < x \leqslant 2$, on a $2\,(x - f(x)) = x - \frac{2}{x} = \frac{x^2-2}{x}$ et $2\left(f(x) - \sqrt{2}\right) = \frac{(x-\sqrt{2})^2}{x}$, donc $\sqrt{2} < f(x) < x \leqslant 2$. On en déduit que la suite (x_n) est *décroissante*. Elle est bornée inférieurement par $\sqrt{2}$, donc, d'après le théorème 44, elle converge ; notons ℓ sa limite. D'après la proposition 39 de la page 553, ℓ est solution de $f(x) = x$, donc $\ell = \sqrt{2}$.

On a $x_{n+1} - \sqrt{2} = \dfrac{(x_n - \sqrt{2})^2}{2x_n} \leqslant \dfrac{(x_n - \sqrt{2})^2}{2\sqrt{2}} = 2\sqrt{2}\left(\dfrac{x_n - \sqrt{2}}{2\sqrt{2}}\right)^2$, ce qui donne :

$$\frac{x_{n+1} - \sqrt{2}}{2\sqrt{2}} \leqslant \left(\frac{x_n - \sqrt{2}}{2\sqrt{2}}\right)^2.$$

On en déduit, par récurrence et en utilisant $\sqrt{2} < x_1 = 3/2$:

$$x_n - \sqrt{2} < 2\sqrt{2}\left(\frac{x_0 - \sqrt{2}}{2\sqrt{2}}\right)^{2^n} = 2\sqrt{2}\left(\frac{\sqrt{2}-1}{2}\right)^{2^n} < 3 \times 2^{-2^{n+1}}. \tag{8}$$

Cette inéquation permet, pour tout $k \in \mathbb{N}$, de calculer n tel que $x_n - \sqrt{2} < 10^{-k}$.
Par exemple, pour $k := 6$, on choisit n tel que $\ln 3 + 6 \ln 10 < 2^{n+1} \ln 2$: $n := 4$ convient.
On en déduit : $1{,}4142135 - 10^{-6} = 1{,}4142125 \leqslant \sqrt{2} < 1{,}4142136$ et l'on obtient 5 décimales *sûrement exactes* : $\sqrt{2} = 1{,}41421\ldots$. On a ainsi *prouvé rigoureusement*, pour 5 décimales, l'estimation babylonienne. En fait, $n = 4$ convient encore pour $k = 9$, ce qui permet d'obtenir 8 décimales *sûrement exactes* : $\sqrt{2} = 1{,}41421356\ldots$.
On peut montrer que la méthode permet d'obtenir, pour tout $k \in \mathbb{N}$, k décimales *sûrement exactes*, mais il faut, pour cela, surmonter une difficulté technique : on a pu enlever 1 à la 6ème décimale $\alpha_6 = 3$ de x_4, mais nous n'aurions pas pu le faire si l'on avait eu $\alpha_6 = 0$.
Pour $k = 101$, $n = 8$ convient, ce qui permet *d'espérer* obtenir 100 décimales. Pour en être sûr il faut voir si la 101ème décimale de x_8 n'est pas nulle. En utilisant un système de calcul formel on trouve que cette décimale est 7. On peut de plus vérifier que x_8 donne en fait 196 décimales exactes de $\sqrt{2}$!
On peut remplacer l'aire 2 du carré par $a \in \mathbb{R}_+^*$ et la fonction f par $f : x \mapsto \frac{1}{2}\left(x + \frac{a}{x}\right)$.
En choisissant $u_0 > \sqrt{a}$, on obtient des suites récurrentes décroissantes qui convergent vers $\sqrt{a}$. Si $a \in \mathbb{Q}$, on obtient des approximations rationnelles de $\sqrt{a}$.

La méthode de Héron peut être interprétée comme un cas particulier de la *méthode de Newton*[11] (que nous étudierons dans le deuxième volume de cette série L2.
Pour résoudre l'équation $g(x) := x^2 - 2 = 0$, on construit une suite récurrente (u_n) par le procédé suivant : on choisit $u_0 > \sqrt{2}$ et l'on détermine u_{n+1} à partir de u_n en prenant l'intersection de la tangente au graphe de g (une parabole) au point $(u_n, g(u_n))$ avec l'axe des abscisses. La tangente à pour équation $y = 2u_n x - u_n^2 - 2$ ($g'(u_n) = 2u_n$) et elle coupe $x = 0$ en $u_{n+1} = u_n - \frac{g(u_n)}{g'(u_n)} = \frac{1}{2}\left(u_n + \frac{2}{u_n}\right)$. On retrouve la suite de Héron.

2.3.3 Suites adjacentes

Définition 24. On dit que deux suites réelles (u_n) et (v_n) sont *adjacentes* si elles vérifient les propriétés suivantes :

1. (u_n) est croissante, et (v_n) est décroissante ;
2. la suite $(v_n - u_n)$ tend vers zéro.

Un exemple typique de suites adjacentes est celui que l'on obtient en considérant respectivemenr les approximations décimales par défaut et par excès d'un même nombre réel (voir page 729) : par exemple pour π on obtient $u_0 = 3$; $u_1 = 3{,}1$; $u_2 = 3{,}14$; $u_3 = 3{,}141$; ... et $v_0 = 4$; $v_1 = 3{,}2$; $v_2 = 3{,}15$; $v_3 = 3{,}142$; ... Il est clair que deux telles suites vérifient bien les propriétés demandées.

Théorème 45. Deux suites adjacentes convergent, et ont la même limite.

L'intérêt de ce théorème réside dans la propriété de convergence, qui n'est pas évidente car on n'a pas supposé les suites bornées. Le fait qu'elles aient la même limite est alors évident, puisque $(v_n - u_n)$ tend vers zéro par hypothèse.

Démonstration. En vertu du théorème 44 de la page 556, il suffit de montrer que les suites sont bornées. Remarquons que la suite $(v_n - u_n)$ est décroissante par décroissance de (v_n) et croissance de (u_n). Comme elle converge vers 0, on en déduit qu'elle est positive (voir la remarque de la page 557), et donc que $\forall n \in \mathbb{N}\,,\ u_n \leqslant v_n$. On a donc, pour tout $n \in \mathbb{N}$:

$$u_0 \leqslant u_1 \leqslant \cdots \leqslant u_n \leqslant v_n \leqslant \cdots \leqslant v_1 \leqslant v_0$$

ce qui montre que les deux suites sont minorées par u_0 et majorées par v_0. ■

Remarquer qu'il résulte de cette démonstration, que les deux suites adjacentes vérifient :

$$\forall n \in \mathbb{N}\,,\ u_n \leqslant v_n \qquad \text{(et plus généralement } \forall m \in \mathbb{N}\,,\ u_n \leqslant v_m\text{)},$$

ce qui n'est pas requis au départ.

Exercice 32.

Soient $u_n := \sum_{p=0}^{n} \frac{1}{p!}$ et $v_n := u_n + \frac{1}{n!}$.

Montrer que les suites (u_n) et (v_n) convergent vers une même limite. Quelle est cette limite ? Même question en remplaçant v_n par $w_n := u_n + \frac{1}{n\,n!}$.

[11] Isaac Newton, *De analysi per aequationes numero terminorum infinitas*, 1669.

Solution. Montrons que les suites (u_n) et (v_n) sont adjacentes.

- La croissance de (u_n) est immédiate puisque $u_n - u_{n-1} = 1/n! \geqslant 0$.
- On a donc $v_n - v_{n-1} = \frac{1}{n!} + \frac{1}{n!} - \frac{1}{(n-1)!} = \frac{2-n}{n!}$.

  La suite (v_n) est donc décroissante à partir du rang 2 ($v_n \leqslant v_{n-1}$ pour tout $n \geqslant 2$).
- On a : $v_n - u_n = \frac{1}{n!} \leqslant \frac{1}{n} \underset{n\to+\infty}{\longrightarrow} 0$.

Le résultat découle donc du théorème précédent. La limite de la suite u_n est $\mathrm{e} = \exp(1)$, la base des logarithmes népériens (*cf.* le théorème et définition 21 de la page 272, le théorème 22 de la page 272 et la proposition et définition 4 de la page 680).
On a $w_{n+1} - w_n = -\frac{1}{n(n+1)(n+1)!}\cdot$; la suite (w_n) est donc décroissante.
Enfin, $w_n - u_n = \frac{1}{n\,n!} \leqslant \frac{1}{n} \underset{n\to+\infty}{\longrightarrow} 0$.

Exercice 33.

En utilisant le résultat de l'exercice précédent, montrer que e est *irrationnel*.
Solution. On raisonne par l'absurde. On suppose que $\mathrm{e} = p/q$, avec $p \in \mathbb{N},\ q \in \mathbb{N}^*$.
On suppose $n \geqslant 2$ et $n > q$. La suite (u_n) est *strictement* croissante.
Alors $u_n < \mathrm{e} < w_n = u_n + \frac{1}{n\,n!}$, donc :

$$u_n n! < p\frac{n!}{q} < w_n n! = u_n n! + \frac{1}{n} \leqslant u_n n! + \frac{1}{2};$$

$u_n n!$ et $p\frac{n!}{q}$ sont entiers, donc $u_n n! + 1 \leqslant p\frac{n!}{q}$, ce qui contredit $p\frac{n!}{q} \leqslant u_n n! + \frac{1}{2}\cdot$

Il est plus difficile de prouver l'irrationalité de π. En 1766 le mathématicien suisse Johann Heinrich Lambert a montré[12] que la tangente d'un nombre rationnel est irrationnel. Il en a déduit l'irrationnalité de π en utilisant $\tan \pi/4 = 1$. Il existe des preuves plus récentes et « plus simples » utilisant des intégrales. On peut aussi montrer, de façon analogue, que si $a \in \mathbb{Q}^*$, alors e^a est irrationnel (*cf.* l'exercice IV.1.60).

Proposition 46 (Théorème des segments emboîtés). Soit $\big([a_n, b_n]\big)$ une suite décroissante de segments non vides dont les longueurs tendent vers 0. Alors l'intersection $\bigcap\limits_{n\in\mathbb{N}} [a_n, b_n]$ est réduite à un point.

Ce résultat est une version plus forte de la proposition 5 de la page 516 qui, sans l'hypothèse sur les longueurs, nous disait que l'intersection des segments $[a_n, b_n]$ était non vide.

Démonstration. Des hypothèses, on déduit facilement que les suites (a_n) et (b_n) sont adjacentes : (a_n) est croissante, (b_n) décroissante et $b_n - a_n$, longueur du segment $[a_n, b_n]$ tend vers 0. Par le théorème 45 de la page précédente, elles convergent donc vers un réel ℓ qui vérifie $\forall n \in \mathbb{N},\ a_n \leqslant \ell \leqslant b_n$ donc qui appartient à l'intersection $I := \bigcap\limits_{n\in\mathbb{N}} [a_n, b_n]$.
Soit maintenant $x \in I$. Par passage à la limite dans les inégalités $a_n \leqslant x \leqslant b_n$, on en déduit $\ell \leqslant x \leqslant \ell$, soit $x = \ell$. Donc $I = \{\ell\}$ est bien réduit à un point. ∎

---

[12] En utilisant le développement en *fractions continues* de cotan.
Les fractions continues sont présentées dans le livre L3.

2.4 Suites bornées

2.4.1 Théorème de Bolzano-Weierstraß

Rappelons qu'une suite numérique (x_n) est bornée s'il existe un réel M tel que $|x_n| \leqslant M$ pour tout n. Naturellement une telle suite ne converge pas en général : penser à la suite $\big((-1)^n\big)$. Pourtant, on voit sur ce contre-exemple qu'elle admet une sous-suite convergente : il suffit de prendre les termes d'indice pair (ou les termes d'indice impair).
Plus généralement, étant donnée une suite (x_n) bornée, on peut sélectionner suffisamment habilement des termes $x_{n_0}, x_{n_1}, \dots$, de sorte que la suite extraite soit convergente.

Théorème 47 (Bolzano-Weierstraß). Toute suite numérique bornée $(x_n)_{n\in\mathbb{N}}$ contient une suite extraite $(x_{n_k})_{k\in\mathbb{N}}$ convergente.

Le résultat ci-dessus a été découvert par le mathématicien autrichien Bernard Bolzano en 1830 et prouvé rigoureusement par Karl Weierstrass trente ans plus tard.

Démonstration dans le cas d'une suite réelle bornée.

Nous allons procéder par dichotomie. Nous supposons que $(x_n)_{n\in\mathbb{N}}$ prend ses valeurs dans un segment $[a, b]$. Nous allons d'abord définir par récurrence une suite de segments emboîtés (I_n). On pose $I_0 := [a, b]$. On considère ensuite les deux segments $[a, \frac{a+b}{2}]$ et $[\frac{a+b}{2}, b]$. Leur réunion est I_0, donc l'un (au moins) des deux contient une infinité de termes de la suite. On le note I_1 (on choisit arbitrairement si nécessaire). On recommence ensuite avec I_1 ce que nous venons de faire avec I_0 en utilisant le milieu de I_1 pour découper I_1 en deux segments. L'un au moins contient une infinité de termes de la suite, on le note I_2. On poursuit par récurrence. Les I_n sont des segments emboîtés, chaque I_n contient une infinité de termes de termes de la suite et la longueur de I_n est $\frac{b-a}{2^n}$ qui tend vers 0 quand n tend vers $+\infty$.
Notons $[a_n, b_n] := I_n$. Les deux suites (a_n) et (b_n) sont évidemment adjacentes donc convergentes. On note ℓ leur limite commune.
Nous allons construire une application strictement croissante $\varphi : \mathbb{N} \mapsto \mathbb{N}$ telle que, pour tout $n \in \mathbb{N}$, $x_{\varphi(n)} \in I_n$. On pose $\varphi(0) := 0$. On définit ensuite $\varphi(1)$ comme le plus petit entier $m \in \mathbb{N}^*$ tel que $x_m \in I_1$, puis $\varphi(2)$ comme le plus petit entier m tel que $m > \varphi(1)$ et $x_m \in I_2$. On poursuit par récurrence. On obtient ainsi une suite extraite $(v_n := x_{\varphi(n)})$.
On a, pour tout $n \in \mathbb{N}$, $a_n \leqslant v_n \leqslant b_n$, donc, d'après le théorème des gendarmes, (v_n) converge vers ℓ. ∎

Démonstration dans le cas d'une suite complexe bornée. Soit (u_n) une suite complexe bornée. Posons, pour tout $n \in \mathbb{N}$, $x_n := \operatorname{Re} u_n$ et $y_n := \operatorname{Im} u_n$. Ce sont des suites bornées puisque $|x_n| \leqslant |u_n|$ et $|y_n| \leqslant |u_n|$ pour tout $n \in \mathbb{N}$.

- Comme (x_n) est une suite réelle bornée, elle admet une sous-suite convergente : notons-la $(x_{n_p})_{p\in\mathbb{N}}$ et notons α sa limite.
- La suite $(z_p) = (y_{n_p})$ est alors une sous-suite de (y_n) bornée, donc est également bornée. Elle admet une sous-suite $(z_{p_k})_{k\in\mathbb{N}}$ convergeant vers un certain réel β.
- La suite $(x_{n_{p_k}})_{k\in\mathbb{N}}$ est alors une sous-suite de $(x_{n_p})_{p\in\mathbb{N}}$ et donc converge vers α, tandis que $(y_{n_{p_k}})_{k\in\mathbb{N}}$ converge, par construction, vers β. On en déduit que la suite $(u_{n_{p_k}})_{k\in\mathbb{N}}$, extraite de (u_n) converge vers $\alpha + i\beta$. ∎

2.4.2 Valeurs d'adhérence

La définition suivante est utilisée pour obtenir des informations sur le *comportement asymptotique* d'une suite.

Définition 25. On dit qu'une suite (x_n) de $\mathbb{K}$ admet $a \in \mathbb{K}$ comme *valeur d'adhérence* s'il existe une suite extraite $(x_{n_k})_{k\in\mathbb{N}}$ convergeant vers a.

Proposition 48. Soit (x_n) une suite *bornée*. Si elle admet une *unique* valeur d'adhérence a, alors elle converge vers a.

Démonstration. On raisonne par l'absurde. Supposons que la suite (x_n) ne converge pas vers a, alors : $\exists\varepsilon > 0\,,\quad \forall N \in \mathbb{N},\ \exists n \geqslant N\,,\quad |x_n - a| > \varepsilon$.
Il existe donc une suite $(x_{n_k})_{k\in\mathbb{N}}$ extraite de (x_n), telle que, pour tout $k \in \mathbb{N}$, l'on ait $|x_{n_k} - a| > \varepsilon$. En appliquant le théorème de Bolzano-Weierstraß, on peut extraire de la suite $(x_{n_k})_{k\in\mathbb{N}}$ une suite convergente. D'après l'hypothèse, la limite de cette dernière ne peut être que a et l'on en déduirait $0 = a - a \geqslant \varepsilon$, d'où une contradiction. ■
Si la suite n'est pas bornée le résultat n'est pas vrai en général. Par exemple la suite $(x_n := n + (-1)^n n)$ admet 0 pour unique valeur d'adhérence. Soit en effet une suite extraite (x_{n_k}). Si l'ensemble des n_k pairs est fini, la suite est nulle à partir d'un certain rang et converge donc vers 0. Si l'ensemble des n_k pairs est infini, la suite n'est pas bornée et elle ne peut donc pas converger.
L'ensemble des valeurs d'adhérence d'une suite peut être infini. Nous en verrons des exemples plus loin.
Pour « deviner » les valeurs d'adhérence d'une suite, il est souvent efficace d'utiliser des outils numériques et graphiques (*cf.* ci-dessous page 567).

2.4.3 Étude des suites récurrentes réelles

Nous allons mettre en place une méthode d'étude des suites récurrentes réelles. Elle est suffisante pour traiter beaucoup d'exemples. Notons toutefois que certains cas d'apparence « anodine » ne peuvent pas être étudiés de façon approfondie sans outils « sophistiqués » (*cf.* la partie 2.4.4).

Applications contractantes et suites récurrentes

Proposition 49. Soient I un segment de $\mathbb{R}$ et $f : I \to I$ une application. On suppose qu'il existe $k \in \mathbb{R}$, $0 \leqslant k < 1$, tel que, pour tous $x,\ y \in I$, on ait $\big|f(x) - f(y)\big| \leqslant k|x - y|$. Alors, il existe un unique $\ell \in I$ tel que $f(\ell) = \ell$. De plus, pour tout $u_0 \in I$, la suite récurrente (u_n), définie par u_0 et f, converge vers ℓ.

Une application vérifiant les conditions ci-dessus est dite *contractante*.

Démonstration.
Existence. Soit $u_0 \in I$. On note (u_n) la suite récurrente définie par u_0 et f.
Montrons d'abord le résultat suivant : pour tous entiers p, q tels que $p \leqslant q$, on a :

$$|u_q - u_p| \leqslant \frac{k^p}{1-k}|u_1 - u_0|. \tag{9}$$

On montre par récurrence $|u_{n+1} - u_n| \leqslant k^n |u_1 - u_0|$. Alors, si $p < q$:

$$\begin{aligned} |u_q - u_p| &\leqslant |u_q - u_{q-1}| + \cdots + |u_{p+1} - u_p| \\ &\leqslant (k^{q-1} + \cdots + k^p)|u_1 - u_0| \leqslant \frac{k^p}{1-k}|u_1 - u_0|. \end{aligned}$$

D'après le théorème de Bolzano-Weierstraß (théorème 47 de la page 562), on peut extraire de (u_n) une suite convergente (u_{n_h}) ; on note $\ell \in I$ sa limite. Nous allons montrer que la suite (u_n) converge vers ℓ.

Soit $\varepsilon > 0$. Il existe $h_0 \in \mathbb{N}$ tel que, pour tout $h \geqslant h_0$, $|u_{n_h} - \ell| \leqslant \varepsilon/2$. Par ailleurs, il existe $h_1 \geqslant h_0$, tel que $\dfrac{k^{n_{h_1}}}{1-k}|u_1 - u_0| \leqslant \varepsilon/2$. On a $|u_{n_{h_1}} - \ell| \leqslant \varepsilon/2$.

Donc, pour tout $n \geqslant n_{h_1}$, on a, en utilisant (9) :

$$|u_n - \ell| \leqslant |u_n - u_{n_{h_1}}| + |u_{n_{h_1}} - \ell| \leqslant \frac{k^{n_{h_1}}}{1-k}|u_1 - u_0| + \frac{\varepsilon}{2} \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

La fonction f est évidemment continue, donc, d'après la proposition 39, $f(\ell) = \ell$.

Soient $v_0 \in I$ et (v_n) la suite récurrente définie par v_0 et f. On prouve, par récurrence sur n : $|v_n - \ell| \leqslant k^n |v_0 - \ell|$. Puisque k^n tend vers 0 quand $n \to +\infty$, la suite (v_n) converge vers ℓ.

Unicité. S'il existe $\ell \in I$ tel que $f(\ell) = \ell$, il est déterminé de façon unique par cette condition. Soit, en effet $\ell_1 \in I$ tel que $f(\ell_1) = \ell_1$. On a : $|\ell_1 - \ell| = |f(\ell_1) - f(\ell)| \leqslant k|\ell_1 - \ell|$ et, puisque $k < 1$, $|\ell_1 - \ell| = 0$, c'est-à-dire $\ell_1 = \ell$. ■

On peut donner une variante de la preuve de la proposition ci-dessus en utilisant l'exercice 29 de la page 554 au lieu du théorème de Bolzano-Weierstraß. Le lecteur le fera à titre d'exercice.

Définition 26. Soient I un intervalle *ouvert* et $f : I \to \mathbb{R}$ une application dérivable sur I. Soit $\ell \in I$ un point fixe de f. On dit que ℓ est *attractif* (resp. *super attractif*) si $|f'(\ell)| < 1$ (resp $f'(\ell) = 0$) et que ℓ est *répulsif* si $|f'(\ell)| > 1$.

La définition ci-dessus est justifiée par le résultat suivant.

Proposition 50. Soient I un intervalle ouvert et $f : I \to \mathbb{R}$ une application dérivable, à dérivée continue, sur I. Soit $\ell \in I$ un point fixe de f.

(i) On suppose que ℓ est *attractif*. Alors il existe $0 < k < 1$ et un segment $J \subset I$, de centre ℓ, non réduit à un point, tel que, pour tout $x \in J$, $|f'(x)| \leqslant k < 1$. S'il en est ainsi, $f(J) \subset J$ et, pour tout $u_0 \in J$, la suite récurrente $(u_{n+1} := f(u_n))$ est bien définie et admet ℓ pour limite.

(ii) On suppose que ℓ est *répulsif*. Soit $u_0 \in I$ tel que la suite récurrente $(u_{n+1} := f(u_n))$ soit bien définie. Si elle converge vers ℓ, elle est constante et égale à ℓ à partir d'un certain rang.

Notons que, dans (i), tout segment $J' \subset J$, de centre ℓ, est invariant par f.

Démonstration. Les preuves des deux assertions utilisent l'égalité des accroissements finis : pour tous $x, y \in I$, $x < y$, il existe $\xi \in]x, y[$ tel que $f(x) - f(y) = f'(\xi)(x - y)$, (*cf.* le théorème 44 de la page 648).

(i) Soit $k \in \mathbb{R}$ tel que $|f'(\ell)| < k < 1$. D'après la continuité de f', Il existe un segment $J \subset I$, de centre ℓ, non réduit à un point, tel que, pour tout $x \in J$, $|f'(x)| < k$. D'après l'égalité des accroissements finis, on a, pour tous $x,\ y \in J$, $\big|f(x) - f(y)\big| = \big|f'(\xi)\big||x - y|$ et, par suite, $\big|f(x) - f(y)\big| \leqslant k|x - y|$.

Alors, pour tout $x \in J$, $\big|f(x) - \ell\big| \leqslant k|x - \ell|$ et $f(J) \subset J$.

On termine la preuve en utilisant la proposition 49 de la page 563.

(ii) Soit $k \in \mathbb{R}$ tel que $|f'(\ell)| > k > 1$. D'après la continuité de f', Il existe un segment $J \subset I$, de centre ℓ, non réduit à un point, tel que, pour tout $x \in J$, $|f'(x)| > k$. D'après l'égalité des accroissements finis, on a, pour tous $x,\ y \in J$, $\big|f(x) - f(y)\big| = \big|f'(\xi)\big||x - y|$ et, par suite, $\big|f(x) - f(y)\big| \geqslant k|x - y|$. Soit $u_0 \in I$ tel que la suite récurrente $(u_{n+1} := f(u_n))$ soit bien définie.

Supposons qu'elle converge vers ℓ ; alors, il existe $N \in \mathbb{N}$ tel que, pour tout $n \geqslant N$, $u_n \in J$. On a $|u_{n+1} - \ell| = \big|f(u_n) - \ell\big| \geqslant k|u_n - \ell|$, donc, par récurrence $|u_{n+1} - \ell| \geqslant k^{n-N}|u_N - \ell|$. Par suite, $|u_{n+1} - \ell|$ ne peut tendre vers 0, quand $n \to +\infty$, que si $u_N = \ell$ (un point entrant dans J sans être égal à ℓ est « repoussé »). Ainsi la suite est constante et égale à ℓ à partir du rang N. ■

Exemples.

1. *Suites arithmético-géométriques réelles :* $f(x) := ax + b,\ a \neq 1$. Le point fixe est $\ell = \frac{b}{1-a}$ et $f'(\ell) = a$. Le point fixe est attractif si $|a| < 1$ et répulsif si $|a| > 1$.
2. On reprend l'exemple de la page 557.

   Soient $I := [-1, 1]$ et l'application $f : x \in I \mapsto -\frac{x^3}{2} + \frac{3}{2}x \in I$. Elle a 3 points fixes $0, \pm 1$; le point 0 est répulsif ($f'(0) = 3/2$) et les points ± 1 sont super attractifs ($f'(\pm 1) = 0$).
3. L'application polynômiale $x \mapsto 4x - x^2$ définit par restriction une application continue $f : [0, 4] \to [0, 4]$. Les points fixes de f sont 0 et 3. On a $f'(x) = 4 - 2x$, donc $f'(0) = 4$ et $f'(3) = -2$, par suite les deux points fixes sont *répulsifs*. Donc, si la suite définie par u_0 et f converge, elle est nécessairement stationnaire : elle est constante et égale à 0 ou à 3 à partir d'un certain rang. Cherchons les points u_0 tels qu'il en soit ainsi. Si $u_0 \in [0, 4]$, il existe $\alpha \in \mathbb{R}$ (unique si l'on impose $\alpha \in [0, \pi/2]$) tel que $u_0 = 4\sin^2\alpha$. On montre par récurrence que, pour tout $n \in \mathbb{N}$, $u_n = 4\sin^2(2^n\alpha)$, en utilisant $4\left(\sin^2(2^n\alpha) - \sin^4(2^n\alpha)\right) = 4\sin^2(2^n\alpha)\cos^2(2^n\alpha) = \sin^2(2^{n+1}\alpha)$. Donc la suite définie par u_0 est stationnaire si, et seulement s'il existe $n \in \mathbb{N}$ tel que $\sin 2^n\alpha = 0$ ou $\sin 2^n\alpha = \pm\sqrt{3}/2$, c'est-à-dire si, et seulement si, $\alpha = \frac{k\pi}{3.2^n}$, avec $k \in \mathbb{Z}$. Donc la suite est stationnaire si, et seulement si, $u_0 = 4\sin^2\frac{k\pi}{3.2^n}$, avec $k \in \mathbb{Z}$ et $n \in \mathbb{N}$. En utilisant l'exercice IV.1.12, on montre que les valeurs u_0 donnant des suites stationnaires sont denses dans $[0, 4]$ et que les valeurs u_0 donnant des suites divergentes sont denses dans $[0, 4]$.
4. *Suites homographiques réelles.* Soit $f : x \mapsto \frac{a\,x+b}{c\,x+d}$, avec $a, b, c, d \in \mathbb{R}$, $c \neq 0$ et $ad - bc \neq 0$. On suppose que les deux points fixes α et β sont distincts et réels. Nous allons préciser les propriétés d'attractivité de ces points.

On rappelle : $\frac{f(x)-\alpha}{f(x)-\beta} = k\,\frac{x-\alpha}{x-\beta}$, avec $k := \frac{c\beta+d}{c\alpha+d} \in \mathbb{R}^* \setminus \{1\}$. Calculons $f'(\alpha)$ et $f'(\beta)$. On a $(f(x)-\alpha)(x-\beta) = k\,(f(x)-\alpha)(x-\beta)$ et, en dérivant : $f'(x)(x-\beta)+f(x)-\alpha = k\,(f'(x)(x-\alpha)+f(x)-\beta)$. Si $x := \alpha$, on en déduit : $f'(\alpha) = k$ et, si $x := \beta$: $f'(\beta) = 1/k$. Par suite, si $|k| \neq 1$ (ce qui est équivalent à $k \neq -1$, puisque $k \in \mathbb{R} \setminus \{-1\}$), on a $|k| > 1$ ou $|k| < 1$ et l'un des points fixes est *attractif* et l'autre *répulsif*. Rappelons que nous avons montré plus haut que le point fixe attractif (α pour fixer les idées) « attire » toutes les suites récurrentes qui sont définies sauf la suite constante β, c'est-à-dire que la limite d'une telle suite est toujours α, et que la suite est définie pour des valeurs initiales appartenant à un intervalle ouvert « suffisamment petit » de centre α (*cf.* la page 551).

Le résultat suivant permet souvent de prouver la convergence des suites extraites des termes de rang pair et de rang impair d'une suit récurrente pour une valeur initiale bien choisie (il existe un énoncé analogue pour les τ-cycles ; la preuve est similaire mais un peu plus compliquée).

Corollaire 51. Soient I un intervalle *ouvert* et $f : I \to I$ une application dérivable, à dérivée continue, sur I. Soit (ℓ_1, ℓ_2) un 2-cycle de f. On suppose $|f'(\ell_1)f'(\ell_2)| < 1$. Alors :

- il existe un segment $J_1 \subset I$, de centre ℓ_1, non réduit à un point, tel que, pour tout $u_0 \in J_1$, les suites $(u_{2p})_{p\in\mathbb{N}}$ et $(u_{2p+1})_{p\in\mathbb{N}}$, extraites de (u_n), convergent respectivement vers ℓ_1 et ℓ_2 ;
- il existe un segment $J_2 \subset I$, de centre ℓ_2, non réduit à un point, tel que, pour tout $u_0 \in J_2$, les suites $(u_{2p})_{p\in\mathbb{N}}$ et $(u_{2p+1})_{p\in\mathbb{N}}$, extraites de (u_n), convergent respectivement vers ℓ_2 et ℓ_1.

Autrement dit, quand $n \to +\infty$, les termes de la suite (u_n) « se rapprochent alternativement des deux points du 2-cycle ». On dit que le 2 cycle est *attractif*.
Plus généralement si $(\ell_1, \ldots, \ell_\tau)$ est un τ-cycle de f, on dit qu'il est *attractif* (resp. *répulsif*) si $\left|(f^{\circ\tau})'(\ell_1)\right| = \cdots = \left|(f^{\circ\tau})'(\ell_\tau)\right| = |f'(\ell_1)\ldots f'(\ell_\tau)| < 1$ (resp. $\left|(f^{\circ\tau})'(\ell_1)\right| > 1$).

Démonstration. En utilisant la formule de dérivation de fonctions composées (*cf.* le théorème 36 de la page 643), On prouve que $f'(\ell_1)f'(\ell_2))$ est la dérivée de $f \circ f$ au point ℓ_1, et aussi au point ℓ_2. Par suite les points ℓ_1 et ℓ_2 sont des points fixes attractifs de $f \circ f$. Soit $0 < k < 1$. En appliquant la proposition ci-dessus à $f \circ f$ en ces points, on obtient des segments $L_q \subset I$ $(q = 1,\ 2)$, de centre ℓ_q, non réduits à un point, tel que, pour tout $x \in L_q$, $\left|(f \circ f))'\right|(x) < k$. Alors, pour tout $q = 1,\ 2$, $f \circ f(L_q) \subset L_q$, et, pour tout $v_{q-1} \in J_q$, la suite récurrente $\left(v_{p+1} := f \circ f(v_p)\right)_{p\geqslant q}$ est une suite de L_q et elle converge vers ℓ_q.

Comme nous l'avons noté plus haut, on peut remplacer le segment L_q par un segment plus petit et les propriétés énoncées restent vraies. En utilisant la continuité de f en ℓ_1 (resp. ℓ_2) et $f(\ell_1) = \ell_2$ (resp. $f(\ell_1) = \ell_2$) on prouve l'existence d'un segment $J_1 \subset L_1$ (resp. $J_2 \subset L_2$), de centre ℓ_1 (resp. ℓ_2), non réduit à un point, tel que $f(J_1) \subset L_2$ (resp. $f(J_2) \subset L_1$)).
On vérifie facilement que les segment J_1 et J_2 répondent à la question. ∎

Exemples.

1. Soit $f : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & x^2 - 1 \end{array}$ qui admet deux points fixes $\dfrac{1 \pm \sqrt{5}}{2}$.

   On a $f(0) = -1$ et $f(-1) = 0$, donc $(0, -1)$ est un 2-cycle pour f. On a $(f \circ f)'(0) = (f \circ f)'(-1) = 4f'(0)f'(-1) = 0$, donc le 2-cycle est *attractif* (et même super attractif). Pour des conditions initiales u_0 suffisamment voisines de 0 (resp. -1), les suites récurrentes définies par $f \circ f : x \mapsto x^4 - 2x^2$ convergent vers 0 (resp. -1) et les suites récurrentes définies par f oscillent entre des points qui se rapprochent du 2-cycle.

2. Soit $g : x \in \mathbb{R} \mapsto -\frac{3}{2}x^2 + \frac{5}{2}x + 1 \in \mathbb{R}$. On a $g(0) = 1,\ g(1) = 2,\ g(2) = 1$, donc $(0, 1, 2)$ est un 3-cycle.

   On a $g'(x) = -3x + \frac{5}{2}$ et $(g \circ g \circ g)'(0) = g'(0)g'(1)g'(2) = (-\frac{7}{2})(\frac{1}{2})(\frac{5}{2}) = \frac{35}{8} > 1$; le 3-cycle est donc *répulsif*.

Étude graphique des suites récurrentes

La méthode graphique décrite ci-dessous est un *guide* pour l'étude des suites récurrentes. Elle ne peut pas servir de preuve.

Soit U un intervalle ouvert de $\mathbb{R}$ (privé éventuellement d'un nombre fini de points). Soit $f : U \to \mathbb{R}$ une application dérivable et à dérivée continue.

Dans un repère orthonormé, on dessine le graphe Γ de f et la droite $\Delta : \{x = y\}$. On utilise ce dessin pour « deviner » les points fixes de f (ils correspondent à $\Gamma \cap \Delta$).

Pour « deviner » le comportement d'une suite récurrente définie par f et u_0, on part du point $(u_0, u_0) \in \Delta$. On trace ensuite la demi-droite *verticale* issue de ce point et dirigée vers Γ ; elle rencontre le graphe en $(u_0, f(u_0))$. On trace ensuite la demi-droite *horizontale* issue de ce point et dirigée vers Δ ; elle rencontre Δ en $(f(u_0), f(u_0)) = (u_1, u_1)$. En recommençant le procédé, à partir de (u_1, u_1), on obtient les points $(u_1, f(u_1))$, puis (u_2, u_2). Ensuite, on itère le processus.

Si l'on part d'une valeur u_0 « assez voisine » d'un point fixe a (tel que $|f'(a)| \neq 1$), on obtient un dessin « en escalier », si $f'(a) > 0$, ou « en escargot » (on dit aussi « en toile d'araignée »), si $f'(a) < 0$. L'escalier (resp. l'escargot) se rapproche du point fixe si a est attractif et s'en éloigne si a est répulsif. On a ainsi 4 dessins possibles (*cf.* la figure IV.1.2 de la page suivante) suivant la position de $f'(a)$ par rapport à $-1, 0, 1$.

En variant le choix de u_0, on essaie de comprendre ce qui peut se passer. On peut s'aider d'outils numériques. On programme la fonction f et, pour diverses valeurs de u_0, on calcule quelques termes de la suite (une vingtaine par exemple). En restreignant, quand c'est possible, f à un segment invariant I, on peut aussi, pour $u_0 \in I$, avoir une idée de l'ensemble des valeurs d'adhérence de la suite en demandant à l'ordinateur de dessiner *l'ensemble* $V_{M,N} = \{u_n\}_{M \leqslant n \leqslant N}$ *des valeurs prises par la suite entre deux rangs* M *et* N *suffisamment élevés* ($M := 100$ et $N := 150$ par exemple). Si l'on obtient « à peu près » un point (ou un nombre fini τ de points), on peut espérer que la suite converge (ou que l'on a τ suites extraites convergentes) et, sinon, il est vraisemblable que la suite ne converge pas et que l'on est dans un cas difficile. La figure IV.1.3 de la page 574 a été obtenue par cette méthode.

FIGURE IV.1.2 – Étude graphique de suites récurrentes

Récapitulatif des méthodes d'étude d'une suite récurrente

Nous donnons une méthode d'étude des suites récurrentes définies, pour simplifier la présentation, par un polynôme ou une fraction rationnelle f. Le lecteur adaptera cette méthode, au cas par cas, pour une fonction f plus générale (continûment dérivable). Dans tous les cas on s'aidera de l'étude graphique (et éventuellement numérique) de f (et, si nécessaire, de $f \circ f$).

Soit f un polynôme ou une fraction rationnelle ; f définit une fonction réelle sur $\mathbb{R}$ ou sur $\mathbb{R}$ privé d'un nombre fini de points.

On cherche d'abord les points fixes de f (ce sont des nombres algébriques dont il faut calculer une approximation décimale, au moins « grossière »). On cherche (s'il y en a...) ceux qui sont attractifs. On cherche ensuite un ou des segments(s) I invariants par f, contenant si possible au moins un point fixe attractif. Attention, il peut ne pas y avoir d'intervalle invariant et l'étude devient alors difficile !

– Si f est *contractante* sur I, il y a un seul point fixe et les suites partant de $u_0 \in I$ convergent vers ce point.
– Si f est *croissante* sur I, (u_n) est monotone et converge vers un point de I. C'est un point fixe, mais il peut y en avoir plusieurs, et il faut alors trouver lequel.
– Si f est *décroissante* sur I, les suites des termes pairs et impairs sont monotones ; l'une est croissante, l'autre est décroissante et elles convergent vers des points fixes de $f \circ f$. La suite diverge si ces points fixes sont différents, elle converge sinon.
– S'il y a un ou plusieurs segments invariants, il faut voir si, partant d'un point u_0 n'appartenant à aucun de ces segments, u_n finit par rentrer dans l'un d'entre eux et déterminer lequel. Il peut aussi arriver arriver que u_n soit l'un des points fixes répulsifs ; alors la suite reste ensuite en ce point fixe. La suite peut aussi tendre vers l'infini ou diverger « de façon compliquée ».

Etude de suites récurrentes par changement de coordonnées

Soient I un intervalle réel et $f : I \to \mathbb{R}$ une fonction réelle. Dans certains cas, il peut être utile de ramener l'étude des suites récurrentes définies par f à celle des suites récurrentes définies par une autre fonction $g : J \to \mathbb{R}$ en utilisant un *changement de variable*. Par exemple si le cas de g est déjà connu ou « plus simple ». Nous avons déjà utilisé cette méthode pour les suites arithmético-géométriques et les suites homographiques. Plus précisément, soit $h : I \to J$ une application continue bijective telle que $h^{-1} : J \to I$ soit continue (rappelons qu'alors h s'appelle est un homéomorphisme de I sur J) ; on dit que $g : J \to \mathbb{R}$ est obtenu à partir de $f : I \to \mathbb{R}$ par le changement de variable h si $g = h \circ f \circ h^{-1}$. On dit aussi que g est obtenu à partir de f par *conjugaison* (par h).
On a alors, pour tout $n \in \mathbb{N}$, $g^{\circ n} = (f \circ h^{-1})^{\circ n} = h \circ f^{\circ n} \circ h^{-1}$, donc $g^{\circ n} \circ h = h \circ f^{\circ n}$.
Soient $x_0 \in I$ et $y_0 := h(x_0)$. Notons $x_n = f^{\circ n}(x_0)$ et $y_n = g^{\circ n}(y_0)$. On a $y_n = h(x_n)$ et $x_n = h^{-1}(y_n)$, donc les études des suites récurrentes (x_n) et (y_n) sont « équivalentes ». Les deux suites ont des propriétés « similaires ». On prouve facilement par exemple que :

- si $\lim x_n = \ell$, alors $\lim y_n = h(\ell)$;
- si $(x_0, \ldots, x_{p-1})$ est un p-cycle de f, $(y_0, \ldots, y_{p-1})$ est un p-cycle de g ;
- si l'image de la suite (x_n) est dense dans I, l'image de la suite (y_n) est dense dans J.

L'inverse d'un homéomorphisme est un homéomorphisme et la composée de deux homéomorphismes est un homéomorphisme. On en déduit que la conjugaison est une relation d'équivalence.

Exemples. On définit une application, dite logistique (*cf.* ci-dessous l'exemple de la page 570) $f_4 : [0,1] \to [0,1]$, par $x \mapsto f_4(x) = 4x(1-x)$.

1. On définit une application $g : [-2,2] \to [-2,2]$ par $x \mapsto g(x) = x^2 - 2$. Montrons que f_4 et g sont conjuguées.
   On définit $h : [0,1] \to [-2,2]$ par $x \mapsto h(x) := -4x + 2$. c'est évidemment un homéomorphisme.
   On a $g \circ h(x) = h(x)^2 - 2 = 16x^2 - 16x + 2 = h\left(4x(1-x)\right) = h \circ f(x)$.

2. Nous allons montrer que l'application logistique f_4 et l'application tente T (définie dans l'exemple de la page 529) sont conjuguées.

   On définit $h : [0,1] \to [0,1]$, par $x \mapsto h(x) = \dfrac{2}{\pi} \arcsin \sqrt{x}$. Cette application est un homéomorphisme de $[0,1]$ sur lui même.
   L'application inverse est $y \mapsto h^{-1}(y) = \sin^2(\pi x/2)$.
   Vérifions que $T \circ h = h \circ f_4$. Cette relation est équivalente à $f_4 \circ h^{-1} = h^{-1} \circ T$.
   Pour $x \in [0, 1/2]$, on a :

$$f_4\left(h^{-1}(x)\right) = 4\sin^2 \frac{\pi x}{2}(1 - \sin^2 \frac{\pi x}{2}) = \sin^2 \pi x = h^{-1}(2x) = h^{-1}\left(T(x)\right).$$

   Pour $x \in [1/2, 1]$, on a :

$$f_4\left(h^{-1}(x)\right) = \sin^2 \pi x = (\sin(\pi - \pi x))^2 = \left(\sin \frac{\pi(2-2x)}{2}\right)^2 = h^{-1}\left(T(x)\right).$$

Exercice 34.
On considère l'application tente $T : [0,1] \to [0,1]$. Montrer que l'ensemble des conditions initiales $u_0 \in [0,1] \to [0,1]$ telles que la suite récurrente définie par T et u_0 converge forment un sous-ensemble dénombrable dense de $[0,1] \to [0,1]$ (on utilisera l'exemple 3 de la page 565).
Solution. Notons $\varphi : [0,4] \to [0,4]$ l'application définie par $\varphi : t \mapsto 4t - t^2$. On fait le changement de variable $h : t \mapsto x := t/4$. On vérifie $f_4 = h \circ \varphi \circ h^{-1}$. Ainsi f_4 et φ sont conjuguées. Par transitivité T et φ sont conjuguées ; d'où le résultat en utilisant l'exemple 3 de la page 565.

2.4.4 Quelques exemples plus difficiles

Suites logistiques et cascade de Feigenbaum

Des lapins et des hommes : la modélisation de l'évolution des populations par des suites. La modélisation de l'évolution d'une population au cours du temps fait apparaitre des suites récurrentes. Le premier exemple est dû à Fibonacci dans son *Liber Abaci* en 1202 (sous forme de problème « récréatif »). Il « modélise » la croissance d'une population de lapins à partir d'un couple initial, avec l'hypothèse qu'un couple de lapins se reproduit tous les mois à partir de l'age de trois mois. Ceci le conduit à la suite dite de Fibonacci. On montre que la croissance de celle-ci est « en gros » exponentielle (*cf.* les exercices 27 et IV.1.44).
Un exemple concret de la croissance « exponentielle » d'une population de lapins est celui observé en Australie. En 1859 un fermier a introduit dans ce pays 27 couples de lapins et 6 ans plus tard leur nombre était de l'ordre de 22 millions !
Le début des études « sérieuses » de la croissance des populations est, au XVIIIe siècle, le travail de Leonhard Euler. En 1760, dans ses *Recherches générales sur la mortalité et la reproduction du genre humain*, il a propose un modèle de l'évolution de la population d'une ville ou d'une province basé sur des études « statistiques »[13]. Il fait l'hypothèse que les probabilités d'atteindre un certain âge et le taux de natalité sont *invariants dans le temps* (il suppose implicitement l'absence d'émigration, d'immigration, de guerre ou d'épidémie catastrophique). Il montre alors que l'évolution de la

[13] Motivées par des préoccupations très concrètes de calculs de rentes viagères...

population suit *une progression géométrique* dont il calcule la raison Λ. Si $\Lambda > 1$ (ce qui et le cas si le taux de natalité est assez grand), la population croît exponentiellement et si $\Lambda < 1$, elle décroît. Plus tard, en 1798, l'idée de croissance exponentielle d'une population a été reprise en Angleterre par l'économiste Thomas Robert Malthus qui a, de nouveau, modélisé la population humaine comme une suite géométrique tout en modélisant les « ressources » par une suite arithmétique. Ceci a conduit a une doctrine de restriction de la natalité appelée malthusianisme. Celle-ci a été appliquée sous diverses formes par certains gouvernements dans le monde, par exemple, en Chine de 1979 à 2015 avec la politique de l'enfant unique.

Le modèle de croissance géométrique d'une population a été remis en cause vers 1840 par le mathématicien belge Pierre François Verhulst [14]. L'idée de départ (déjà présente chez Malthus) est que, si la population augmente, des facteurs de limitation apparaissent (manque de ressources, place disponible...). Verhulst suppose qu'il y a une population maximale P et que le coefficient Λ d'augmentation de la population p n'est plus constant et qu'il diminue quand p se rapproche de P. Il choisit l'hypothèse la plus simple, c'est-à-dire que Λ est proportionnel à $P-p$: $\Lambda := \lambda(P-p)$. Si p_n est la population la $n^{ème}$ année, alors $p_{n+1} = \lambda p_n(P-p_n)$. En utilisant la variable $x := p/P$, on obtient $x_{n+1} = \mu x(1-x)$. En fait Verhulst fait une modélisation continue : au lieu de suites il utilise des équations différentielles, qu'il appelle *logistiques*[15] ; *cf.* l'exercice IV.1.48. Par analogie, on appelle *suite logistique* la suite (x_n). Le modèle de Verhulst s'est révélé assez efficace : il a par exemple permis de prédire en 1837 la population de la France en 1930 avec une bonne précision. Au début du XXe siècle, divers scientifiques ont retrouvé le modèle logistique et l'ont appliqué efficacement à des problèmes divers : croissance d'organismes individuels, de populations de micro organismes ou d'humains (population des États Unis d'Amérique). Plus récemment, le modèle logistique a perdu de sa pertinence en dynamique des populations : on a constaté, entre autres, que, des facteurs sociologiques et idéologiques influent fortement sur le taux de natalité[16] et qu'ils sont à prendre en compte dans les modèles. Mais parallèlement divers scientifiques[17] ont constaté en étudiant les *suites* logistiques qu'elles peuvent avoir des comportements « compliqués et intéressants » (dépendant fortement du paramètre μ). Ainsi, en même temps que les suites logistiques perdaient de leur intérêt en écologie, elles faisaient le bonheur des mathématiciens[18]...

On note $I := [0,1]$. Pour tout $\mu > 0$ fixé, l'application logistique f_μ est l'application quadratique $f_\mu : \mathbb{R} \to \mathbb{R}$ définie par $f_\mu : x \mapsto \mu x(1-x)$. On considère μ comme le paramètre d'une *famille d'applications*. Le comportement des suites récurrentes définies par f_μ dépend des valeurs du paramètre μ. Pour suivre graphiquement l'analyse ci-dessous, le lecteur dessinera le graphe de quelques unes des fonctions f_μ. Ce sont des paraboles passant par les points $(0,0)$, $(0,1)$ et de sommet $(1/2, \mu/4)$. Les fonctions f_μ sont croissantes sur $[0,1/2]$ et décroissantes sur $[1/2,1]$.

Si $x \in I$, $f_\mu(x) \geqslant 0$. En utilisant l'équation $\mu t^2 - \mu t + 1 = 0$, on vérifie que $f_\mu(I) \subset I$ si, et seulement si, $\mu \leqslant 4$. On suppose donc dans la suite $0 < \mu \leqslant 4$ et l'on inter-

[14] Recherches mathématiques sur la loi d'accroissement de la population, Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles, 1845.

[15] L'origine de cette terminologie semble inconnue.

[16] Ceci apparaît déjà dans les réflexions de Verhulst.

[17] Le premier fut, en 1976, le biologiste australien Robert M. May : *Simple mathematical models with very complicated dynamics*, Nature vol. 261, n° 5560, p. $459-467$.

[18] Et de certains physiciens : les subtilités du comportement des suites logistiques apparaissent dans certains dispositifs expérimentaux comme par exemple les lasers à CO_2.

prète f_μ comme une application de I dans I. Les *points fixes* (possibles) de f_μ sont donnés par $f_\mu(x) = x$, c'est-à-dire par $x(\mu x + 1 - \mu) = 0$. On obtient $\alpha_0 = 0$ et $\alpha_1 := 1 - \frac{1}{\mu}\cdot$
On a $f'_\mu(x) = \mu(1 - 2x)$, donc $f'_\mu(0) = \mu$ et $f'_\mu(\alpha_1) = 2 - \mu$.
Quand μ varie, pour certaines valeurs de μ : $\mu_0 := 1$, $\mu_1 := 3$, $\mu_2 := 1 + \sqrt{6}, \ldots$, on observe un « changement de comportement » des suites récurrentes ; c'est ce que l'on appelle des *phénomènes de bifurcations*.
Si $0 < \mu \leqslant 3$, pour toute valeur initiale $u_0 \in I$, la suite logistique est convergente. Pour $\mu > 3$, la suite ne converge plus, en revanche les suites d'ordres pair et impair convergent : un 2-cycle apparaît. Pour $\mu > 1 + \sqrt{6} = \mu_2$, on constate, quand μ augmente, des phénomènes de plus en plus complexes, surprenants pour le débutant...

- Si $0 < \mu \leqslant 1$ il n'y a qu'un point fixe dans I : $\alpha_0 = 0$.
  On a $f_\mu(x) - x = \mu x(\alpha_1 - x) \leqslant 0$. Toutes les suites récurrentes sont décroissantes et tendent vers 0. Si $\mu < 1$, alors 0 est attractif : $0 < f'_\mu(0) = \mu < 1$.

- Si $\mu > 1$, $f'_\mu(0) = \mu$ donc 0 est devenu *répulsif* et le second point fixe $\alpha_1 = \frac{\mu-1}{\mu}$ appartient à I.

- Si $1 < \mu < 3$, $|f'_\mu(\alpha_1)| = |2 - \mu| < 1$: le point fixe α_1 est *attractif*.
  On distingue les cas $1 < \mu \leqslant 2$ et $2 < \mu < 3$. Nous donnons seulement les résultats (les preuves sont élémentaires mais nécessitent un peu de travail, elles font l'objet de l'exercice IV.1.49 de la page 608).
  – Supposons $1 < \mu \leqslant 2$. Si $u_0 = 0$ ou $u_0 = \alpha_1$, la suite est constante. Si $u_0 = 1$ (resp. $u_0 = 1/\mu$), la suite est constante et égale à 0 (resp. α_1) à partir du rang 1. Si $u_0 \in]0, \alpha_1[$ (resp. $u_0 \in]\alpha_1, 1/2[$), alors la suite est croissante (resp. décroissante) et converge vers α_1. Si $u_0 \in]1/2, 1/\mu[$ (resp. $u_0 \in]1/\mu, 1[$), alors $u_1 \in]\alpha_1, 1/2[$ (resp. $u_1 \in]0, \alpha_1[$) et l'on est ramené au cas précédent. Si $\mu = 2$, $f'_\mu(\alpha_1) = 0$ et le point α_1 est *super attractif*.
  – Supposons $2 < \mu < 3$. Le point fixe α_1 est *attractif*, avec une dérivée $f'_\mu(\alpha_1) = 2 - \mu$ *négative*. D'après la proposition 50, pour u_0 suffisamment voisin de α_1, la suite converge vers α_1 avec un comportement « en escargot ». On montre qu'en fait, *pour tout* $u_0 \in]0, 1[$, la suite converge vers α_1.

- Si $\mu = 3$, $f'_\mu(\alpha_1) = -1$. Les suites récurrentes convergent vers α_1 (lentement...) alternativement par valeurs supérieures et inférieures. Nous l'admettrons.

- Si $\mu > 3$, $f'_\mu(\alpha_1) = 2 - \mu < -1$. Les deux points fixes 0 et α_1 sont maintenant *répulsifs* (donc les suites récurrentes convergentes sont nécessairement stationnaires et de limite 0 ou α_1). Nous allons étudier (dans le cas où (u_n) ne converge pas) les suites récurrentes des termes pairs et impairs de (u_n) définies par $f_\mu \circ f_\mu$. Le graphe de $f_\mu \circ f_\mu$ est « en bosse de chameau » et il a 4 points d'intersection avec la droite $\{y = x\}$. On a :
  $$f_\mu \circ f_\mu(x) = f_\mu\left(f_\mu(x)\right) = f_\mu(\mu x - \mu x^2) = \mu^2 x - (\mu^2 + \mu^3)x^2 + 2\mu^3 x^3 - \mu^3 x^4.$$
  L'équation aux points fixes de $f_\mu \circ f_\mu$ est $f_\mu \circ f_\mu(a) = a$. Elle s'écrit :
  $$\mu^3 a\left(a - (1 - \mu^{-1})\right)\left(a^2 - (1 + \mu^{-1})a + \mu^{-1}(1 + \mu^{-1})\right) = 0.$$

Après élimination des points fixes 0 et $\alpha_1 = 1 - \mu^{-1}$ de f, il reste l'équation du second degré (où $\mu \geqslant 3$) : $a^2 - (1+\mu^{-1})a + \mu^{-1}(1+\mu^{-1}) = 0$ dont les solutions sont $\alpha_{2,\pm} = \frac{1}{2} + \frac{1}{2\mu} \pm \frac{\sqrt{(\mu-3)(\mu+1)}}{2\mu}$. Pour $\mu = 3$, on a une racine double $\alpha_{2,+} = \alpha_{2,-}$. Pour $\mu > 3$, $\alpha_{2,+} \neq \alpha_{2,-}$ et, en prenant $\alpha_{2,\pm}$ comme condition initiale, on obtient un 2-cycle. Un calcul simple donne :

$$\begin{aligned}(f_\mu \circ f_\mu)'(\alpha_{2,\pm}) &= f'_\mu(\alpha_{2,+}) f'_\mu(\alpha_{2,-}) = (\mu - 2\mu\alpha_{2,+})(\mu - 2\mu\alpha_{2,-}) \\ &= \mu^2 - 2\mu^2(\alpha_{2,+} + \alpha_{2,-}) + 4\mu^2\alpha_{2,+}\alpha_{2,-} = -\mu^2 + 2\mu + 4.\end{aligned}$$

Pour $\mu = 3$, on a $\alpha_{2,+} = \alpha_{2,-}$ et $(f_\mu \circ f_\mu)'(\alpha_{2,\pm}) = 1$.

Si $3 < \mu < 1 + \sqrt{6} = \mu_2$, on a : $-1 < (f_\mu \circ f_\mu)'(\alpha_{2,\pm}) < 1$: le 2-cycle est attractif. Si $u_0 \neq 0$ et $u_0 \neq 1$, on peut montrer que les suites extraîtes de (u_n), respectivement formées des termes pairs et impairs, convergent chacune vers l'un des deux points $\alpha_{2,\pm}$ (à l'exception de certaines suites constantes et égales à 0 ou α_1 à partir d'un certain rang). Si $\mu > \mu_2 := 1 + \sqrt{6} \sim 3{,}449490$, le 2-cycle devient *répulsif*.

Il est très intéressant de voir ce qui se passe après la troisième bifurcation, c'est-à-dire pour $\mu_2 < \mu \leqslant 4$, comme l'a fait le mathématicien américain Mitchell Feigenbaum en 1975. Nous allons le décrire, sommairement et sans démonstration.

Pour $\mu > \mu_2$, le 2-cycle devient répulsif. Par contre un 4-cycle apparaît : on dit qu'il y a eu un nouveau *doublement de période*. Le cycle est abord attractif, puis il devient répulsif, après une nouvelle bifurcation en $\mu_3 \sim 3{,}544090$ et un 8-cycle apparaît. En pratique, on étudie les points fixes de $f_\mu^{\circ 4}$. Cette application a 4 « nouveaux » points fixes formant le 4-cycle.

On étudie ensuite successivement les points fixes de $f^{\circ 2^m}$ $(m > 2)$. La suite croissante (μ_m) des valeurs de bifurcation est infinie et majorée par 4. Elle est donc convergente. On note μ_∞ sa limite : $\mu_\infty = 3{,}569946\ldots$.

On montre que, pour tout $m \in \mathbb{N}^*$, μ_m est la borne inférieure des μ tels que f_μ admette un 2^m-cycle. Pour $\mu_m < \mu < \mu_\infty$, f_μ admet un 2^m-cycle qui est attractif pour $\mu < \mu_{m+1}$ et répulsif pour $\mu > \mu_{m+1}$. Pour $\mu < \mu_\infty$, f_μ n'admet pas de τ-cycle si τ n'est pas une puissance de 2.

Pour $\mu_\infty \leqslant \mu \leqslant 4$, la situation devient extrêmement compliquée et l'étude nécessite des outils mathématiques sophistiqués, on dit que l'on est dans la « zone chaotique ». On peut, par exemple, montrer que pour $\mu \geqslant \mu_\star := 1 + 2\sqrt{2} = 3{,}828427\ldots > \mu_\infty$, il existe au moins un cycle d'ordre 3 (unique pour $\mu_\star$), et qu'il n'en existe pas pour $\mu < \mu_\star$. En utilisant un résultat dû à Sarkovskii (*cf.*, ci-dessous, la proposition 52), on peut en déduire que, pour $\mu_\star \leqslant \mu \leqslant 4, f_\mu$ a des cycles de *toutes les périodes*.

La *cascade de Feigenbaum* est le graphe obtenu en portant en ordonnée, pour chaque valeur de $\mu \in [0,4]$, l'ensemble V des valeurs d'adhérence possibles pour une suite récurrente définie par f_μ avec une même condition initiale u_0. En pratique on remplace V par un ensemble $V_{M,N}$, défini comme à la fin de l'étude graphique ci-dessus (*cf.* la page 567), avec la condition initiale $u_0 := 1/2$ (les figures IV.1.3 de la page suivante ont été obtenues par cette méthode). Le lecteur trouvera un très grand nombre d'images de la cascade sur internet... On observera que la cascade donne un « brouillard de gouttelettes » pour $\mu \geqslant \mu_\infty$ et que dans ce brouillard s'ouvrent des fenêtres où l'on peut observer des « modèles réduits de la cascade ».

FIGURE IV.1.3 – Cascade de Feigenbaum

Remarque.
Il existe (au moins...) un millier d'articles, passages de livres, documents..., consacrés à l'étude des suites récurrentes définies par un polynôme quadratique réel ou complexe $f : x \mapsto ax^2 + bx + c$. Moyennant un *changement affine de coordonnée* sur $\mathbb{K}$: $x \mapsto \alpha x + \beta$ (ou conjugaison affine), une telle étude se ramène à celle du cas logistique (avec $\mu \in \mathbb{K}$), ou si l'on préfère, à celle du cas de la famille $g_\gamma : x \mapsto x^2 - \gamma$, paramétrée par $\gamma \in \mathbb{K}$, que l'on rencontre très souvent dans la « littérature ».[19] Ceci fait l'objet de l'exercice IV.1.50 de la page 609.

Exemples de suites périodiques et de suites d'image dense

Cette partie est consacrée à des exemples de comportements « compliqués » de suites récurrentes définies par des fonctions « simples ». Elle peut être omise en première lecture. Les preuves les plus délicates sont proposées en exercice.

Remarque. Soit $f : X \to X$. S'il existe *une* condition initiale $u_0 \in X$ telle que la suite récurrente (u_n) soit d'image dense dans X, alors il existe un ensemble *dense* de conditions initiales v_0 telles que (v_n) soit d'image dense. En effet, pour tout $k \in \mathbb{N}$, la suite récurrente (v_n) définie par $v_0 := u_k$ et $v_{n+1} = f(v_n)$ est d'image dense dans X.

Exemple 1 : suites géométriques complexes.
Notons r la raison d'une progression géométrique. On suppose $u_0 \neq 0$ et l'on définit : $(u_n := u_0 r^n)$.
On suppose $\mathbb{K} := \mathbb{C}$ et $|r| = 1$. Il existe $\xi \in [-1/2, 1/2[$, unique, tel que $r = e^{2i\pi\xi}$. Il faut distinguer les cas $\xi \in \mathbb{Q}$ et $\xi \notin \mathbb{Q}$.
– Si $\xi = 0$, la suite est constante.

[19] Le célèbre *ensemble de Mandelbrot* apparaît dans l'étude de cette famille : c'est l'ensemble des valeurs de $\gamma \in \mathbb{C}$ telles que la suite récurrente définie par g_γ et $u_0 := 0$ reste bornée. Le lecteur trouvera sur internet une myriade de magnifiques dessins et animations en couleurs en relation avec cet ensemble.

– Si $\xi \in \mathbb{Q}^*$, $\xi = p/q$ ($p \in \mathbb{N}^*$, $q \in \mathbb{N}^*$, p et q premiers entre eux). Alors r est une racine primitive $q^{ème}$ de l'unité (*cf.* le module II.5, définition 13 de la page 260) et la suite géométrique (u_n) est périodique de plus petite période q.
– Si $\xi \notin \mathbb{Q}$, on peut montrer que l'image de la suite géométrique est *dense* dans le cercle $\mathcal{C} \subset \mathbb{C}$, de centre 0 et de rayon $|u_0|$ (pour une preuve *cf.* l'exercice IV.1.52 de la page 609).

Exemple 2 : suites homographiques réelles.
Soit (u_n) une suite homographique définie par une relation de récurrence de la forme $u_{n+1} = f(u_n)$ avec $f : x \mapsto \frac{a\,x+b}{c\,x+d}$. On suppose $a, b, c, d \in \mathbb{R}$, $ad - bc \neq 0$ et $c \neq 0$. On suppose, de plus, que f admet deux points fixes *distincts* α et β (réels ou complexes conjugués). Si $u_0 \neq \beta$, on pose $v_n := \frac{u_n-\alpha}{u_n-\beta}$ et l'on se ramène à l'étude d'une suite géométrique $(v_n := k^n v_0)$, avec $k := \frac{c\beta+d}{c\alpha+d} \in \mathbb{C}^*$. On a $k \neq 1$. Par suite les suites récurrentes sont bornées si, et seulement si, $|k| \leqslant 1$.
Alors :

1. Si $|k| < 1$, la suite (v_n) converge vers 0 et la suite (u_n) converge vers le point fixe α ($\alpha = \frac{0-\alpha}{0-1}$). Pour le cas où $\alpha \in \mathbb{R}$, ceci généralise le résultat obtenu plus haut (*cf.* l'exemple 1 ci-dessus) en utilisant le fait que le point fixe α est *attractif* ($k = f'(\alpha)$) ; la preuve est *mutatis mutandis* la même.
2. Si β et α sont réels et $|k| = 1$. Alors k est réel, donc $k = -1$. On a alors $a + d = 0$ et $f = f^{-1}$, donc, pour tout $u_0 \in \mathbb{R} \setminus \{\alpha, \beta\}$ la suite u_n est périodique de période deux.
3. Si $\beta = \bar{\alpha}$, il n'y a pas de point fixe réel. La suite ne converge donc jamais, quelle que soit la valeur initiale $u_0 \in \mathbb{R}$. On peut préciser le type de divergence.

   Nous avons vu que, dans ce cas, on a $|k| = 1$. Il existe donc $\xi \in \mathbb{R}$ tel que $k = e^{2i\pi\xi}$.

   L'homographie *complexe* $h : x \mapsto \frac{x-\alpha}{x-\bar{\alpha}}$ induit une bijection entre la droite rélle $\mathbb{R}$ et le cercle unité privé d'un point $U \setminus \{1\}$. L'homographie inverse est $h^{-1} : \zeta \mapsto \frac{\bar{\alpha}\zeta-\alpha}{y-1}$; elle induit la bijection inverse entre $U \setminus \{1\}$ et $\mathbb{R}$.

   - Si $\xi \in \mathbb{Q} \setminus \mathbb{N}$, $\xi = p/q$ ($p \in \mathbb{Z}$, $q \in \mathbb{N}^*$, $q \geqslant 2$), k est une racine $q^{ème}$ de l'unité et la suite récurrente est périodique de période q. Si p et q sont premiers entre eux, k est une racine *primitive* et q est *la plus petite période*. On note $\mathcal{Z} := \{k^m \mid m \in [\![1, q-1]\!]\}$; c'est l'ensemble des racines $q^{èmes}$ de l'unité *différentes de* 1. Pour tout point v_0 de cet ensemble il existe un terme de la suite géomérique $(v_n := v_0 k^n)$ qui est égal à 1. L'ensemble Z des points u_0 pour lesquels la suite n'est pas définie est $h^{-1}(\mathcal{Z})$. Cet ensemble est *fini* ($\operatorname{card} Z = q - 1$). Pour toute condition initiale $u_0 \in X := \mathbb{R} \setminus Z$, on obtient un q*-cycle*.
   - Si $\xi \in \mathbb{R} \setminus \mathbb{Q}$, les suites récurrentes ont un comportement compliqué. En particulier une telle suite est dense dans $\mathbb{R}$ et l'ensemble des valeurs initiales telles que la suite ne soit pas définie est dénombrable et dense dans $\mathbb{R}$ (on trouvera des preuves dans l'exercice IV.1.53). Le complémentaire dans $\mathbb{R}$ d'un ensemble dénombrable est non dénombrable et dense (*cf.* l'exercice 26) , donc les suites récurrentes sont définies pour un ensemble de valeurs initiales non dénombrable et dense dans $\mathbb{R}$.

On peut montrer que, pour tout $\alpha \in \mathbb{C} \setminus \mathbb{R}$ et tout $k \in \mathbb{C} \setminus \{0, 1\}$, il existe une homographie *réelle* admettant α et $\bar{\alpha}$ pour points fixes et k pour « invariant » (*cf.* l'exercice 26 de la page 548). Ainsi, *tous* les cas décrits ci-dessus existent (ce qui n'était pas *a priori* évident...). Les deux exercices suivant donnent des exemples.

Exercice 35.

On considère l'homographie réelle $f : x \mapsto f(x) = \frac{x+\sqrt{3}}{-\sqrt{3}x+1}$. Calculer $f \circ f \circ f$. Calculer k. Que peut-on dire des suites récurrentes réelles définies par f ?

Solution. Pour tout $x \in \mathbb{R}$, $x \neq \pm 1/\sqrt{3}$, on a : $f \circ f(x) = \frac{x-\sqrt{3}}{\sqrt{3}x+1}$ et $f \circ f \circ f(x) = x$. On en déduit $g \circ g \circ g = 1$ (avec $g : y \mapsto ky$) donc $k^3 = 1$, c'est-à-dire $k = \pm j = -\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$. On a $\alpha = \pm i$. Si l'on choisit $\alpha := i$, on obtient $k = j^2$. Les suites récurrentes réelles définies par f sont périodiques de période 3. Plus précisément *toutes* les conditions initiales réelles différentes de $\pm 1/\sqrt{3}$ donnent des 3-cycles (puisque les points fixes $\pm i$ sont imaginaires purs). L'application f induit une *bijection* de $\mathbb{R} \setminus \{\pm 1/\sqrt{3}\}$ sur lui même ; l'application inverse est $y \mapsto \frac{\sqrt{3}-y}{\sqrt{3}y+1}$.

Exercice 36.

On considère l'homographie réelle $f : x \mapsto f(x) = \dfrac{\left(\cos\frac{\pi}{\sqrt{2}}\right) x + \sin\frac{\pi}{\sqrt{2}}}{-\left(\sin\frac{\pi}{\sqrt{2}}\right) x + \cos\frac{\pi}{\sqrt{2}}}$.

Calculer les points fixes et k. Montrer que k n'est pas une racine de l'unité. Que peut-on dire des suites récurrentes réelles définies par f ? Que peut-on dire de l'ensemble des conditions initiales pour lesquelles la suite n'est pas définie.

Solution. Les points fixes sont $\pm i$ et, si l'on choisit $\alpha := i$, on obtient :

$$k = \frac{\cos\frac{\pi}{\sqrt{2}} + i\sin\frac{\pi}{\sqrt{2}}}{\cos\frac{\pi}{\sqrt{2}} - i\sin\frac{\pi}{\sqrt{2}}} = e^{i\pi\sqrt{2}} = e^{2i\pi\xi}, \qquad \text{avec} \qquad \xi = 1/\sqrt{2}.$$

Nous savons que $1/\sqrt{2} \notin \mathbb{Q}$, donc k n'est pas une racine de l'unité. Par suite, les suites récurrentes qui sont définies sont d'image dense dans $\mathbb{R}$. Les conditions initiales pour lesquelles la suite n'est pas définie forment aussi un ensemble dense de $\mathbb{R}$.

Il existe des applications continues *admettant des cycles de toutes les périodes*. Nous allons en donner des exemples dont la définition est « simple » (ce qui ne veut pas dire que l'étude l'est...).

Exemple 3 : doublement de l'angle.

On note $U := \{z \in \mathbb{C} \mid |z| = 1\}$. On considère l'application $\varphi : \begin{array}{ccc} U & \longrightarrow & U. \\ z & \longmapsto & z^2 \end{array}$

On l'interprète souvent en identifiant $z \in U$ à son argument θ défini modulo 2π ; φ est alors l'application de *doublement de l'angle* : $(\theta \mod 2\pi) \mapsto (2\theta \mod 2\pi)$.

Son itérée $n^{\text{ème}}$ est $\varphi : z \in U \mapsto z^{2^n}$ que l'on interprète par $(\theta \mod 2\pi) \mapsto (2^n\theta \mod 2\pi)$.

L'application φ est un homomorphisme surjectif de groupes. Son noyau est $\{\pm 1\}$. Si la suite récurrente définie par $u_0 \in U$ et φ est constante et égale à 1 à partir d'un certain rang, il existe $m \in \mathbb{N}$ tel que $1 = \varphi^{\circ m}(u_0) = u_0^{2^m}$.

Le comportement des suites récurrentes définies par φ et $u_0 \in U$ est compliqué. On a les résultats suivants, dont des preuves sont proposées dans l'exercice IV.1.66 de la page 616 (ces preuves utilisent l'écriture en base 2 des nombres réels).

- La valeur initiale $z_0 \in U$ définit une suite récurrente périodique à partir d'un certain rang si, et seulement si, $z_0 = \mathrm{e}^{2i\pi t}$ avec $t \in \mathbb{Q}$.
- Parmi les suites récurrentes définies par φ il y a des cycles *de toutes les périodes*.

- Les suites récurrentes définies par φ ont des propriétés de *sensibilité aux conditions initiales* : une modification « aussi petite que l'on veut » d'une condition initiale peut changer complètement le comportement asymptotique de la suite récurrente. Plus précisèment, si l'on note :
  - P l'ensemble des conditions initiales $z_0 \in U$ telles que la suite récurrente (z_n) définie par φ soit *périodique* ;
  - pour tout $p \in \mathbb{N}$, tel que $n \geqslant 2$, P_p l'ensemble des conditions initiales $z_0 \in U$ telles que (z_n) soit *périodique*, de plus petite période p, à partir d'un certain rang ;
  - P_1 l'ensemble des conditions initiales $z_0 \in U$ telles que (z_n) soit *stationnaire* ;
  - D l'ensemble des conditions initiales $z_0 \in U$ telles que (z_n) soit *dense* dans U ;

  alors les ensembles P, P_p $(p \in \mathbb{N}^*)$ et D sont *denses* dans U.

  Par suite, pour tout $z_0 \in U$ et tout $\varepsilon > 0$, l'ensemble $\{z \in U \mid |u_0 - z_0| < \varepsilon\}$ contient un point (et même une infinité de points) de chacun des ensembles P, D_p $(p \in \mathbb{N}^*)$ et D.

On admet le résultat suivant, dû, en 1964, au mathématicien ukrainien Sarkovsky[20] (une preuve simple fait l'objet de l'exercice IV.1.64). Il nous permettra de construire facilement des exemples d'applications continues admettant des cycles de toutes les périodes.

Proposition 52. Soient I un *segment* de $\mathbb{R}$ et $f : I \to I$ une application *continue*. On suppose que f admet un 3-cycle, alors, pour tout $\tau \in \mathbb{N}^*$, $T \geqslant 2$, f admet un τ-cycle.

Attention : la proposition ci-dessus n'est pas vraie en général si I n'est pas un segment (*cf.* l'exercice 35 de la page ci-contre).

Exemple 4. On reprend l'application tente $T : [0,1] \to [0,1]$ (*cf.* l'exemple de la page 529). Nous avons vu qu'elle admet deux 3-cycles. Elle admet donc des cycles de toutes les périodes.
On vérifie que le graphe de $T^{\circ n}$ est obtenu par la juxtaposition de 2^n « tentes » de hauteur 1 et de base de longueur $1/2^n$ (*cf.* la figure IV.1.1 de la page 530). On en déduit que $T^{\circ n}$ possède 2^n points fixes. Plus précisément la première bissectrice rencontre chaque tente en deux points (distincts), donc chaque segment obtenu en divisant $[0,1]$ en 2^n parties égales contient deux points fixes de $T^{\circ n}$. On en déduit que la réunion des images des τ-cycles de T, pour $\tau \in \mathbb{N}^*$, $\tau \geqslant 2$, est dense dans $[0,1]$. Les dérivées aux points fixes de $T^{\circ n}$ sont de module 2^n, ces points fixes sont donc *répulsifs* et les cycles de T sont tous *répulsifs*. Par suite les suites récurrentes convergentes définies par $T^{\circ n}$ sont nécessairement stationnaires.
L'application logistique f_4 et l'application quadratique $g_2 : \begin{array}{ccc} [-2,2] & \longrightarrow & [-2,2] \\ x & \longmapsto & x^2 - 2 \end{array}$ sont conjuguées à T (*cf.* les exemples de la page 569), elles admettent donc aussi des cycles de toutes les périodes et la réunion des images des τ-cycles de f_4 (resp. g_2), pour $\tau \in \mathbb{N}^*$, $\tau \geqslant 2$, est dense dans $[0,1]$ (resp. $[-2,2]$).

[20] Ce résultat a été retrouvé en 1975 par Li et Yorke, qui l'ont popularisé avec le slogan « la période 3 entraîne le chaos ».

2.5 Limites infinies — formes indéterminées

2.5.1 Limites infinies

Définition 27. On dit qu'une suite réelle (u_n) *tend vers l'infini* si $(1/u_n)$ est définie à partir d'un certain rang, et a pour limite zéro. Si de plus la suite est positive à partir d'un certain rang, on dit que la suite tend vers $+\infty$; si elle est négative à partir d'un certain rang, on dit qu'elle tend vers $-\infty$.

Dans ces deux derniers cas, on écrit par abus de notation $\lim u_n = +\infty$ ou $\lim u_n = -\infty$, mais on prendra garde que $\pm\infty$ ne sont pas des réels, donc ne sont pas à proprement parler des limites : la suite (u_n) reste divergente.
On peut donner une définition similaire pour une suite complexe, mais dans ce cas le signe ne veut rien dire. Une suite complexe (u_n) tend vers l'infini si, et seulement si, la suite réelle $(|u_n|)$ tend vers l'infini.
Le résultat suivant est très utile.

Proposition 53. Pour une suite réelle (u_n) on a l'équivalence :

$$(u_n) \text{ tend vers } +\infty \quad \Longleftrightarrow \quad \forall A > 0, \exists N \in \mathbb{N}, \forall n \geqslant N\,,\ u_n \geqslant A.$$

Il existe une propriété similaire pour une limite $-\infty$.

Démonstration. Pour une suite réelle, une autre façon de dire que (u_n) tend vers $+\infty$ est d'écrire en même temps que $(1/u_n)$ est positive (à partir d'un certain rang) et qu'elle tend vers zéro, donc :

$$\forall \varepsilon > 0\,,\ \exists N \in \mathbb{N}\ :\ \forall n \geqslant N\,,\ 0 < (1/u_n) \leqslant \varepsilon.$$

Mais la dernière inégalité équivaut à $u_n \geqslant 1/\varepsilon$ (tenir compte du signe !). Le nombre $1/\varepsilon$ est aussi grand que l'on veut à condition de prendre ε suffisamment petit. En appliquant cela avec $\varepsilon := 1/A$, pour n'importe quel réel $A > 0$, nous avons donc :

$$(u_n) \text{ tend vers } +\infty \Longleftrightarrow \forall A > 0, \exists N \in \mathbb{N}, \forall n \geqslant N\,,\ u_n \geqslant A.$$

■

2.5.2 Opérations et formes indéterminées

L'intérêt des « limites infinies » est de permettre certaines opérations similaires à celles vues dans le théorème 33 de la page 546.

Théorème 54. Soient (u_n) une suite réelle tendant vers $+\infty$ et (v_n) une suite réelle bornée. Alors $(u_n + v_n)$ et $(u_n - v_n)$ tendent vers $+\infty$. Le même résultat vaut si (u_n) tend vers $-\infty$, les limites sont alors $-\infty$.

Naturellement on en déduit aussi la limite de $(v_n - u_n)$ en changeant le signe. Pour le produit et quotient, une propriété similaire ne peut être vraie (considérer $u_n = n$ et $v_n = a/n$). Mais l'on a :

Théorème 55. Soient (u_n) une suite tendant vers $+\infty$ et (v_n) une suite bornée ; alors (v_n/u_n) tend vers zéro. Si (v_n) est convergente, de limite non nulle, alors $(u_n v_n)$ et (u_n/v_n) tendent vers l'infini, avec le signe de la limite de (v_n).

La première assertion n'est qu'une réécriture de la proposition 32 de la page 546, car $(1/u_n)$ tend vers zéro.

Démonstration. Les deux dernières assertions expriment en fait la même chose, car $(1/v_n)$ a aussi une limite non nulle. Il suffit donc de faire la preuve pour $(u_n v_n)$. De plus, il suffit de traiter le cas où la limite ℓ de (v_n) est strictement positive, l'autre cas s'en déduit en considérant $(-v_n)$.

Comme $\ell > 0$ nous pouvons utiliser $\varepsilon = \ell/2$ dans la définition (5) de la convergence de (v_n) ; ainsi il existe $N_0 > 0$ tel que, pour tout $n \geqslant N_0$, on ait $|v_n - \ell| \leqslant \ell/2$, ce qui implique notamment $v_n \geqslant \ell/2$.

Alors $u_n v_n \geqslant \frac{\ell}{2} u_n$. Soit $A > 0$ donné. Comme (u_n) tend vers l'infini, il existe $P \in \mathbb{N}$ tel que $u_n \geqslant \frac{2A}{\ell}$. Soit $N := \max(P, N_0)$; si $n \geqslant N$, alors $u_n v_n \geqslant \frac{2A}{\ell}\frac{\ell}{2} = A$. On a ainsi montré que $(u_n v_n)$ tend vers $+\infty$. ∎

On voit dans la démonstration qu'il suffit en fait que (v_n) soit minorée par un nombre strictement positif, ou majorée par un nombre strictement négatif.

Pour des opérations additives sur des suites convergentes ou tendant vers $\pm\infty$, les différents cas évoqués ci-avant ainsi que quelques autres faciles à montrer, peuvent être résumés dans un tableau :

	$u_n \to -\infty$	u_n bornée	$u_n \to +\infty$
$v_n \to -\infty$	$u_n + v_n \to -\infty$ $u_n - v_n$?	$u_n + v_n \to -\infty$ $u_n - v_n \to +\infty$	$u_n + v_n$? $u_n - v_n \to +\infty$
v_n bornée	$u_n + v_n \to -\infty$ $u_n - v_n \to -\infty$	$u_n + v_n$ bornée $u_n - v_n$ bornée	$u_n + v_n \to +\infty$ $u_n - v_n \to +\infty$
$v_n \to +\infty$	$u_n + v_n$? $u_n - v_n \to -\infty$	$u_n + v_n \to +\infty$ $u_n - v_n \to -\infty$	$u_n + v_n \to +\infty$ $u_n - v_n$?

Les « ? » correspondent tous à la même « forme indéterminée » : $\infty - \infty$. La différence de deux suites tendant vers l'infini peut avoir n'importe quel comportement : bornée ou non, convergente ou non, etc. Le lecteur est invité à trouver des exemples des différents cas possibles.

Pour le produit et le quotient la situation est plus complexe, en raison des signes. Nous donnons dans un tableau le cas du produit, celui du quotient se déduit en notant que (u_n/v_n) est le produit de (u_n) par $(1/v_n)$, et que cette dernière tend vers zéro si (v_n) tend vers l'infini, et réciproquement. Dans tous les cas faire attention aux signes !

$\lim u_n v_n$	$u_n \to -\infty$	$u_n \to a < 0$	$u_n \to 0$	$u_n \to a > 0$	$u_n \to +\infty$
$v_n \to -\infty$	$+\infty$	$+\infty$	?	$-\infty$	$-\infty$
$v_n \to b < 0$	$+\infty$	ab	0	ab	$-\infty$
$v_n \to 0$	?	0	0	0	?
$v_n \to b > 0$	$-\infty$	ab	0	ab	$+\infty$
$v_n \to +\infty$	$-\infty$	$-\infty$	?	$+\infty$	$+\infty$

Ici aussi nous avons une forme indéterminée, à savoir « $0.\infty$ ».

Pour le quotient, la même forme s'écrit « ∞/∞ » ou « $0/0$ ».
Terminons avec ce qui arrive pour $u_n^{v_n}$. Notons que, sauf dans les cas très particuliers où la suite v_n est constamment entière, cela n'est qu'une abréviation pour $\exp(v_n \ln u_n)$ et n'a donc de sens que si $u_n > 0$. Les règles se déduisent du cas du produit et des propriétés de l'exponentielle et du logarithme :

$\lim u_n^{v_n}$	$u_n \overset{>}{\to} 0$	$u_n \to a \in]0,1[$	$u_n \to 1$	$u_n \to a > 1$	$u_n \to +\infty$
$v_n \to -\infty$	$+\infty$	$+\infty$	?	0	0
$v_n \to b < 0$	$+\infty$	a^b	1	a^b	0
$v_n \to 0$	?	1	1	1	?
$v_n \to b > 0$	0	a^b	1	a^b	$+\infty$
$v_n \to +\infty$	0	0	?	$+\infty$	$+\infty$

Ici les formes indéterminées sont « 0^0 », « ∞^0 » et « 1^∞ ».
Comment résoudre les cas où l'on trouve une forme indéterminée ? Il n'y a pas de méthode systématique, mais certaines règles peuvent aider. En voici quelques-unes.

Cas $\infty - \infty$

Souvent les deux termes ne vont pas à l'infini « au même rythme » : il suffit alors de factoriser le terme dominant (typiquement la plus grande puissance de n). Par exemple $n^2 - n = n^2(1 - 1/n) \to (+\infty) \cdot (1 - 0) = +\infty$. Lorsque cette méthode échoue, il faut soit se montrer plus précis en utilisant par exemple un développement limité, soit utiliser quelques factorisations plus habiles. Un cas classique est celui de la différence de deux racines carrées, où l'on peut multiplier et diviser par la « quantité conjuguée », c'est-à-dire la somme des deux racines, et ensuite utiliser l'identité $(\sqrt{a} - \sqrt{b})(\sqrt{a} + \sqrt{b}) = a - b$ en espérant une simplification :

$$\begin{aligned}\sqrt{n^2+1} - \sqrt{n^2-1} &= (\sqrt{n^2+1} - \sqrt{n^2-1})\frac{\sqrt{n^2+1} + \sqrt{n^2-1}}{\sqrt{n^2+1} + \sqrt{n^2-1}} \\ &= \frac{n^2+1-(n^2-1)}{\sqrt{n^2+1} + \sqrt{n^2-1}} \to \frac{2}{\infty + \infty} = 0.\end{aligned}$$

Cas $0.\infty$, ∞/∞, $0/0$

Il faut factoriser le terme dominant dans chaque partie du produit ou quotient, et simplifier :

$$\frac{2n^4 - n^2 + n - 5}{\sqrt{n^8 - 6}} = \frac{n^4\left(2 - \frac{1}{n^2} + \frac{1}{n^3} - \frac{5}{n^4}\right)}{n^4\sqrt{1 - \frac{6}{n^8}}} = \frac{2 - \frac{1}{n^2} + \frac{1}{n^3} - \frac{5}{n^4}}{\sqrt{1 - \frac{6}{n^8}}} \to \frac{2}{\sqrt{1}} = 2.$$

Cas 0^0, ∞^∞, 1^∞

Prendre le logarithme de l'expression, ce qui ramène aux formes indéterminées précédentes. Ne pas oublier de reprendre l'exponentielle du résultat ! Exemple :

$$\left(1 + \frac{\alpha}{n}\right)^n = \exp\left[n \ln\left(1 + \frac{\alpha}{n}\right)\right] \to \exp(\alpha) = e^\alpha.$$

3 Un exemple de construction de $\mathbb{R}$

Cette partie est destinée au lecteur souhaitant approfondir les choses et elle nécessite un certain effort d'abstraction. Elle peut être omise sans inconvénient pour la lecture du reste de l'ouvrage. La lecture de la partie 3.1 peut toutefois être utile en relation avec la partie 2.3 du module IV.4 (page 727).

Comme nous l'avons dit plus haut, dès l'antiquité grecque les mathématiciens ont constaté (et prouvé) que les rationnels n'étaient pas suffisants pour mesurer certaines longueurs apparaissant dans des figures géométriques simples. Dans leur langage ils savaient que la diagonale d'un carré ou d'un cube n'est pas commensurable au côté[21] et que la diagonale d'un pentagone régulier, l'emblème de la secte des Pythagoriciens, n'est pas non plus commensurable au côté[22]. Pour des raisons « géométriques », on avait donc besoin de « nouveaux nombres », dits irrationnels, comme $\sqrt{2}$, $\sqrt{3}$, $\frac{1+\sqrt{5}}{2}$ ou π. Au XVIIe siècle avec l'apparition du calcul infinitésimal, l'introduction d'autres nombres irrationnels comme e ou $\ln 2$ devint nécessaire. Comme ceux de l'antiquité, les mathématiciens de l'époque ne se posaient pas la question de formaliser les notions, se contentant, pour chacun de ces nombres, de savoir l'écrire « sous forme décimale », avec un nombre aussi grand que voulu de chiffres après la virgule, en utilisant diverses formules, comme, par exemple, la formule de Machin [23] pour π : $\frac{\pi}{4} = 4\arctan\frac{1}{5} - \arctan\frac{1}{239}$. Ce n'est qu'à fin du XIXe siècle que les mathématiciens se sont proposés de donner un statut formel rigoureux à la notion de nombre irrationnel. Une des premières réponses est due au mathématicien allemand Richard Dedekind en 1872 [24] avec sa notion de *coupures*. L'idée de départ est de formaliser la notion de « trous » dans $\mathbb{Q}$, que nous avons évoquée plus haut dans l'introduction de ce module, par la donnée d'une *partition* de $\mathbb{Q}$ en deux sous-ensembles A_1 et A_2 (non vides) tels que pour tout $a_1 \in A_1$ et tout $a_2 \in A_2$, on ait $a_1 < a_2$. Si A_1 admet un plus grand élément ou si A_2 admet un plus petit élément, on peut définir (A_1, A_2) à partir de cet élément (de deux façons) et « il n'y a rien de nouveau » (par exemple $A_1 := \{x \in \mathbb{Q} \mid x \leqslant 3/2\}$ et $A_2 := \{x \in \mathbb{Q} \mid x > 3/2\}$). Sinon, la donnée de (A_1, A_2), la *coupure*, manifeste l'existence d'un « trou » dans $\mathbb{Q}$, comme nous l'avons vu avec l'exemple $A_1 = \{x \in \mathbb{Q} \mid x^2 < 2\}$ et $A_2 = \{x \in \mathbb{Q} \mid x^2 > 2\}$. L'idée de Dedekind est de construire les nombres réels à partir de l'ensemble des coupures, autrement dit « d'ajouter tous les trous » aux rationnels pour construire $\mathbb{R}$, en décidant qu'une coupure qui n'est pas définie par un rationnel *est* un irrationnel. Plus tard la construction de Dedekind a été un peu simplifiée en remplaçant la notion de coupure par celle de *section commençante*.En fait, en 1869, un peu avant Dedekind, le mathématicien français Charles Meray avait proposé une construction rigoureuse de $\mathbb{R}$ basée sur une idée tout à fait différente : la notion de *suite de Cauchy* de rationnels[25], que nous définirons dans le livre L2. L'idée de Meray a été redécouverte un peu plus tard (vers 1872) par le mathématicien allemand Georg Cantor.

Nous avons choisi de présenter une construction de $\mathbb{R}$ basée sur la notion de *nombre décimal*. Ce n'est pas la plus simple, ni la plus élégante, mais elle est conforme à l'intuition que l'on a de ces nombres, et à leur pratique dans l'enseignement primaire et secondaire en mathématiques et dans les sciences expérimentales. Tout au long de la construction, nous mettrons en évidence la différence entre les définitions et constructions théoriques (basées sur la théorie des ensembles) et les calculs effectifs (c'est-à-dire programmables) en l'illustrant par des exemples simples.

[21] Cf. le livre V des *Élements* d'Euclide

[22] Ce rapport est le *nombre d'or* ; il figure chez Euclide qui l'utilise pour une construction géométrique du pentagone.

[23] Découverte en 1706 par le mathématicien anglais John Machin, cette formule lui a permis de calculer 100 décimales de π, *cf.* l'exercice IV.3.14 de la page 709.

[24] Cf. *Stetigkeit und irrationale Zahlen* : *Continuité et nombres irrationnels*.

[25] Son article *Remarques sur la nature des quantités définies par la condition de servir de limites à des variables données* est passé assez inaperçu à l'époque.

Nous insisterons, un peu plus que strictement nécessaire pour la construction de $\mathbb{R}$, sur le maniement des réels sous forme décimale. Au delà de la construction proprement dite, il est utile pour de nombreuses applications de maîtriser ce maniement. Pour des raisons évidentes nous avons choisi d'utiliser l'écriture en base 10. Mais l'on peut construire $\mathbb{R}$ en reprenant, *mutatis mutandis*, la construction présentée en utilisant la numération dans une base arbitraire b (*cf.* le théorème 31 de la page 81). Le cas $b = 2$ est particulièrement intéressant en relation avec l'informatique et le maniement algorithmique des réels (*cf.* la *représentation flottante* des réels dans le livre L2).

Convention de lecture. Jusqu'à la fin de cette partie 3, le lecteur doit « oublier » que l'on a admis l'existence d'un ensemble vérifiant les axiomes des nombres réels et « oublier » également la « définition » de $\mathbb{R}$ que nous en avons déduit (*cf.* page 514) . « Repartant de zéro », nous allons construire, à partir de la seule donnée de $\mathbb{Q}$ et de sa structure de corps totalement ordonné, un ensemble $\mathbb{R}$, appelé l'ensemble des nombres réels, puis le munir d'une structure de corps totalement ordonné « contenant $\mathbb{Q}$ comme sous-corps » et vérifier qu'il satisfait la propriété de la borne supérieure. Certains points sont délicats et tout doit être fait avec beaucoup de soin.

3.1 Écritures décimales

3.1.1 Nombres décimaux et écritures décimales

Notre point de départ est l'écriture décimale d'un entier (*cf.* le théorème 31 de la page 81). On montre, en utilisant la division euclidienne, que tout $n \in \mathbb{N}^*$ s'écrit, de façon unique : $n = a_k 10^k + a_{k-1}10^{k-1} + \cdots + a_0$, avec $a_0, \ldots, a_k \in [\![0, 9]\!]$, $a_k \neq 0$. Par définition la chaîne de caractères (ou le « mot ») $a_k a_{k-1} \ldots a_0$ est l'écriture décimale de n et l'écriture décimale de 0 est la chaîne 0. En pratique on identifie le plus souvent un entier n et son écriture décimale $a_k a_{k-1} \ldots a_0$, ce que nous ferons dans la suite.

On appelle *nombre décimal* un rationnel qui peut s'écrire sous la forme $n/10^k$, avec $n \in \mathbb{Z}$, $k \in \mathbb{N}$. En particulier un entier est un nombre décimal.

Un rationnel τ est un nombre décimal si, et seulement s'il s'écrit sous forme réduite (c'est-à-dire telle que les entiers p et q sont premiers entre eux et $q > 0$) $\tau = \frac{p}{q}$, où la décomposition en facteurs premiers de q est de la forme $q = 2^a 5^b$ (a, $b \in \mathbb{N}$). En effet un décimal est de cette forme et si, inversement $r = \frac{p}{2^a 5^b}$, on a $r = \frac{2^b 5^a p}{10^{a+b}}$. Il résulte, par exemple, de ce critère que $\frac{1}{3}$ n'est pas un nombre décimal.

On montre facilement le résultat suivant.

Proposition 56. L'ensemble $\mathbb{D}$ des nombres décimaux est un *sous-anneau* de $\mathbb{Q}$.

Si $d \in \mathbb{D}$ et $k \in \mathbb{N}$, $10^{-k} d \in \mathbb{D}$: c'est le produit de d et du décimal 10^{-k}.

Muni de la relation d'ordre induite par celle de $\mathbb{Q}$, $\mathbb{D}$ est un *anneau ordonné*.

L'inverse d'un nombre décimal n'est pas *en général* un nombre décimal.

On peut caractériser facilement les éléments inversibles de $\mathbb{D}$. Un nombre décimal de la forme $\pm 2^m 5^n$ est inversible dans $\mathbb{D}$: $(\pm 2^m 5^n)\left(\pm \frac{2^n 5^n}{10^{m+n}}\right) = 1$. On montre qu'inversement tout élément inversible est de cette forme (*cf.* l'exercice IV.1.54).

Rappelons qu'un nombre décimal se note aussi sous la forme, dite décimale : $\pm n_0{,}\alpha_1\alpha_2 \ldots \alpha_k$, avec $n_0 \in \mathbb{N}$ (n_0 écrit sous-forme décimale), $\alpha_i \in \{0, \ldots, 9\}$. Le signe $\pm$ est ici soit $+$ (nombres positifs) soit $-$ (nombres négatifs), n_0 s'appelle la *partie principale* et les α_i les *décimales* du nombre.

Lorsque le nombre est positif ou nul, la partie principale s'appelle également *partie entière*.

Pour les décimaux positifs, on note le plus souvent $n_0{,}\alpha_1\alpha_2 \ldots \alpha_k$ au lieu de $+n_0{,}\alpha_1\alpha_2 \ldots \alpha_k$.

On a par exemple $1/2 = 5/10^1 = 0{,}5$; $-31789/1000 = -31{,}789$, etc. Nous avons tous appris à l'école primaire comment manipuler les décimaux sous forme décimale (comparaison, addition, multiplication).
Regardons plus en détail ce que signifie la notation $n_0{,}\alpha_1\alpha_2\ldots\alpha_k$ (nous nous limitons aux nombres positifs pour simplifier). C'est en fait une notation abrégée :

$$n_0{,}\alpha_1\alpha_2\ldots\alpha_k \overset{\text{déf}}{=} n_0 + \frac{\alpha_1}{10} + \frac{\alpha_2}{100} + \cdots + \frac{\alpha_k}{10^k} = n_0 + \sum_{i=1}^{k} \frac{\alpha_i}{10^i}.$$

Par exemple $12{,}008 = 12 + 8/1000$. En utilisant l'égalité ci-dessus, on montre que l'on peut ajouter un nombre (fini) arbitraire de zéros après la dernière décimale α_k sans changer la valeur du nombre : $n_0{,}\alpha_1\alpha_2\ldots\alpha_k 0\ldots 0 = n_0{,}\alpha_1\alpha_2\ldots\alpha_k$. On a aussi $-0,0\ldots 0 = 0,0\ldots 0 = 0$.
Nous allons construire $\mathbb{R}$ en utilisant des « écritures décimales infinies », comme, par exemple, $0,3333\ldots$, qui, comme on le devine, correspondra au rationnel $1/3$: on obtient $0,3333\ldots$ en divisant 1 par 3 par l'algorithme de division appris à l'école primaire et « en ne s'arrêtant pas ». On définit ces écritures infinies de la façon suivante.

Définition 28. On appelle écriture décimale toute « suite » de la forme $\pm n_0{,}\alpha_1\alpha_2\ldots\alpha_s\ldots$, où $\pm$ est $+$ ou $-$, $n_0 \in \mathbb{N}$, et $\alpha_i \in [\![0,9]\!]$ pour tout $i \in \mathbb{N}^*$.

Plus formellement, une écriture décimale est un élément de l'ensemble $\{\pm\} \times \mathbb{N} \times [\![0,9]\!]^{\mathbb{N}}$. On peut convenir que $n_0 \in \mathbb{N}$ est écrit en écriture décimale.
Par définition, on obtient l'écriture décimale $dec(\delta)$, ou développement décimal, d'un nombre décimal δ en complétant la représentation décimale usuelle avec un nombre *infini* de zéros.
L'application dec de $\mathbb{D}$ dans l'ensemble des écritures décimales est injective. On peut donc *identifier* $\mathbb{D}$ à son image dans l'ensemble des écritures décimales en « décidant » que $\delta = dec(\delta)$, ce que nous ferons dans la suite.
Le lecteur notera toutefois que cette identification n'est pas neutre du point de vue « pratique ». La connaissance de $\delta := n/10^k$ permet d'écrire $dec(\delta)$. Mais inversement, si l'on connaît *seulement le développement infini* $dec(\delta)$, il n'y a pas en général de moyen algorithmique de retrouver la fraction $n/10^k$. Si l'on voit apparaître des zéros dans le développement (donné par exemple par un programme), on ne sait pas en général si les décimales suivantes seront toutes nulles ou pas.
Nous verrons plus loin que, parmi les rationnels, les nombres décimaux sont un peu particuliers, car (sauf le nombre zéro) ils admettent *une autre* représentation décimale, avec un nombre infini de 9. Par exemple, 1 admet les deux représentations décimales : $1{,}0000\ldots$ et $0{,}9999\ldots$, 0 admet les deux représentations décimales : $0{,}0000\ldots$ et $-0{,}0000\ldots$. Cela représente un inconvénient pour la suite (mineur au niveau théorique mais « redoutable » au plan algorithmique...), ce qui justifie l'introduction de la terminologie suivante :

Définition 29. On dit qu'une écriture décimale est *impropre* si elle comprend des décimales toutes égales à 9 à partir d'un certain rang (donc $\exists n \in \mathbb{N}^* \ : \ \forall i > n\,,\ \alpha_i = 9$), ou bien si elle est de la forme $-0{,}0000\ldots$.
Dans le cas contraire, on parle d'écriture décimale *propre*

La notion de *troncature* est très utile pour manipuler les écritures décimales.

Définition 30. Pour tout $k \in \mathbb{N}^*$, on appelle troncature à l'ordre k (ou à la $k^{\text{ème}}$ décimale) d'une écriture décimale $x := \pm n_0{,}\alpha_1\alpha_2\ldots\alpha_s\ldots$ le *décimal* $x_k := \pm n_0{,}\alpha_1\alpha_2\ldots\alpha_k$. On convient de définir la troncature à l'ordre 0 par : $x_0 := \pm n_0 \in Z$.

Si $x := \pm n_0{,}\alpha_1\alpha_2\ldots\alpha_l\ldots$, on a $x_n = \pm n_0 + \sum_{i=1}^{n} \frac{\alpha_i}{10^i}$. On prendra garde à ne pas confondre les indices des termes d'une suite et les indices de troncature. Pour la troncature à l'ordre k du $n^{ème}$ terme x_n d'une suite, on notera en général $(x_n)_k$.

Dans la suite nous utiliserons la notion de *limite dans* $\mathbb{Q}$. La définition de cette notion est similaire à celle de limite dans $\mathbb{R}$ donnée plus haut dans la définition 19 (que nous ne pouvons pas utiliser ici puisque nous avons « oublié » $\mathbb{R}$...). Il faut simplement supposer que les u_n, ℓ et ε sont des *rationnels*. On prouve sans difficulté, mutatis mutandis, les propriétés élémentaires des limites dans $\mathbb{Q}$: unicité, somme, inégalités ... Nous laissons les détails au lecteur.

Exercice 37.

1. On note, pour tout $h \in \mathbb{N}^*$, $u_h := 0{,}\overbrace{99\ldots9}^{h \text{ répétitions}}$. Montrer que $u_h < 1$ et que la suite $(u_h)_{h\in\mathbb{N}^*}$ converge, dans $\mathbb{Q}$, vers 1.

2. Soient $d := 0{,}\alpha_1, \alpha_2\ldots\alpha_s\ldots$ une écriture décimale et $\tau \in \mathbb{Q}$. On suppose que τ est la limite *dans* $\mathbb{Q}$ de la suite des troncatures de d : $(d_h = 0, \alpha_1\ldots\alpha_h)_{h\in\mathbb{N}}$. Montrer que $\tau \leqslant 1$. Montrer que si d est *propre*, alors $\tau < 1$.

Solution.

1. On a, d'après les propriétés des progressions géométriques (*cf.* la proposition 23) :
$$u_h = \sum_{i=1}^{h} \frac{9}{10^i} = 9\sum_{i=1}^{h} \frac{1}{10^i} = \frac{9}{10}\frac{1-10^{-h}}{1-\frac{1}{10}} = 1 - 10^{-h} < 1.$$
La limite, dans $\mathbb{Q}$, de la suite $(u_h)_{h\in\mathbb{N}}$ est évidemment 1.

2. On a, pour tout $h \in \mathbb{N}$, $d_h \leqslant 0{,}\overbrace{99\ldots9}^{h \text{ répétitions}} = u_h$. D'où, en utilisant (i) et en passant à la limite, $\tau \leqslant 1$.
Si l'on suppose que d est *propre*, il existe $k \in \mathbb{N}^*$ tel que $\alpha_k \neq 9$. On a, pour tout $l \in \mathbb{N}^*$, $d_{k+l} = d_k + 10^{-k}v_l$, où $v_l := 0, \alpha_{k+1}\alpha_{k+2}\ldots\alpha_{k+l}$. On en déduit que la suite (v_l) admet une limite ω dans $\mathbb{Q}$ et que $\tau = d_k + 10^{-k}\omega$. On a (en utilisant $\alpha_k < 9$) :
$$d_k < u_k \leqslant 1 - 10^{-k}, \quad \text{et} \quad \omega \leqslant 1,$$
d'où $\tau < 1 - 10^{-k} + 10^{-k} = 1$.

3.1.2 Représentation décimale d'un nombre rationnel.

Parmi les écritures décimales, certaines sont déjà familières au lecteur (au moins dans certains cas particuliers). Ce sont celles associées aux nombres rationnels. Nous allons revenir sur ces écritures et les caractériser par une propriété simple parmi *toutes* les écritures décimales.

Soit $(p, q) \in \mathbb{N} \times \mathbb{N}^*$. L'algorithme élémentaire de division de l'entier p par l'entier q, poursuivi indéfiniment, permet d'associer à (p, q) une écriture décimale $+n_0{,}\beta_1\beta_2\ldots\beta_s\ldots$.
On a un résultat plus précis.

Proposition et définition 57. Soit $(p,q) \in \mathbb{N} \times \mathbb{N}^*$. L'algorithme élémentaire de division de l'entier p par l'entier q, poursuivi indéfiniment, permet d'associer à (p,q) une écriture décimale de la forme :

$$+n_0{,}\beta_1\beta_2\ldots\beta_k\underline{\alpha_1\ldots\alpha_m}\,\alpha_1\ldots\alpha_m\ldots, \tag{10}$$

où le groupe de m décimales $\alpha_1\ldots\alpha_m$ est répété indéfiniment. On a, de plus, $m < q$. L'écriture décimale (10) ne dépend que du *rationnel* p/q. On l'appelle représentation décimale (ou développement décimal) du rationnel p/q et on la note $dec(p/q)$.

Une écriture décimale de la forme (10) est dite *périodique à partir d'un certain rang* (ici $k+1$) ; le groupe de décimales $\alpha_1\ldots\alpha_m$ est la *période*, m est la longueur de cette période.
On peut aussi définir une représentation décimale associée à (p,q) quand $p \in -\mathbb{N}^*$: on change $+$ en $-$ dans celle associée à $(-p,q)$.

Démonstration. La première étape de l'algorithme élémentaire de division de p par q « est » la division euclidienne $p = qn_0 + r_0$, $0 \leqslant r_0 < q$. Ensuite, on « abaisse un zéro » et l'on fait une nouvelle division euclidienne, ce qui donne $10r_0 = q\beta_1 + r_1$, avec $0 \leqslant r_1 < q$ et $\beta_1 \in [\![0,9]\!]$. On poursuit de façon évidente : $10r_s = q\beta_{s+1} + r_{s+1}$. On obtient une écriture décimale $n_0{,}\beta_1\beta_2\ldots\beta_s\ldots$.
Les restes successifs r_s, $s \in \mathbb{N}^*$, appartiennent à l'ensemble *fini* $[\![0, q-1]\!]$. Ils ne peuvent donc pas être tous distincts. Il existe donc nécessairement $k,\ m \in \mathbb{N}^*$ tels que $r_k = r_{k+m-1}$ et l'algorithme se répète cycliquement à partir de r_k. Le résultat s'en déduit. En notant $\alpha_1 := \beta_{k+1}, \ldots, \alpha_m := \beta_{k+m}$, on obtient, pour tout $l \in \mathbb{N}$, $\beta_{k+lm+1} = \alpha_1, \ldots, \beta_{k+lm+m} = \alpha_m$. On a $0 \leqslant m < q$.
Supposons p et q premiers entre eux. Si $\frac{p'}{q'} = \frac{p}{q}$, il existe $u \in \mathbb{N}^*$ tel que $(p',q') = (up,uq)$. Si l'on remplace (p,q) par (up,uq), dans l'algorithme de division les restes successifs r_s sont remplacés par ur_s et n_0 et les quotients β_s ne changent pas. L'écriture décimale (10) ne change donc pas. ■

Exemples.

1. $p := 1$, $q := 3$. L'écriture décimale associée est : $0{,}333\ldots$ $(k=0,\ m=1)$;
2. $p := 829$, $q := 70$.
   L'écriture décimale associée est : $11{,}8\underline{428571}428571428571\ldots$ $(k=1,\ m=6)$.

Exercice 38.

Soit $x := +n_0{,}\beta_1\beta_2\ldots\beta_h\ldots$ une écriture décimale.

(i) Montrer que x est périodique, de période m, à partir du rang $k+1$ si, et seulement si, pour tout $h > k$, on a, $\beta_{h+m} = \beta_h$.

(ii) On suppose que x est périodique, de période m, à partir du rang $k+1$. Montrer que, pour tout $k' \geqslant k$, x est périodique de période m, à partir du rang $k'+1$ et comparer les deux suites finies $(\beta_{k+1},\ldots,\beta_{k+m})$ et $(\beta_{k'+1},\ldots,\beta_{k'+m})$.

(iii) On suppose que x est périodique, de période m, à partir du rang $k+1$ et qu'il existe $k' \geqslant k$ tel que $\beta_{k'+1} = \cdots = \beta_{k'+m} = 0$. Montrer que $x \in \mathbb{D}$ et qu'il s'écrit $x := +n_0{,}\beta_1\beta_2\ldots\beta_k$.

Solution.

(i) C'est évident en utilisant la division euclidienne de h par m.

(ii) On suppose que x est périodique, de période m, à partir du rang $k+1$ et que $k' \geqslant k$. En utilisant (i), on montre que x est périodique, de période m, à partir du rang $k'+1$. On vérifie que les deux suites finies se déduisent l'une de l'autre par permutation circulaire.

(iii) D'après (ii), on a $\beta_{k+1} = \cdots = \beta_{k+m} = 0$, donc $\beta_h = 0$ pour tout $h > k$. Ainsi $x \in \mathbb{D}$ et il s'écrit $x := +n_0{,}\beta_1\beta_2\ldots\beta_k$.

La représentation décimale d'un nombre décimal est son écriture ordinaire suivie d'une infinité de zéros. On rappelle que, pour tout $d \in \mathbb{D}$, on a identifié d à $dec(d)$.
Revenant à la description de l'algorithme de division, on peut écrire, pour tout $s \in \mathbb{N}^*$, $r_s = q\frac{\beta_{s+1}}{10} + \frac{r_{s+1}}{10}$ et :

$$p = qn_0 + r_0 = qn_0 + q\frac{\beta_1}{10} + \frac{r_1}{10} = qn_0 + q\frac{\beta_1}{10} + q\frac{\beta_2}{100} + \frac{r_2}{100} = qn_0 + q\frac{\beta_1}{10} + q\frac{\beta_2}{100}\cdots + q\frac{\beta_s}{10^s} + \frac{r_s}{10^s}\cdot$$

$$p = qn_0 + q\left(\frac{\beta_1}{10} + \frac{\beta_2}{100}\cdots + \frac{\beta_s}{10^s}\right) + \frac{r_s}{10^s} = q(n_0, \beta_1\ldots\beta_s) + \frac{r_s}{10^s}\cdot \tag{11}$$

Proposition 58.
Soient $\tau := \frac{p}{q} \in \mathbb{Q}^+$ et $dec(\tau) = n_0, \beta_1\ldots\beta_s\ldots$ sa représentation décimale :

(i) τ est la borne supérieure *dans l'ensemble ordonné* $\mathbb{Q}$ du sous-ensemble A de décimaux défini par $A := \{n_0, \beta_1\ldots\beta_h\}_{h\in\mathbb{N}^*}$;

(ii) τ est la limite, *dans* $\mathbb{Q}$, de la suite croissante de nombres décimaux ; $(n_0, \beta_1\ldots\beta_h)_{h\in\mathbb{N}^*}$.

(iii) l'écriture décimale $dec(\tau)$ est *propre*.

Démonstration. D'après (11), on a, pour tout $h \in \mathbb{N}^*$, une égalité *dans* $\mathbb{Q}$:

$$n_0, \beta_1\ldots\beta_h = \frac{p}{q} - \frac{r_h}{q}10^{-h}.$$

Puisque $0 \leqslant \frac{r_h}{q} < 1$, on en déduit :

$$\frac{p}{q} - 10^{-h} < n_0, \beta_1\ldots\beta_h \leqslant \frac{p}{q}\cdot \tag{12}$$

Les assertions (i) et (ii) en résultent. Montrons (iii).
On raisonne par l'absurde. Supposons que $\tau \neq 0$ et que son écriture est impropre. Alors, elle s'écrit $dec(\tau) = d = n_0{,}\beta_1\beta_2\ldots\beta_k 99\ldots9\ldots$. D'après (ii), τ est la limite de la suite $(d_h)_{h\in\mathbb{N}^*}$, où, si $h > k$: $d_h := n_0{,}\beta_1\beta_2\ldots\beta_k\overbrace{99\ldots9}^{h-k \text{ répétitions}}$. On a, en calculant dans $\mathbb{D}$: $d_h = n_0{,}\beta_1\beta_2\ldots\beta_k + \dfrac{0{,}\overbrace{99\ldots9}^{h-k \text{ répétitions}}}{10^k}$ et la limite de la suite $(d_h)_{h\in\mathbb{N}^*}$ est $n_0{,}\beta_1\beta_2\ldots\beta_k + 10^{-k}$, puisque la limite *dans* $\mathbb{Q}$ de la suite de décimaux $(0{,}\overbrace{99\ldots9}^{h-k \text{ répétitions}})_{h\in\mathbb{N}^*}$ est 1 (*cf.* l'exercice 37). Ainsi $\tau = n_0{,}\beta_1\beta_2\ldots\beta_k + 10^{-k}$ est un *décimal* et son développement décimal est propre, d'où une contradiction. ∎

Proposition 59. Soient $\tau \in \mathbb{Q}$, et $m_0, \alpha_1, \alpha_2 \ldots \alpha_s \ldots$, $n_0, \beta_1, \beta_2, \ldots \beta_s \ldots$ deux écritures décimales *propres*. On suppose que τ est la limite *dans* $\mathbb{Q}$ de la suite $(m_0, \alpha_1 \ldots \alpha_h)_{h \in \mathbb{N}^*}$ et aussi de la suite $(n_0, \beta_1 \ldots \beta_h)_{h \in \mathbb{N}^*}$. Alors les deux écritures sont égales :

$$m_0, \alpha_1, \alpha_2 \ldots \alpha_s \ldots = n_0, \beta_1, \beta_2, \ldots \beta_s \ldots.$$

La proposition ci-dessus est fausse si l'on ne suppose plus que les deux écritures sont *propres* (*cf.* l'exercice 37 de la page 584).

Démonstration. On raisonne par l'absurde en supposant que les deux écritures décimales sont différentes.
Supposons d'abord que $n_0 < m_0$. L'écriture décimale $n_0, \beta_1, \beta_2, \ldots \beta_s \ldots$ est *propre*, donc, d'après l'exercice 37, la limite ω, dans $\mathbb{Q}$, de la suite $(0, \beta_1 \ldots \beta_h)_{h \in \mathbb{N}^*}$ est *strictement* inférieure à 1 et $\tau = n_0 + \omega < n_0 + 1 \leqslant m_0 \leqslant \tau$, d'où une contradiction.
On montrerait de même que l'on ne peut pas avoir $m_0 < n_0$.
On a donc $m_0 = n_0$. Il suffit de traiter le cas $m_0 = n_0 = 0$. Il existe $k \in \mathbb{N}^*$ tel que $\alpha_k \neq \beta_k$. On peut supposer que k est le plus petit entier possible possédant cette propriété. On suppose que $\beta_k < \alpha_k$ (le cas $\alpha_k < \beta_k$ est similaire). La suite $(0, \beta_{k+1} \ldots \beta_{k+h})_{h \in \mathbb{N}^*}$ converge, dans $\mathbb{Q}$, vers un rationnel ω tel que $\tau = 0, \beta_1 \ldots \beta_k + \omega$. On a $0 \leqslant \omega < 1$, donc

$$\tau < 0, \beta_1 \ldots \beta_k + 10^{-k} \leqslant 0, \beta_1 \ldots \alpha_k = 0, \alpha_1 \ldots \alpha_k \leqslant \tau,$$

et l'on obtient une contradiction. ∎

Le résultat suivant permet de caractériser les écritures décimales qui sont représentations décimales de rationnels. Nous n'utiliserons pas ce résultat pour la construction de $\mathbb{R}$ et sa preuve peut être omise en première lecture.

Proposition 60. Soit une écriture décimale *propre* périodique à partir d'un certain rang :

$$+n_0{,}\beta_1\beta_2 \ldots \beta_k \underline{\alpha_1 \ldots \alpha_m}\, \alpha_1 \ldots \alpha_m \ldots.$$

Alors elle est la représentation décimale d'un rationnel.

Démonstration.
Considérons d'abord le cas particulier de l'écriture décimale $u = 0{,}\underline{\alpha_1 \ldots \alpha_m}\, \alpha_1 \ldots \alpha_m \ldots$.
On définit, pour tout $n \in \mathbb{N}^*$:

$$\begin{aligned} a_n := 0{,}\overbrace{\underline{\alpha_1 \ldots \alpha_m}\, \alpha_1 \ldots \alpha_m \ldots \alpha_1 \ldots \alpha_m}^{n \text{ répétitions}} &= \sum_{i=1}^{m} \frac{\alpha_i}{10^i} + \sum_{i=1}^{m} \frac{\alpha_i}{10^{i+m}} + \cdots + \sum_{i=1}^{m} \frac{\alpha_i}{10^{i+m(n-1)}} \\ &= \left(\sum_{i=1}^{m} \frac{\alpha_i}{10^i} \right) \cdot \left(1 + \frac{1}{10^m} + \cdots + \frac{1}{10^{m(n-1)}} \right). \end{aligned}$$

Le décimal a_n est donc de la forme $a_1(1 + x + x^2 + \cdots + x^{n-1})$ avec $x = 1/10^m = 10^{-m}$. En utilisant $1 + x + \cdots + x^{n-1} = (1 - x^n)/(1 - x)$, on obtient :

$$a_n = b - 10^{-mn} b \leqslant b \qquad \text{où} \qquad b := \frac{a_1}{1 - 10^{-m}}.$$

On en déduit que la suite croissante de décimaux $(a_n)_{n \in \mathbb{N}^*}$ converge, *dans* $\mathbb{Q}$, vers le *rationnel* b, puis que la suite $(u_h)_{h \in \mathbb{N}}$ converge aussi, *dans* $\mathbb{Q}$, vers le *rationnel* b. Alors, d'après la proposition 59, $u = dec(b)$.

Considérons maintenant le cas général.
Soit l'écriture décimale $u := n_0{,}\beta_1\beta_2\ldots\beta_k\underline{\alpha_1\ldots\alpha_m}\,\alpha_1\ldots\alpha_m\ldots$. On définit, pour tout $n \in \mathbb{N}^*$:

$$a'_n := n_0{,}\beta_1\beta_2\ldots\beta_k\overbrace{\underline{\alpha_1\ldots\alpha_m}\,\alpha_1\ldots\alpha_m\ldots\,\alpha_1\ldots\alpha_m}^{n \text{ répétitions}}.$$

On a $a'_n = u_{k+mn} = n_0{,}\beta_1\beta_2\ldots\beta_k + 10^{-k}a_n$. On en déduit que la suite croissante de décimaux $(a'_n)_{n\in\mathbb{N}^*}$ converge, *dans* $\mathbb{Q}$, vers le *rationnel* :

$$b' := n_0{,}\beta_1\beta_2\ldots\beta_k + 10^{-k}b = n_0{,}\beta_1\beta_2\ldots\beta_k + 10^{-k}\frac{a_1}{1-10^{-m}}. \tag{13}$$

Notons $m_0{,}\gamma_1\gamma_2\ldots\gamma_s\ldots := dec(b')$. On vérifie facilement que la limite de la suite croissante de décimaux $(n_0, \beta_1\ldots\beta_h)_{h\in\mathbb{N}^*}$ est b'. D'après la proposition 58, celle de la suite $(m_0, \gamma_1\ldots\gamma_h)_{h\in\mathbb{N}^*}$ est aussi b'. Alors, d'après la proposition 59, $u = dec(b')$. ■

L'algorithme de divisions successives qui permet de calculer $dec(\tau)$ à partir de $\tau = \frac{p}{q}$ est facile à programmer. nous engageons le lecteur à le faire à titre d'exercice. On peut inversement se demander comment *calculer* un rationnel τ à partir de la donnée d'une écriture décimale u dont on sait *abstraitement* qu'elle est périodique à partir d'un certain rang. C'est facile, on vient de le voir, si l'on connaît, *en plus de* u, le point de départ $k+1$ de la partie périodique et la longueur m de la période : on utilise (13). Par contre *si l'on ne connaît pas* k *et* m, il est en général *impossible* de les calculer à partir de la seule connaissance de u (par exemple quand u est engendré par un programme). Ainsi, au plan *algorithmique*, la donnée d'un rationnel τ et celle du développement $u := d(\tau)$ ne sont pas équivalentes.

Exemple. Soit $u := 2,135\underline{74}7474\ldots$ $(k = 3,\ m = 2)$. On a : $\tau := \frac{p}{q} = \frac{2135}{1000} + \frac{1}{1000}\frac{74}{99} = \frac{211439}{99000}$ et $dec(\tau) = u$.

3.2 Définition des nombres réels à partir des écritures décimales

3.2.1 Définition de $\mathbb{R}$

Nous disposons maintenant des éléments nécessaires pour donner une définition de *l'ensemble* $\mathbb{R}$ des nombres réels. Nous le munirons ensuite d'une structure d'ensemble totalement ordonné satisfaisant la propriété de la borne supérieure, puis d'une structure de corps totalement ordonné. Nous vérifierons aussi que $\mathbb{R}$ « contient » $\mathbb{Q}$ comme sous-corps.

Définition 31. On appelle *nombre réel* toute écriture décimale *propre* : $\pm n_0{,}\alpha_1\alpha_2\ldots\alpha_k\ldots$, où $\pm$ est $+$ ou $-$, $n_0 \in \mathbb{N}$, et $\alpha_i \in [\![0, 9]\!]$ pour tout $i \in \mathbb{N}^*$.
On note $\mathbb{R}$ l'ensemble des nombres réels.

Si les décimales d'un réel sont nulles à partir d'un certain rang, c'est un décimal. Si toutes les décimales sont nulles, c'est un entier (égal à $\pm n_0$). Le plus souvent, on continuera à écrire les nombres entiers sous leur forme classique et non leur développement décimal (par exemple on écrira 0 et non $+0{,}0000\ldots$). On écrira souvent $n_0{,}\alpha_1\alpha_2\ldots$ au lieu de $+n_0{,}\alpha_1\alpha_2\ldots$.
Notons, qu'au plan algorithmique, la vérification de *l'égalité* de deux réels (dont par exemple les décimales successives sont données par deux programmes) pose un problème : il faudrait tester *toutes* les décimales, c'est-à-dire une infinité.
La proposition suivante nous permettra *plus tard* d'identifier $\mathbb{Q}$ à un sous-ensemble de $\mathbb{R}$. Nous ne le ferons qu'en fin de construction pour éviter les confusions.

Proposition 61. L'application $dec : \tau \in \mathbb{Q} \to dec(\tau) \in \mathbb{R}$ qui à un rationnel associe son développement décimal est une injection.

Nous appellerons cette application dec *l'injection naturelle de* $\mathbb{Q}$ *dans* $\mathbb{R}$. Un réel qui n'appartient pas à l'image de dec est dit *irrationnel*.

Démonstration. Ceci résulte de l'assertion (ii) de la proposition 58. Si $dec(\tau) = dec(\tau')$, les suites associées sont les mêmes, elles ont la même limite et $\tau = \tau'$. ■

Les rationnels correspondent aux écritures décimales périodiques à partir d'un certain rang. « Boucher les trous » de $\mathbb{Q}$ revient alors à considérer *toutes* les écritures décimales au lieu de se limiter à celles qui sont périodiques à partir d'un certain rang.

Compte tenu de ce qui se passe pour les rationnels, on est tenté d'interpréter un nombre réel $x := \pm n_0{,}\alpha_1\alpha_2\ldots\alpha_s\ldots$ comme la limite de la suite de décimaux $x_n := (\pm n_0{,}\alpha_1\alpha_2\ldots\alpha_k)_{k\in\mathbb{N}}$. Malheureusement, au point où nous en sommes, nous ne disposons pas de structures sur $\mathbb{R}$ permettant de définir la notion de limite et de prouver les propriétés des limites. Cette interprétation ne sera possible que lorsque nous aurons suffisamment avancé dans la construction.

Pour l'instant nous ne savons pas s'il existe un irrationnel[26]. On en donne un exemple dans l'exercice suivant.

Exercice 39.

On définit une écriture décimale $x = 0,\alpha_1\alpha_2\ldots\alpha_h\ldots$ par :

- $\alpha_k = 1$ s'il existe $n \in \mathbb{N}$ tel que $k = n!$,
- $\alpha_k = 0$ sinon.

Montrer que, pour tout $n \in \mathbb{N}^*$, la troncature $x_{n!}$ de x à l'ordre $n!$ vérifie :

$$x_{n!} := 0,\alpha_1\alpha_2\ldots\alpha_{n!} = \sum_{k=1}^{n} \frac{1}{10^{k!}}.$$

Montrer que l'écriture décimale x est propre et que le nombre réel x est irrationnel.

Solution. On doit avoir $x_1 = 0,\alpha_1 = 0,1$, $x_2 = 0,\alpha_1\alpha_2 = 0,11$, $x_3 = 0,\alpha_1\alpha_2\alpha_3 = 0,110$, $x_6 = 0,\alpha_1\alpha_2\ldots\alpha_6 = 0,110001$, $x_{24} = 0,\alpha_1\alpha_2\ldots\alpha_{24} = 0,110001000000000000000001\ldots$

L'écriture décimale x vérifie clairement $x_{n!} := 0,\alpha_1\alpha_2\ldots\alpha_{n!} = \sum\limits_{k=1}^{n} 10^{-k!}$. Les décimales de x sont égales à 0 ou 1 donc jamais égales à 9 et l'écriture est propre. Pour tout $k \in \mathbb{N}$, il existe $n \in \mathbb{N}$ tel que $n! > k$, donc $\alpha_{n!} = 1 \neq 0$; le réel x n'est donc pas un décimal.

Montrons, par l'absurde, que x est irrationnel. Supposons qu'il existe un rationnel τ tel que $dec(\tau) = x$. Alors x est périodique à partir d'un certain rang $k+1$. Notons m la longueur de la période ($m \in \mathbb{N}^*$ et $m \geqslant 2$).

Quand $n \to +\infty$, $n! - (n-1)! = (n-1)(n-1)!$ tend vers $+\infty$. Il existe donc $n \in \mathbb{N}^*$, tel que $k < (n-1)!$ et $m < n! - (n-1)!$. Alors $(n-1)! + m < n!$, donc les m décimales de x qui suivent la $(n-1)!^{ème}$ sont toutes nulles : $\alpha_{(n-1)!+1} = \alpha_{(n-1)!+2} = \cdots = \alpha_{(n-1)!+m} = 0$. En utilisant l'exercice 38, on en déduit que x est un décimal. On a donc une contradiction

[26] Rappelons que nous avons convenu « d'oublier » que $\sqrt{2}$ est irrationel !

Le nombre $x = 0,110001000000000000000010000\ldots$ est appelé un *nombre de Liouville* . Le mathématicien français Joseph Liouville a montré en 1844 que ce nombre est en fait *transcendant* (pour une preuve, *cf.* la module II.6, page 314). C'est le premier exemple historique de nombre transcendant[27].

On a pour le moment un ensemble $\mathbb{R}$ qui est un ensemble « en vrac », sans aucune structure existante. Nous allons ajouter les structures que l'on souhaite avoir sur $\mathbb{R}$, à savoir : la relation d'ordre, les opérations $+$ et $\times$ et vérifier leurs propriétés. Nous allons commencer par la relation d'ordre.

3.2.2 Relation d'ordre sur $\mathbb{R}$

Définitions et premières propriétés

Nous savons comment comparer deux nombres décimaux *positifs* écrits en représentation décimale : on regarde la partie principale (devant la virgule), puis éventuellement les décimales successives. Autrement dit, on utilise l'ordre lexicographique (comme dans un dictionnaire) en considérant une écriture décimale finie comme un « mot ». Sur $\mathbb{N}$, l'ordre est l'ordre ordinaire. Sur $[\![0,9]\!]$ l'ordre est l'ordre ordinaire. Pour comparer deux nombres décimaux négatifs x et y, on compare $-x$ et $-y$: $x \leqslant y$ si, et seulement si, $-y \leqslant -x$. Enfin un nombre décimal négatif est inférieur à un nombre décimal positif.
On va prolonger (de façon « naturelle ») cette règle aux écritures décimales infinies.

La comparaison à zéro revient simplement à examiner le signe :

Définition 32. On dit qu'un réel $x = \pm n_0{,}\alpha_1\alpha_2\ldots$ est positif ou nul (et l'on écrit $x \geqslant 0$) si $\pm$ est $+$; et l'on dit qu'il est strictement négatif (et l'on écrit $x < 0$) dans le cas contraire.
On note $-x$ le nombre réel égal à $\mp n_0{,}\alpha_1\alpha_2\ldots$ si $x \neq 0$, ou à zéro si $x = 0$ (ici $\mp$ signifie le signe opposé à $\pm$).
On dit que x est strictement positif (et l'on écrit $x > 0$) si $-x < 0$; dans le cas contraire on le dit négatif ou nul (et l'on écrit $x \leqslant 0$).

On note : $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x \geqslant 0\}$ et $\mathbb{R}_- := \{x \in \mathbb{R} \mid x \leqslant 0\}$.

Définition 33. Soient deux réels $x = \pm m_0{,}\alpha_1\alpha_2\ldots$ et $y = \pm n_0{,}\beta_1\beta_2\ldots$, non nuls. On dit que $x < y$ si l'un des cas suivants se présente :

1. ou bien $x < 0$ et $y > 0$;
2. ou bien $0 < x$ et $0 < y$ et $m_0 < n_0$;
3. ou bien $0 < x$ et $0 < y$ et $m_0 = n_0$ et il existe $k \in \mathbb{N}^*$ tel que $\alpha_i = \beta_i$ pour tout $i < k$, et $\alpha_k < \beta_k$;
4. ou bien $x < 0$, $y < 0$ et $-y < -x$ par les règles précédentes.

On dit que $x > y$ si $-x < -y$; que $x \leqslant y$ si $x < y$ ou $x = y$; que $x \geqslant y$ si $x > y$ ou $x = y$.

Ainsi, pour deux réels positifs, $m_0{,}\alpha_1\alpha_2\ldots > n_0{,}\beta_1\beta_2\ldots$ équivaut à dire que, ou bien $m_0 > n_0$ (partie entière plus grande), ou bien $m_0 = n_0$ mais il existe un indice entier $k \geqslant 1$ tel que $\alpha_k > \beta_k$ (décimale strictement plus grande), mais $\alpha_i = \beta_i$ pour $i \in [\![1, k-1]\!]$ (les décimales précédentes sont identiques).
On a $x \leqslant y$ si, et seulement si, pour tout $h \in \mathbb{N}$, on a l'inégalité $x_h \leqslant y_h$ sur les troncatures à l'ordre h. On a, pour tout $h \in \mathbb{N}$, $x_h \leqslant x$.

[27] Les preuves de la transcendance de e, par Hermite, et de π, par Lindemann, sont postérieures : en 1873 et 1882.

On vérifie facilement que $\leqslant$ est une relation d'ordre sur $\mathbb{R}$.

Exemples.

1. On a : $0 \leqslant 0,\alpha_1\alpha_2\ldots < 1$.
2. Soit $m \in \mathbb{N}^*$. On a : $0,0\ldots0\,\alpha_{m+1}\alpha_{m+2}\ldots < 10^{-m}$. Soit, en effet, le nombre décimal $0,0\ldots010\ldots0\ldots$, où 1 est la $m^{\text{ème}}$ décimale et où les autres décimales sont nulles ; on a : $0,0\ldots0\,\alpha_{m+1}\alpha_{m+2}\ldots < 0,0\ldots010\ldots0\ldots = 10^{-m}$.
   Inversement, si $0,\alpha_1\alpha_2\ldots\alpha_k\ldots < 10^{-m}$, on a $\alpha_1 = \alpha_2 = \ldots\alpha_m = 0$.
   Sinon, il existerait $k \in [\![1,m]\!]$, tel que $\alpha_k \neq 0$ et l'on aurait :
   $$10^{-m} \leqslant 10^{-k} \leqslant \alpha_k 10^{-k} = 0,00\ldots\alpha_k 0\ldots \leqslant 0,\alpha_1\alpha_2\ldots\alpha_k\alpha_{k+1}\ldots < 10^{-m},$$
   d'où une contradiction.

Dans la définition 33 ci-dessus on a épuisé tous les cas possibles, donc :

Proposition 62.
Deux réels sont toujours comparables, c'est-à-dire que, pour tous $x,\ y \in \mathbb{R}$, $x \leqslant y$ ou $x \geqslant y$.

Ainsi $(\mathbb{R},\leqslant)$ est un ensemble *totalement ordonné*.
Une conséquence importante de ce résultat est que toute partie *finie* non vide de $\mathbb{R}$ possède un plus grand et un plus petit élément. La démonstration est immédiate par récurrence sur le nombre d'éléments.

Proposition 63. Sur $\mathbb{Q}$ la relation d'ordre « induite par celle de $\mathbb{R}$ » et la relation d'ordre ordinaire coïncident. C'est à dire que, pour tous $\tau,\ \tau' \in \mathbb{Q}$, on a $\tau \leqslant \tau'$ dans $\mathbb{Q}$ si, et seulement si, $dec(\tau) \leqslant dec(\tau')$ dans $\mathbb{R}$.

Démonstration. Soient $\tau,\ \tau' \in \mathbb{Q}$. Supposons que $dec(\tau) \leqslant dec(\tau')$ au sens de la relation d'ordre de $\mathbb{R}$. Pour tout $h \in \mathbb{N}^*$, on note τ_h et τ'_h les troncatures respectives à l'ordre h de $dec(\tau)$ et $dec(\tau')$. On a, pour tout $h \in \mathbb{N}^*$, $\tau_h \leqslant \tau'_h$. En utilisant la proposition 58 et en passant à la limite sur h *dans* $\mathbb{Q}$, on en déduit que $\tau \leqslant \tau'$ au sens de la relation d'ordre de $\mathbb{Q}$. Inversement, supposons $\tau \neq \tau'$. Si $\tau < \tau'$, dans $\mathbb{Q}$, on a $d(\tau) < d(\tau')$. Sinon, l'on aurait $dec(\tau') < dec(\tau)$ et $\tau' < \tau$, ce qui est impossible. Notons qu'*a priori* la relation d'ordre sur $\mathbb{Q}$ induite par celle de $\mathbb{R}$ n'en fait pas nécessairement un corps ordonné ; c'est vrai *a posteriori*. ■

Proposition 64. Pour tout réel x, il existe un entier $n \in \mathbb{Z}$ tel que $n > x$.

Démonstration. Si $x = m_0,\alpha_1\alpha_2\ldots \geqslant 0$, on vérifie facilement que $n := m_0 + 1$ convient. Si $x = -m_0,\alpha_1\alpha_2\ldots < 0$, alors $n := 0$ convient. ■

La propriété suivante montre qu'entre deux réels distincts, on peut en trouver un autre, et même un décimal.

Proposition 65. Si x et y sont deux réels tels que $x < y$, alors il existe un nombre réel, et même un nombre décimal z, tel que $x < z < y$. Il existe même une infinité de nombres décimaux vérifiant cette double inégalité.

Démonstration. Commençons par trouver un z adéquat. Si $x < 0 < y$, il suffit de choisir $z := 0$. Dans la suite, on examine le cas où x et y sont de même signe. Supposons par exemple que $x \geqslant 0$, l'autre cas étant similaire. Nous écrivons les développements décimaux propres de x et y comme dans la définition 33 de la page ci-contre.

Les décimales de x ne sont pas toutes égales à 9 à partir d'un certain rang (sinon on aurait un développement impropre) ni toutes égales à celles de y (sinon $x = y$). Soit k le rang de la première décimale de x telle que $\alpha_k \neq \beta_k$ (ou $k = 0$ si $m_0 \neq n_0$) et soit ℓ le premier entier *strictement* supérieur à k tel que $\alpha_\ell \neq 9$. Posons $\widetilde{\alpha}_\ell := 1 + \alpha_\ell$ et $z := m_0{,}\alpha_1 \dots \alpha_{\ell-1}\widetilde{\alpha}_\ell$ de sorte qu'il est évident que z est décimal et $z > x$. Mais on a aussi $z < y$. En effet, si $k = 0$, la partie entière de z est $m_0 < n_0$, et si $k > 0$, les k premières décimales de z sont les mêmes que celles de x, et qu'elles sont identiques à celles de y sauf la $k^{ème}$ qui est strictement plus grande ($\alpha_k < \beta_k$).
Maintenant que nous savons qu'il existe toujours un décimal entre deux réels distincts, appelons z_0 un premier décimal tel que $x < z_0 < y$. Puis appliquons la même propriété à x et z_0 : il existe donc un z_1 tel que $x < z_1 < z_0 < y$. On continue ainsi à construire par récurrence z_{k+1} décimal tel que $x < z_{k+1} < z_k < y$. Les décimaux construits sont tous distincts et compris entre x et y ; il y en a une infinité. ■

Exercice 40.
Soit $x \in \mathbb{R}$, $x > 0$. Montrer qu'il existe $k \in \mathbb{N}$ tel que $0 < 10^{-k} < x$.
Solution. D'après la proposition ci-dessus, il existe un nombre décimal $d := \frac{n}{10^k}$ ($n \in \mathbb{N}^*$, $k \in \mathbb{N}$) tel que $0 < d < x$. On a alors $0 < 10^{-k} \leqslant \frac{n}{10^k} < x$.

Non dénombrabilité de $\mathbb{R}$

Proposition 66. L'ensemble des nombres réels n'est pas dénombrable.

Démonstration. La preuve ci-dessous est une variante du célèbre *argument de la diagonale* dû au mathématicien allemand Georg Cantor (en 1891). On va définir un sous-ensemble A du segment $[0, 1]$ et montrer qu'il n'est pas dénombrable, *a fortiori* $\mathbb{R}$ ne le sera pas non plus.
Soit A l'ensemble des réels compris entre 0 et 1 dont le développement décimal propre ne comporte, après la virgule, que des 1 et des 8. On raisonne par l'absurde. Supposons que l'on sache énumérer l'ensemble (infini) A, c'est-à-dire qu'il existe une bijection $f : \mathbb{N} \to A$ et notons, pour tout $n \in \mathbb{N}$:

$$f(n) = 0, a_{n1}a_{n2} \dots a_{nk} \dots,$$

où a_{nk} est égal à 1 ou 8. Soit x le nombre réel dont le développement décimal s'écrit :

$$x := 0, a_{11}a_{22} \dots a_{kk} \dots.$$

Soit y le nombre réel dont le développement décimal est obtenu en remplaçant dans celui de x, pour tout $k \in \mathbb{N}$, a_{kk} par $9 - a_{kk}$. On a $y \in A$, il existe donc un entier $n \in \mathbb{N}$ tel que $f(n) = y$. Le $n^{ème}$ chiffre de $f(n)$ après la virgule est a_{nn} tandis que celui de y est $9 - a_{nn}$. On a donc une contradiction. ■

> ***Remarque***. On déduit immédiatement de la proposition qu'il existe au moins un nombre irrationnel, ce que nous avons montré plus haut en donnant un exemple explicite (*cf.* l'exercice 39 de la page 589). Il y a en fait « beaucoup » d'irrationnels, comme on le verra dans l'exercice suivant.

Exercice 41.
Montrer que l'ensemble des nombres réels irrationnels n'est pas dénombrable.
Solution. La réunion de deux ensembles dénombrables est dénombrable. Si $\mathbb{R}\backslash\mathbb{Q}$ était dénombrable, $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$ le serait, ce qui n'est pas.

3.3 Théorème de la borne supérieure

C'est *le* point fondamental de la construction de $\mathbb{R}$ à partir des écritures décimales. Il permet de voir que $\mathbb{R}$ « n'a pas de trous ». Nous admettons le théorème suivant. Nous en donnons une preuve dans le livre L3.

Théorème 67 (Borne supérieure).
Tout sous-ensemble non vide et majoré de $\mathbb{R}$ admet une borne supérieure.
De même, tout sous-ensemble non vide et minoré de $\mathbb{R}$ admet une borne inférieure.

Le résultat suivant permet de comparer les notions de borne supérieure dans $\mathbb{Q}$ et dans $\mathbb{R}$.

Corollaire 68. Soient $X \subset \mathbb{Q}$ et $\mathcal{X}$ son image dans $\mathbb{R}$, c'est-à-dire le sous-ensemble de $\mathbb{R}$ formé des développements décimaux des éléments de X. On suppose que X admet une borne supérieure β *dans* $\mathbb{Q}$. Alors $\mathcal{X}$ admet une borne supérieure b dans $\mathbb{R}$ et $b = dec(\beta)$.

Démonstration. L'ensemble X n'est pas vide et il est borné dans $\mathbb{Q}$: β est un majorant. Alors $\mathcal{X}$ n'est pas vide et il est majoré par $dec(\beta)$ dans $\mathbb{R}$. Par suite l'ensemble $\mathcal{X}$ admet une borne supérieure b dans $\mathbb{R}$ et l'on a $b \leqslant dec(\beta)$. Montrons $b = dec(\beta)$. On raisonne par l'absurde. Supposons que $b \neq dec(\beta)$. On a $b < dec(\beta)$. Il existe un *décimal* a tel que $b < a < dec(\beta)$. Ce décimal majore X, on a donc $\beta \leqslant a$ et $dec(\beta) \leqslant a$, d'où une contradiction. ∎

Remarque. Les notions de borne supérieure et de borne inférieure ont été définies à l'aide d'outils « théoriques », mais ces bornes ne sont pas en général « calculables » (au sens où l'on ne peut pas calculer leurs troncatures à un ordre arbitraire).

Voici un exemple. Soit $(x_n)_{n\in\mathbb{N}}$ une suite[28] de décimaux. On note $a \in \mathbb{R}$ la borne supérieure de l'image X de la suite. Alors la connaissance d'un nombre fini m, si grand soit-il, de termes de la suite ne permet pas en général de calculer la partie entière de a. Si, par exemple $x_n := 0,99\ldots9$ (où 9 figure n fois), la borne supérieure a de X est 1 dont la partie entière est 1, mais la connaisance *d'un seul* x_m permet seulement de savoir que la partie entière de a est 0 ou 1 sans aucun moyen de décider. On ne peut le faire que parce-que l'on connait *tous* les termes de la suite.

3.4 Opérations sur les réels

Nous savons comment additionner, soustraire, et multiplier des décimaux. Nous allons en déduire les définitions des opérations analogues sur les réels. Notons toutefois que *l'on ne peut pas donner une méthode directe*. En effet, l'addition de décimaux doit se faire en partant de la dernière décimale (pour propager les retenues). Mais dans le cas des réels, il n'y a pas, en général, de dernière décimale ! Évidemment, si l'on a $\pi = 3{,}141592654\ldots$ et $\mathrm{e} = 2{,}718281828\ldots$, on se doute que $\pi + \mathrm{e} = 5{,}859874482\ldots$, mais comment être sûr que les décimales omises ne modifient pas le résultat ?

On peut voir sur des exemples simples que la propagation des retenues « vers la gauche » peut apparaître sur une longueur arbitrairement longue. Par suite le calcul *exact* des premières décimales de la somme de deux nombres décimaux peut nécessiter de connaitre les troncatures à un ordre élevé (impossible en général à borner *a priori*) de leur développement décimal.

Exemple. Soit $x := 0,0541375$. Calculons la première décimale de $x + y$ et $x + y'$ pour $y := 0,0458622$ et pour $y' := 0,0458626$.

On a $x+y = 0,0999997$, la première décimale est donc 0. On a, par ailleurs, $x+y' = 0,1000001$, la première décimale est donc 1. Il faut connaître les 7 premières décimales de x et y pour calculer la première décimale de leur somme.

[28] Attention : n est un indice et ne désigne pas une troncature.

3.4.1 Addition des réels

Somme de deux rationnels sous forme d'écriture décimale

Pour « deviner » la définition de la somme de deux réels nous allons d'abord étudier ce qui se passe pour la somme de deux *rationnels* donnés par leurs *écritures décimales*.
Soient $\xi,\ \eta \in \mathbb{Q}$. On note $x := dec(\xi)$ et $y = dec(\eta)$. La somme $\zeta := \xi + \eta$ au sens de $\mathbb{Q}$ est sans mystère, mais nous voudrions trouver l'écriture décimale $z := dec(\zeta)$ « sans utiliser $\mathbb{Q}$ », c'est-à-dire à partir de la seule donnée des écritures décimales x et y. On va pour cela utiliser la proposition 58 de la page 586.
On note $x := m_0, \alpha_1, \alpha_2 \ldots \alpha_s \ldots$ et $y := n_0, \beta_1, \beta_2, \ldots \beta_s \ldots$ et l'on définit les deux ensembles de *décimaux* $X := \{x_h := m_0, \alpha_1 \ldots \alpha_h\}_{h \in \mathbb{N}^*}$ et $Y := \{y_h := n_0, \beta_1 \ldots \beta_h\}_{h \in \mathbb{N}^*}$. D'après la proposition 58, ξ est la borne supérieure, dans $\mathbb{Q}$, de l'ensemble de décimaux X et η est la borne supérieure, dans $\mathbb{Q}$, de l'ensemble de décimaux Y. On va voir qu'il est possible de retrouver le réel $z = dec(\zeta)$ à partir de l'ensemble de décimaux $A := \{x_h + y_h\}_{k \in \mathbb{N}^*}$, et, *a fortiori*, à partir des développements x et y.
Montrons que A admet une borne supérieure dans $\mathbb{Q}$ et qu'elle est égale à $\zeta = \xi + \eta$. On utilise la variante pour le corps archimédien $\mathbb{Q}$ du point méthode de la page 518. Le rationnel $\zeta = \xi + \eta$ est évidemment un majorant de A dans $\mathbb{Q}$. Montrons que, pour tout $n \in \mathbb{N}^*$, il existe $h \in \mathbb{N}^*$ tel que $\zeta - \frac{1}{n} < x_h + y_h \leqslant \zeta$.
Soit $n \in \mathbb{N}^*$. Il existe $h',\ h'' \in \mathbb{N}^*$ tels que $\xi - \frac{2}{n} < x_{h'} \leqslant \xi$ et $\eta - \frac{2}{n} < y_{h''} \leqslant \eta$. Soit $h := \max\{h',\ h''\}$. On a $\xi + \eta - \frac{1}{n} < x_h + y_h \leqslant \xi + \eta$. Ainsi $\xi + \eta = \zeta$ est la borne supérieure de A *dans* $\mathbb{Q}$. D'après le corollaire 68, $z := dec(\zeta)$ est la borne supérieure de A *dans* $\mathbb{R}$.

Exemple. Soient $x = dev(\frac{1}{3}) = 0,333333\ldots$ et $y = dev(\frac{2}{3}) = 0,666666\ldots$.
On a $x_h + y_h = 0,999999\ldots$ (où 9 figure h fois). La borne supérieure de $A := \{x_h + y_h\}_{k \in \mathbb{N}^*}$ est $1 = \frac{1}{3} + \frac{2}{3}$.

Somme de deux réels

On admet le résultat suivant (dont la preuve est donnée dans le livre L3).

Théorème et définition 69. Soient $x = m_0{,}\alpha_1\alpha_2\ldots$ et $y = n_0{,}\beta_1\beta_2\ldots$ deux réels positifs. Pour tout $k \in \mathbb{N}^*$, notons $x_k := m_0{,}\alpha_1 \ldots \alpha_k$ et $y_k := n_0{,}\beta_1 \ldots \beta_k$, qui sont des nombres décimaux. Alors l'ensemble :

$$A := \{x_1 + y_1, x_2 + y_2, \ldots\} = \{x_k + y_k \mid k \in \mathbb{N}^*\}$$

est non vide et majoré.
Sa borne supérieure (dans $\mathbb{R}$) est appelée la *somme* de x et y : on la note $x + y$.
Notons $\tilde{y}_k := y_k + 10^{-k}$; l'ensemble :

$$B := \{x_1 - \tilde{y}_1, x_2 - \tilde{y}_2, \ldots\} = \{x_k - \tilde{y}_k \mid k \in \mathbb{N}^*\}$$

est également non vide et majoré. Sa borne supérieure est appelée *différence* de x et y, on la note $x - y$.
Si x et y sont des décimaux la somme et la différence définies ci-dessus coïncident avec la somme et la différence usuelles.

Autrement dit, pour additionner, on *tronque* les nombres à la $k^{ème}$ décimale, on ajoute les décimaux obtenus, et finalement on prend le plus petit réel dépassant tous les résultats obtenus de cette façon.

Contrairement à ce qui se passe pour les nombres décimaux (et plus généralement pour les nombres rationnels), la somme de deux nombres réels n'est pas, en général « effective ». On la définit en utilisant le théorème de la borne supérieure (*cf.* plus loin la remarque de la page 595).

Notons que nous n'avons défini ci-dessus que la somme et la différence de réels positifs, mais on étend ces définitions à tous les réels en se ramenant à des nombres positifs avec les règles habituelles. Par exemple si $x > 0$, $y < 0$, alors $x + y := x - (-y)$ et $x - y := x + (-y)$; si $x < 0$, $y < 0$, alors $x+y := -\big((-x)+(-y)\big)$ et $x-y := (-y)-(-x)$, etc. Avec ces extensions, la soustraction n'apparaît plus que comme un cas particulier de l'addition, et en particulier on peut facilement montrer que $x - y = x + (-y)$ pour tous x, y.
Compte tenu de ce que nous avons vu plus haut et de la définition de l'addition dans $\mathbb{R}$, on a, pour tous ξ, $\eta \in \mathbb{Q}$, $dec(\xi+\eta) = dec(\xi)+dec(\eta)$. Autrement dit, si l'on identifie $\mathbb{Q}$ à un sous-ensemble de $\mathbb{R}$, l'addition de deux rationnels dans $\mathbb{Q}$ et dans $\mathbb{R}$ « donnent le même résultat ».
La plupart du temps, l'addition n'a rien de surprenant. Pourtant, si, comme dans l'exemple 3.4.1 ci-dessus l'on considère $x = dev(\frac{1}{3}) = 0,333333\ldots$ et $y = dev(\frac{2}{3}) = 0,666666\ldots$, on trouve que la somme $x + y$ est la borne supérieure de l'ensemble des nombres de la forme $0,999\ldots9$ (avec un nombre fini, mais arbitraire, de 9) et cette borne supérieure *n'est pas* $0,9999\ldots$, qui est une représentation décimale impropre, mais le nombre entier 1.

Exercice 42.
Soit $a = 0,\alpha_1\ldots\alpha_h\ldots \in \mathbb{R}$, $a \notin \mathbb{D}$. Pour tout $h \in \mathbb{N}$, on pose $\beta_h := 9 - \alpha_h$ et l'on définit une écriture décimale $b = 0,\beta_1\ldots\beta_h\ldots$ Montrer que b est propre et calculer $a + b$.
Solution. Si b n'était pas propre, a serait un décimal, ce qui est contraire à l'hypothèse.
On a, pour tout $h \in \mathbb{N}^*$, $a_h + b_h = 0,\underbrace{99\ldots9}_{h \text{ répétitions}}$.
D'après la définition de la somme : $a + b = sup_h\, 0,\overbrace{99\ldots9}^{h \text{ répétitions}} = 1$.

Remarque. La somme de deux réels est définie à l'aide d'une borne supérieure, par suite, comme la borne supérieure, la somme n'est pas en général « calculable » (*cf.* la remarque 3.3 de la page 593). Pour $k \in \mathbb{N}$ donné, on ne peut pas en général calculer la troncature $(x + y)_k$ de la somme à l'aide de la donnée des deux troncatures x_m et y_m si grand que soit m. Par exemple, si $x = dev(\frac{1}{3}) = 0,333333\ldots$ et $y = dev(\frac{2}{3}) = 0,666666\ldots$, la donnée *d'un seul* $x_m + y_m = 0,999999\ldots$ ne permet pas de calculer la partie entière de $x + y$: on ne peut pas choisir entre 0 et 1. Pour savoir que c'est 1, il faut connaître *toutes* les décimales de x et y.

Propriétés de l'addition.

Proposition 70. Pour tous réels x, y, z, on a $x + y = y + x$ et $x + (y + z) = (x + y) + z$.

On dit que l'addition est *commutative* et *associative*. La démonstration, facile, est laissée au lecteur : elle repose sur le fait que ces égalités sont déjà vraies pour les décimaux.
Notons aussi que $x + 0 = x$ pour tout x, ce qui évident d'après la définition. On dit que 0 est *élément neutre* de l'addition.
On a, en appliquant la définition de l'addition, $x + (-x) = 0$. Ainsi $-x$ est *l'opposé* de x. Par ailleurs :

$$x - x = \sup\{x_k - \tilde{x}_k\}_{k\in\mathbb{N}} = \sup\{x_k - x_k - 10^{-k}\}_{k\in\mathbb{N}} = \sup\{-10^{-k}\}_{k\in\mathbb{N}} = 0.$$

On peut énoncer le résultat suivant.

Proposition 71. L'ensemble $\mathbb{R}$ des réels muni de l'addition est un *groupe commutatif* et l'image de $\mathbb{Q}$ par dec est un sous-groupe.

Proposition 72. Soient x, y des réels tels que $x \leqslant y$.
Alors, pour tout $z \in \mathbb{R}$, on a $x + z \leqslant y + z$.

On dit que l'addition *préserve l'ordre*.

Démonstration. Examinons par exemple le cas où les trois nombres x, y, z sont positifs (les autres cas sont similaires). Notons x_k, y_k, z_k leurs troncatures à la $k^{ème}$ décimale. Comme $x \leqslant y$, on a $x_k \leqslant y_k$. Donc aussi $x_k + z_k \leqslant y_k + z_k$. Comme $y_k + z_k \leqslant y + z$ par définition de la somme, on a donc $x_k + z_k \leqslant y + z$: ainsi $y + z$ est un majorant de l'ensemble des $(x_k + z_k)$; il est donc plus grand que sa borne supérieure $x + z$. ∎

On déduit facilement de ce qui précède que si $x \leqslant y$ et $a \leqslant b$, alors $x + a \leqslant y + b$ (faire deux étapes). De plus, toutes ces propriétés sur les inégalités larges sont aussi vraies avec des inégalités strictes.
D'après l'écriture décimale $\pi = 3{,}141592654\ldots$ et $\mathrm{e} = 2{,}718281828\ldots$, on voit que $3{,}141592654 \leqslant \pi \leqslant 3{,}141592655$ et $2{,}718281828 \leqslant \mathrm{e} \leqslant 2{,}718281829$; donc par addition, on trouve que $5{,}859874482 \leqslant \pi + \mathrm{e} \leqslant 5{,}859874484$. Ainsi les premières décimales de $\pi + \mathrm{e}$ sont bien $5{,}85987448$, mais on n'est pas sûr de la suivante (il faudrait connaître des décimales supplémentaires dans les nombres initiaux).
C'est là un problème classique (et grave) en informatique, où les calculs sur les « nombres à virgule flottante » (c'est-à-dire des « décimaux » en base 2) impliquent des troncatures.
Nous disposons maintenant de structures sur $\mathbb{R}$ suffisantes pour définir la notion de limite dans $\mathbb{R}$ (*cf.* la partie 2.2.2 à la page 541) et prouver les propriétés élémentaires de ces limites.
En utilisant la définition et la notation des séries (*cf.* le module IV.4), on a le résultat suivant que nous admettrons.

Proposition 73. Soit $x := \pm n_0, \alpha_1\alpha_2\ldots\alpha_l\ldots \in \mathbb{R}$. Les suites (x_n) et $(\tilde{x}_n) = x_n + 10^{-n}$ sont adjacentes. Elles convergent dans $\mathbb{R}$ et ont pour limite x. On a : $x = \pm n_0 + \sum\limits_{n\in\mathbb{N}} \frac{\alpha_n}{10^n}$.

On reprend l'énoncé de la proposition 3 de la page 515. La preuve de cette proposition n'utilise que les propriétés de la relation d'ordre et de l'addition. Elle reste donc valable ici.

Proposition 74. Pour tout sous-ensemble majoré non-vide $A \subset \mathbb{R}$, $b \in \mathbb{R}$ est la borne supérieure de A si, et seulement si, b vérifie les deux conditions suivantes :

(i) b est un majorant de A ;

(ii) pour tout $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$, il existe un élément $x \in A$ tel que $b - \varepsilon < x \leqslant b$.

Le lecteur écrira l'énoncé analogue pour la borne inférieure.

3.4.2 Multiplication des réels. Inverses

Produit de réels

La multiplication de deux réels se définit de manière similaire à ce que nous avons fait pour l'addition et la soustraction. Il faut commencer par des nombres positifs. Les définitions s'étendent ensuite aux autres réels avec les règles de signes habituelles.

Théorème et définition 75. Soient x et y deux réels positifs. Pour tout $k \in \mathbb{N}^*$, notons x_k et y_k leurs troncatures respectives à l'ordre k. Alors l'ensemble de nombres rationnels :

$$C := \{x_1y_1, x_2y_2, \ldots\} = \{x_ky_k \mid k \in \mathbb{N}^*\}$$

est non vide et majoré. Sa borne supérieure est appelée le *produit* de x et y : on le note xy ou $x \times y$.

Démonstration. Soient n et m deux entiers tels que $n > x$ et $m > y$. Comme pour tout k, on a $x_k \leqslant x < n$, $y_k \leqslant y < m$, on a $x_ky_k < nm$ et nm est donc un majorant de C. ■

Montrons que pour tous $\xi,\ \eta \in \mathbb{Q}$, $dec(\xi\eta) = dec(\xi)dec(\eta)$. Autrement dit, que si l'on identifie $\mathbb{Q}$ à un sous-ensemble de $\mathbb{R}$, les produits respectifs de deux rationnels dans $\mathbb{Q}$ et dans $\mathbb{R}$ donnent le même résultat.
Il suffit de montrer le résultat pour $\xi, \eta > 0$. Montrons d'abord $\xi\eta = \sup\limits_{k\in\mathbb{N}} (dec(\xi))_k\,(dec(\eta))_k$, le sup étant pris dans $\mathbb{Q}$. On a $0 \leqslant (dec(\xi))_k \leqslant \xi$ et $0 \leqslant (dec(\eta))_k \leqslant \eta$, donc :

$$(dec(\xi\eta))_k \leqslant (dec(\xi))_k\,(dec(\xi))_k \leqslant \xi\eta.$$

D'après la proposition 58 de la page 586, on a $\xi\eta = \sup\limits_{k\in\mathbb{N}} (dec(\xi\eta))_k$ (le sup étant pris dans $\mathbb{Q}$). En utilisant les inégalités ci-dessus, on en déduit $\xi\eta = \sup\limits_{k\in\mathbb{N}} (dec(\xi))_k\,(dec(\eta))_k$. En utilisant le corollaire 68 et la définition de la multiplication de $\mathbb{R}$, on en déduit $dec(\xi\eta) = dec(\xi)dec(\eta)$.
En s'inspirant de ce qui a été fait pour l'addition, on vérifie sans difficulté que $(\mathbb{R}, +, \times, \leqslant)$ est un *anneau* commutatif totalement ordonné et que l'image de $\mathbb{Q}$ par dec est un sous-anneau.

Inverse d'un réel pour la multiplication

Il reste à montrer que tout élément non nul de $\mathbb{R}$ admet un inverse pour la multiplication. La définition de l'inverse est plus délicate que celle de l'addition et de la multiplication car l'inverse d'un décimal est un rationnel mais n'est pas, en général, un décimal.
Soit $x \in \mathbb{R}$, $x > 0$. Pour tout $k \in \mathbb{N}$, $\tilde{x}_k = x_k + 10^{-k} > 0$. On a $\tilde{x}_k \in \mathbb{Q}^*$, donc $\frac{1}{\tilde{x}_k} \in \mathbb{Q}$. Pour allèger la notation on identifie $\frac{1}{\tilde{x}_k} \in \mathbb{Q}$ à $dec\left(\frac{1}{\tilde{x}_k}\right) \in \mathbb{R}$. Nous admettrons le résultat suivant (dont la preuve est donnée dans le livre L3).

Théorème et définition 76. Soit $x \in \mathbb{R}$. Supposons $x > 0$ et notons, comme plus haut, $\tilde{x}_k := x_k + 10^{-k}$. Alors l'ensemble :

$$D := \{1/\tilde{x}_0, 1/\tilde{x}_1, \ldots\} = \left\{1/\tilde{x}_k \;\middle|\; k \in \mathbb{N}\right\}.$$

est non vide et majoré dans $\mathbb{R}$. Sa borne supérieure est appelée *l'inverse* de x, on le note $1/x$ ou $\frac{1}{x}$. On a $x\frac{1}{x} = 1$.

Si $x < 0$, on définit son inverse par $\frac{1}{x} := -\frac{1}{-x}$.

Fin de la construction de $\mathbb{R}$

En *identifiant* $\mathbb{Q}$ à son image dans $\mathbb{R}$ par l'injection naturelle dec, on obtient le théorème suivant qui termine notre construction de $\mathbb{R}$.

Théorème 77. L'ensemble $\mathbb{R}$ muni de la structure de corps totalement ordonné définie ci-dessus satisfait les axiomes des nombres réels.

Remarque.
Dans toutes les constructions « classiques » de $\mathbb{R}$ (décimaux, coupures, sections commençantes, suites de Cauchy de rationnels), on n'obtient pas en fait un corps *contenant* $\mathbb{Q}$ comme sous-corps. Il est nécessaire *d'identifier* $\mathbb{Q}$ à un sous-corps de $\mathbb{R}$ pour vérifier les axiomes des nombres réels. Ceci ne pose pas de problème grâce au résultat suivant que nous admettons ici.

Proposition 78. Soient $\mathcal{R}$ un corps satisfaisant les axiomes des nombres réels et $\mathcal{K}$ un corps totalement ordonné vérifiant la propriété de la borne supérieure. Alors existe un *unique isomorphisme* croissant d'anneaux de $\mathcal{R}$ dans $\mathcal{K}$.

Exercice type corrigé

IV.1.0 Soit (u_n) une suite réelle bornée. Pour tout $n \in \mathbb{N}$, on note :

$$x_n := \inf\{u_p \mid p \geqslant n\} \qquad \text{et} \qquad y_n := \sup\{u_p \mid p \geqslant n\}.$$

1. Montrer que les suites (x_n) et (y_n) sont bien définies, que (x_n) est croissante et (y_n) décroissante.
2. En déduire qu'elles convergent.
   On pose $\alpha := \lim\limits_{n\to+\infty} x_n$ et $\beta := \lim\limits_{n\to+\infty} y_n$. Montrer que $\alpha \leqslant \beta$.
3. On suppose $\alpha = \beta$. Montrer que (u_n) converge.
4. Montrer que si (u_n) admet une sous-suite convergeant vers un certain réel ℓ, alors $\ell \in [\alpha, \beta]$.
5. Montrer que pour tout $\varepsilon > 0$ et pour tout $n \in \mathbb{N}$, il existe $p \geqslant n$ tel que $y_n - \varepsilon \leqslant u_p \leqslant y_n$.
6. En déduire qu'il existe une sous-suite de (u_n) qui converge vers β.
   Quel théorème vient-on de redémontrer ?

Solution.

1. Pour tout $n \in \mathbb{N}$, notons $A_n := \{u_p \mid p \geqslant n\}$. Comme (u_n) est bornée, il existe $M \in \mathbb{R}_+$ tel que $\forall n \in \mathbb{N}, |u_n| \leqslant M$ et les ensembles A_n sont aussi bornés par M. Puisqu'ils sont évidemment non vides, ils admettent des bornes inférieures et supérieures comprises entre $-M$ et M.
   La borne inférieure x_n de A_n est un minorant de A_n donc de A_{n+1} puisque $A_{n+1} \subset A_n$. On en déduit $x_{n+1} \geqslant x_n$ puisque x_{n+1} est le plus grand des minorants de A_n. Ainsi la suite (x_n) est croissante. De le même façon, on montre que (y_n) est décroissante.
2. La suite (x_n) est alors croissante et majorée par M, donc converge. De même, (y_n), étant décroissante et minorée par $-M$, converge. De plus, $x_n = \inf A_n \leqslant \sup A_n = y_n$, donc par passage à la limite, $\alpha \leqslant \beta$.
3. Soit $n \in \mathbb{N}$. Comme $u_n \in A_n$, on a $x_n \leqslant u_n \leqslant y_n$. Donc si $\alpha = \beta$, le théorème des gendarmes montre que (u_n) converge également vers α.
4. Soit (u_{n_k}) une suite extraite de (u_n) convergeant vers un réel ℓ.
   On a $\forall k \in \mathbb{N},, \ x_{n_k} \leqslant u_{n_k} \leqslant y_{n_k}$ et comme ces trois suites convergent, on en déduit, par passage à la limite : $\alpha \leqslant \ell \leqslant \beta$.

5. Soient $\varepsilon > 0$ et $n \in \mathbb{N}$. Comme y_n est le plus petit des majorants de A_n, on en déduit que $y_n - \varepsilon$ n'est pas un majorant de A_n. Il existe alors dans A_n un élément, donc de la forme u_p avec $p \geqslant n$, supérieur à $y_n - \varepsilon$. De plus, y_n majore A_n donc u_p. On en déduit finalement $y_n - \varepsilon \leqslant u_p \leqslant y_n$.
6. Construisons par récurrence, à l'aide du résultat de la question précédente, une suite extraite u_{p_k} convergeant vers β. Posons $p_0 := 0$. Supposons construits $p_0 < p_1 \cdots < p_k$ tels que $\forall i \in [\![1, k]\!]\,,\ y_{1+p_{i-1}} - \frac{1}{i} \leqslant u_{p_i} \leqslant y_{1+p_{i-1}}$. Prenons $\varepsilon = 1/(k+1)$ et $n = 1 + p_k$. La question précédente nous donne l'existence d'un entier $p \geqslant n$ tel que :
$$y_{1+p_k} - \frac{1}{k+1} \leqslant u_p \leqslant y_{1+p_k}.$$
En notant p_{k+1} un tel p, on a alors $p_0 < p_1 \cdots < p_k < p_{k+1}$ et :
$$\forall i \in [\![1, k+1]\!]\,,\ y_{1+p_{i-1}} - \frac{1}{i} \leqslant u_{p_i} \leqslant y_{1+p_{i-1}}.$$
La suite extraite (u_{p_k}) converge donc vers β d'après le théorème des gendarmes puisque l'on a $\lim y_{1+p_{k-1}} = \beta$ (sous-suite d'une suite convergeant vers β) et $\lim \frac{1}{k} = 0$.
On retrouve ainsi le théorème de Bolzano-Weierstraß.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Axiomatique de $\mathbb{R}$

IV.1.1

1. Soient A et B deux parties non vides de $\mathbb{R}$ telles que B majore A, c'est-à-dire telles que tout élément de A soit inférieur ou égal à tout élément de B. Montrer qu'il existe un élément $x \in \mathbb{R}$ tel que $\forall a \in A\,,\ \forall b \in B\,,\ a \leqslant x \leqslant b$.
2. Réciproquement, montrer que tout corps totalement ordonné vérifiant cette propriété vérifie également la propriété de la borne supérieure.
3. Soit K un corps totalement ordonné archimédien vérifiant la *propriété des segments emboîtés*, que nous rappelons : si (a_n) et (b_n) sont deux suites de K telles que (a_n) soit croissante, (b_n) décroissante et que $\forall n \in \mathbb{N}\,,\ a_n \leqslant b_n$, alors il existe un élément $x \in K$ tel que $\forall n \in \mathbb{N}\,,\ a_n \leqslant x \leqslant b_n$.
Montrer que K vérifie la propriété de la borne supérieure.
On pourra procéder par dichotomie : si A est une partie non vide majorée, on considérera une suite (a_n) de A et une suite (b_n) de majorants de A telles que $b_{n+1} - a_{n+1} = (b_n - a_n)/2$.

Rationnels et irrationnels. Corps ordonnés

IV.1.2

1. Soient $p, q \in \mathbb{Q}$, strictement positifs. On suppose que $\sqrt{p}$ est irrationnel.
Montrer que $\sqrt{p} + \sqrt{q}$ est irrationnel.
2. Montrer que $\sqrt{2} + \sqrt{3} + \sqrt{5}$ est irrationnel.
3. Soient n un nombre premier et $m \in \mathbb{N}$, $m \geqslant 2$. Montrer que $\sqrt[m]{n}$ est irrationnel.

IV.1.3 Soit $n \in \mathbb{N}^*$. Donner une condition nécessaire et suffisante sur la décomposition en facteurs premiers de n pour que $\sqrt{n}$ soit irrationnel.

IV.1.4 Montrer que $\ln 3/\ln 2 = \log_2 3$ est irrationnel.
L'irrationalité de $\log_2 3$ est liée au problème de fabrication des gammes musicales étudié depuis Pythagore.

IV.1.5

1. Calculer $\cos 2\pi/5$ et montrer qu'il est irrationnel.
2. Montrer que $\cos 2\pi/7$ est solution d'une équation du troisième degré à coefficients entiers. En déduire que $\cos 2\pi/7$ est irrationnel.

IV.1.6

1. Montrer qu'il existe un corps archimédien totalement ordonné K tel que $\mathbb{Q} \subsetneq K \subsetneq \mathbb{R}$.
   On pourra par exemple considérer l'ensemble $\{a + b\sqrt{2} \in \mathbb{Q} \mid a, b \in \mathbb{Q}\}$.
2. Soit L un corps tel que $\mathbb{R} \subset L \subset \mathbb{C}$, l'inclusion $\mathbb{R} \subset L$ étant *stricte*. Montrer qu'il n'existe pas de structure de corps totalement ordonné sur L.

IV.1.7 Soient $K \subset \mathbb{R}$ un sous-corps et $f : \mathbb{R} \to \mathbb{R}$ une application induisant un isomorphisme du corps $\mathbb{R}$ sur le corps K. Montrer que $\mathbb{K} = \mathbb{R}$ et que f est l'identité. En particulier un automorphisme du corps $\mathbb{R}$ est l'identité.
Montrer que s'il existe un isomorphisme du corps $\mathbb{R}$ sur un corps L, il est unique.

Majorer, minorer

IV.1.8 Encadrer $x + y$, $x - y$, xy, $\dfrac{x}{y}$, sachant que $x \in [3, 6]$ et $y \in [-4, -2]$.

IV.1.9

1. Montrer que, pour tout $n \in \mathbb{N}^*$, $\sqrt{n+1} - \sqrt{n} < \dfrac{1}{2\sqrt{n}} < \sqrt{n} - \sqrt{n-1}$.
2. Calculer la partie entière du nombre réel $a := \dfrac{1}{2} \displaystyle\sum_{n=1}^{10000} \frac{1}{\sqrt{n}}$.

Sous-groupes de $\mathbb{R}$

IV.1.10 $*$ On considère le sous-ensemble $A \subset \mathbb{R}$ défini par $A := \{p + q\sqrt{2} \mid p, q \in \mathbb{Z}\}$ et l'on pose $\alpha := \sqrt{2} - 1$.

1. Vérifier que A est un sous-anneau de $\mathbb{R}$. Montrer que, pour tout $n \in \mathbb{N}^*$, $\alpha^n \in A$.
2. Montrer que $0 < \alpha < 1/2$ et que, pour tout $n \in \mathbb{N}^*$, $0 < \alpha^n < 1/n$.
   Montrer que, pour tous $x, y \in \mathbb{R}$ tels que $0 \leqslant x < y$, il existe $n \in \mathbb{N}^*$ tel que $\alpha^n < y - x$.
3. Montrer que, pour tous $x, y \in \mathbb{R}$, tels que $x < y$, il existe $a \in A$ tel que $x < a < y$, autrement dit que A est dense dans $\mathbb{R}$.

IV.1.11 Soit G un sous-groupe additif de $\mathbb{R}$. On cherche à montrer que l'on a l'alternative suivante : ou bien G est dense dans $\mathbb{R}$, ou bien il existe $a \geqslant 0$ tel que $G = a\,\mathbb{Z}$.

1. Traiter le cas où $G = \{0\}$. Dans la suite on suppose $G \neq \{0\}$.
2. On note $A := \big\{x \in G \;\big|\; x > 0\big\}$.
   Montrer que $A \neq \varnothing$ et qu'il admet une borne inférieure $a \geqslant 0$.
3. On suppose $a > 0$.
   a) Montrer qu'il existe $x \in A$ tel que $a \leqslant x < 2a$. Montrer qu'un tel x est nécessairement égal à a et que, par suite, $a \in A$.
   b) Montrer que $G = a\,\mathbb{Z}$.
4. On suppose $a = 0$. Montrer que G est dense dans $\mathbb{R}$.

IV.1.12 On considère $B \subset \mathbb{R}$ défini par $B := \{\frac{k\pi}{3.2^n} \mid k \in \mathbb{Z},\ n \in \mathbb{N}\}$.

1. Montrer que B est un sous-groupe de $(\mathbb{R}, +)$. Montrer, en utilisant l'exercice IV.1.11 que B est dense dans $\mathbb{R}$. Montrer que $\mathbb{R} \setminus B$ est dense dans $\mathbb{R}$.
2. On définit une application $f : \mathbb{R} \mapsto [0,1]$ par $f : x \mapsto f(x) := \sin^2 x$. Montrer que $f(B)$ est dense dans $[0,1]$. Montrer que $f(\mathbb{R} \setminus B)$ est dense dans $[0,1]$.
   On utilisera la proposition 35.

Suites

IV.1.13 Soit (u_n) la suite définie par $u_n := \dfrac{3n}{2n^2 - 1}$, pour tout $n \in \mathbb{N}$.

Trouver $N_1 \in \mathbb{N}$ tel que $\forall n \geqslant N_1$, $|u_n| \leqslant 10^{-4}$. Puis trouver $N_2 \in \mathbb{N}$ tel que $\forall n \geqslant N_2$, $|u_n| \leqslant \varepsilon$ pour $\varepsilon > 0$ donné.

IV.1.14 Soit (u_n) une suite définie par $u_n := \dfrac{n^2 - 2}{n^2 + n}$, pour tout $n \in \mathbb{N}$. Montrer que (u_n) est croissante et calculer sa limite.

IV.1.15 Soit (u_n) la suite définie par $u_n := \dfrac{2 + (-1)^n}{3^n}$, pour tout $n \in \mathbb{N}$. Montrer que (u_n) est décroissante et calculer sa limite.

IV.1.16 Soit (u_n) la suite réelle définie par $u_n := \dfrac{2}{9n^2 + 15n + 4}$, pour tout $n \in \mathbb{N}$.

1. Montrer que l'on peut déterminer deux réels a et b tels que $\forall n \in \mathbb{N}$, $u_n = \dfrac{a}{3n+1} + \dfrac{b}{3n+4}$.
2. En déduire la valeur de $S_n := \displaystyle\sum_{i=0}^{n} u_i$ pour $n \in \mathbb{N}$, ainsi que la limite éventuelle de S_n.

IV.1.17

Calculer les limites éventuelles des suites réelles suivantes définies par leur terme général :

$$\frac{\sin n}{n}, \quad \frac{n^3 + 2n - 1}{2n^3 - 3n + 2}, \quad \frac{n!}{n^n}, \quad \sqrt{n+1} - \sqrt{n}, \quad \frac{n - (-1)^n}{n + (-1)^n}, \quad \sum_{k=1}^{n} \frac{n}{n^2 + k}, \quad \frac{a^n - b^n}{a^n + b^n}.$$

IV.1.18

1. Montrer que la suite $(u_n := \sin n)$ diverge.
2. Soit $(v_n := \sin(\ln n))$. Montrer qu'il existe une suite extraite de (v_n) qui converge vers 1 (resp. -1). En déduire la divergence de (v_n).

IV.1.19 Trouver une suite réelle divergente $u = (u_n)$ telle que pour tout entier $k \geqslant 2$, la suite extraite (u_{kn}) converge.

IV.1.20 Soit (u_n) une suite réelle. Dire si les énoncés suivants sont vrais, et sinon donner un contre-exemple :

1. Si (u_n) converge, alors elle est monotone.
2. Si (u_n) diverge, alors elle est monotone.
3. Si (u_n) diverge, alors elle est non bornée.
4. Si (u_n) est décroissante et non minorée, alors elle tend vers $-\infty$.

IV.1.21 Soient (u_n) et (v_n) deux suites réelles, et $\ell \in \mathbb{R}$. Discuter les assertions suivantes.

1. Si $(|u_n|)$ converge vers 0, alors (u_n) converge vers 0.
2. Si $(|u_n|)$ converge vers ℓ, alors (u_n) converge vers ℓ ou $-\ell$.
3. Si (u_n) converge vers ℓ, alors $(|u_n|)$ converge vers $|\ell|$.
4. Si (u_n) est à termes strictement positifs et $\lim u_n = \ell$, alors ℓ est strictement positif.
5. Si (v_n) converge vers 0, alors $(u_n v_n)$ converge vers 0.

IV.1.22 Soit (u_n) une suite géométrique telle que $u_0 := 90$ et $\lim\limits_{n\to+\infty}(u_0+u_1+\cdots+u_n) = 150$. Quelle est sa raison ?

IV.1.23 On considère la suite positive (u_n) définie par : $u_0 := 1$ et $\forall n \in \mathbb{N},\ u_{n+1}^2 := 2u_n$, et (v_n) la suite définie par $v_n := \ln u_n - \ln 2$.

1. Montrer que (v_n) est une suite géométrique.
2. Calculer la limite de la suite (v_n), puis celle de (u_n).

IV.1.24 Dans tout l'exercice on suppose $k \in \mathbb{N}$ et $n \in \mathbb{N}^*$. On note $S_k(n) := 1^k+2^k+\cdots+n^k$. On se propose de calculer $S_k(n)$ sous forme de polynôme en n en utilisant une récurrence sur k.

1. Montrer que, pour tout $i \in \mathbb{N}^*$, $i^3 - (i-1)^3 = 3i^2 - 3i + 1$.
   En déduire $n^3 = 3S_2(n) - 3S_1(n) + S_0(n)$ et calculer $S_2(n)$.
2. Montrer que $n^{k+1} = \sum\limits_{i=1}^{n} \left(i^{k+1} - (i-1)^{k+1}\right)$. En déduire :
   $$n^{k+1} = \binom{k+1}{1} S_k(n) - \binom{k+1}{2} S_{k-1}(n) + \cdots + (-1)^k S_0(n)$$
   et la formule suivante (due à Blaise Pascal) :
   $$(k+1)S_k(n) = n^{k+1} + \binom{k+1}{2} S_{k-1}(n) - \binom{k+1}{3} S_{k-2}(n) + \cdots + (-1)^{k+1} S_0(n). \quad (14)$$
3. Montrer que, pour tout $k \in \mathbb{N}$, l'on peut calculer $S_k(n)$ sous forme de *polynôme en* n, c'est-à-dire qu'il existe un polynôme (unique) $P_k \in \mathbb{Q}[X]$, tel que, pour tout $n \in \mathbb{N}^*$, $S_k(n) = P_k(n)$. Calculer $S_k(n)$ pour $k = 3$ et $k = 4$. Comparer $S_3(n)$ et $S_1(n)$.

IV.1.25 Pour tout $k \in \mathbb{N}$ et $n \in \mathbb{N}^*$, on note $S_k(n) := 1^k + 2^k + \cdots + n^k$. On se propose de calculer $S_k(n)$ sous forme de *polynôme en* n, par une méthode de *sommes télescopiques* mais sans utiliser de récurrence (*cf.* l'exercice IV.1.61 pour une approche voisine).

On définit une application $\Delta : \begin{array}{ccc} \mathbb{Q}[X] & \longrightarrow & \mathbb{Q}[X] \\ P & \longmapsto & \Delta(P) : x \mapsto P(x) - P(x-1). \end{array}$

On note, pour tout $r \in \mathbb{N}$, $\mathbb{Q}_r[X] := \{P \in \mathbb{Q}[X] \mid \deg P \leqslant r\}$ (en convenant de $\deg 0 = -\infty$).

1. Montrer que, pour tout $r \in \mathbb{N}^*$, $\deg \Delta(x^r) = r - 1$. Montrer que Δ est une application linéraire. Quel est son noyau ? Quelle est l'image de $\mathbb{Q}_r[X]$ par Δ ?

   Montrer que si $Q \in \mathbb{Q}[X]$ ($Q \neq 0$) est un polynôme de degré k, il existe un unique polynôme $P \in \mathbb{Q}[X]$ tel que $\Delta(P) = Q$ et $P(0) = 0$. Quel est le degré de P ?

2. Pour tout $k \in \mathbb{N}$, on note $P_k \in \mathbb{Q}[X]$ l'unique polynôme tel que $\Delta(P_k) = x^k$ et $P_k(0) = 0$. Calculer $S_k(n)$ en fonction de $P_k(n)$. Calculer P_1 et P_2. En déduire $S_1(n)$ et $S_2(n)$.

3. Donner une méthode de calcul de P_k. Calculer P_3 et $S_3(n)$.

IV.1.26 On considère la fonction homographique $f : x \mapsto \frac{3x+2}{x+4}$. On note $I := [0, 1]$.

1. Montrer que $f(I) \subset I$. Calculer $f(x) - x$. Quel est le point fixe de $f : I \to I$?

2. On considère une suite récurrente (u_n) définie par f et $u_0 \in I$.

   Montrer que, pour tout $n \in \mathbb{N}$, on a $u_{n+1} - u_n = \frac{(1-u_n)(u_n+2)}{u_n+4}$.

   En déduire que la suite (u_n) est convergente. Calculer sa limite.

3. Retrouver le résultat de 2 en utilisant le point méthode de la page 539.

IV.1.27 Soit (u_n) une suite récurrente définie par u_0 et une application f. Montrer le résultat suivant. S'il existe $\tau \in \mathbb{N}$ tel que les τ suites : $(u_{p\tau+q})_{p\in\mathbb{N}^*}$, $q \in [\![0, \tau - 1]\!]$, extraites de (u_n), convergent respectivement vers ℓ_0, $\ell_1, \ldots, \ell_{\tau-1}$, les ℓ_0, $\ell_1, \ldots,$ $\ell_{\tau-1}$ sont des *points fixes* de $f^{\circ\tau}$ et les suites définies par f et les conditions initiales ℓ_q ($q \in [\![0, \tau - 1]\!]$) sont périodiques de période τ. Montrer que, dans ce cas la suite (u_n) est *bornée* et qu'elle est convergente si, et seulement si, $\ell_0 = \ell_1 = \cdots = \ell_{\tau-1}$.

IV.1.28 *Cycles de l'application tente.* L'exercice porte sur l'application tente T, *cf.* l'exemple de la page 529.

1. Trouver le nombre de 4-cycles de T, celui de 5-cycles et celui de 6-cycles.

2. On note $\omega(1) = 2$ et, pour tout $\tau \in \mathbb{N}^*$, $\tau \geqslant 2$, on note $\omega(\tau)$ le nombre de τ-cycles de T. Calculer $\sum\limits_{\tau'|\tau} \omega(\tau')$.

3. Soit p un nombre premier. Montrer que le nombre de p-cycles de T est $\frac{2^p-2}{p}$.

   Vérifier directement que $\frac{2^p-2}{p}$ est un entier. Calculer $\omega(17)$.

IV.1.29 *Suites de Collatz.*[29]
On considère l'application $f : \mathbb{Z} \to \mathbb{Z}$ définie par $f(m) = m/2$ si m est pair et $f(m) = 3m + 1$ si m est impair (on l'appelle *application de Collatz*). On a $f(\mathbb{N}) \subset \mathbb{N}$.

1. Que peut-on dire des suites récurrentes (u_n) définies par f et les conditions initiales respectives $u_0 = 0$, $u_0 = 1$, $u_0 = -1$, $u_0 = -5$, $u_0 = -17$?
2. On conjecture que pour tout $u_0 \in \mathbb{N}^*$, il existe $n \in \mathbb{N}$ tel que $u_n = 1$. Le mathématicien portugais Tomás Oliveira e Silva a vérifié en 2009 (à l'aide d'un ordinateur) que c'est vrai si $u_0 \leqslant 20 \times 2^{58}$ ($20 \times 2^{58} > 5 \times 10^{18}$). Vérifier (plus modestement...) que le résultat est vrai pour $u_0 \leqslant 26$. Programmer le calcul de la suite et vérifier le résultat pour quelques valeurs initiales, par exemple $u_0 = 12345$, $u_0 = 99999$, $u_0 = 837799$, $u_0 = 100759293214567$ et $u_0 = 279731455495736617$.
3. Montrer que si une suite récurrente définie par f et $u_0 \in \mathbb{Z}^*$ est bornée, alors son image est un cycle.
4. On suppose que A est un cycle de f contenu dans $\mathbb{N}$ (on dira alors que A est un cycle positif). Ceci équivaut évidemment à $u_0 \in \mathbb{N}$.
   a) On note a le plus petit élément de A (pour la relation d'ordre usuelle sur $\mathbb{N}$). Montrer que a est impair.
   b) On suppose que $\operatorname{card} A = 5$ et l'on note $A = \{a_1, a_2, a_3, a_4, a_5\}$ en supposant que a_1 est le plus petit élément de A.
   Montrer que $a_1 \geqslant 3$. Montrer que a_2 est pair, a_3 impair, a_4 et a_5 pairs.
   En considérant $P := a_1a_2a_3a_4a_5$, montrer que $8a_1a_3 = (3a_1 + 1)(3a_3 + 1)$ et en déduire une contradiction. Conclure.

IV.1.30 Soient (x_n) une suite de $\mathbb{K}$ et $a \in \mathbb{K}$. Montrer que les trois conditions suivantes sont équivalentes.

(i) La suite (x_n) admet a comme valeur d'adhérence.

(ii) Pour tout $\varepsilon > 0$ et tout $N \in \mathbb{N}$, il existe $n \geqslant N$ tel que $|x_n - a| < \varepsilon$.

(iii) Pour tout $\varepsilon > 0$, il existe une infinité de valeurs de n telles que $|x_n - a| < \varepsilon$.

IV.1.31 $*$

1. Soit (u_n) une suite réelle telle que $u_{n+1} - u_n \to 0$ quand $n \to +\infty$. Montrer que l'ensemble des valeurs d'adhérence de (u_n) est un intervalle. On utilisera le résultat de l'exercice précédent.
2. Soit $v := (v_n = \sin(\ln n))$. Montrer que l'ensemble des valeurs d'adhérence de la suite v est l'intervalle $[-1, 1]$ (on utilisera l'exercice IV.1.18).

IV.1.32 Soient $I \subset \mathbb{R}$ un segment et $f, g : I \to I$ deux fonctions continues telles que $f \circ g = g \circ f$. Montrer qu'il existe $x \in I$ tel que $f(x) = g(x)$.

IV.1.33 Soient α un nombre *irrationnel* positif et deux suites (p_n) et (q_n) d'entiers strictement positifs telles que la suite (p_n/q_n) converge vers α. Montrer que les suites (p_n) et (q_n) tendent vers $+\infty$.

[29] Le problème dit de Collatz, ou de Hasse, ou de Ulam, ou de Syracuse, ou $3n + 1$... est apparu aux États Unis vers 1935.

IV.1.34 Soient $\alpha \in \mathbb{R}$ et (p_n/q_n) une suite de rationnels telle que $0 < \left|\alpha - p_n/q_n\right| < \varepsilon_n/q_n$, la suite (ε_n) tendant vers 0. Montrer que le réel α est irrationnel.

IV.1.35 $*$ Pour tout $x \in \mathbb{R}$, on note $\langle x \rangle := x - E(x)$ où $E(x)$ est la partie entière de x.
Soit $\alpha \in \mathbb{R}$.

1. On suppose α irrationnel. Soient $m_1, m_2 \in \mathbb{N}$. Montrer que si $\langle m_1\alpha \rangle = \langle m_2\alpha \rangle$, alors $m_1 = m_2$.
   Soit $Q \in \mathbb{N}^*$. Montrer qu'il existe deux entiers distincts $m_1, m_2 \in [\![1, Q]\!]$ tels que $\left|\langle m_1\alpha \rangle - \langle m_2\alpha \rangle\right| < 1/Q$ (on utilisera le « principe des tiroirs »).
   Montrer qu'il existe $p \in \mathbb{Z}$ et $q \in \mathbb{N}^*$, tels que $|p/q - \alpha| < 1/q\,Q < 1/q^2$.
2. On suppose α irrationnel. Montrer qu'il existe un ensemble *infini* de rationnels p/q ($p \in \mathbb{Z}$, $q \in \mathbb{N}^*$) tels que $|p/q - \alpha| < 1/q^2$ (ce résultat est dû à Peter Gustav Dirichlet 1842).
3. On suppose α rationnel. Montrer que l'ensemble des rationnels p/q ($p \in \mathbb{Z}$, $q \in \mathbb{N}^*$) tels que $|p/q - \alpha| < 1/q^2$ est fini.

IV.1.36 Soit $0 < q < 1$.

1. Pour tout $n \in \mathbb{N}$, on note $w_n := (1+q)(1+q^2)(1+q^4)\dots(1+q^{2^n})$. Calculer $(1-q)w_n$. Étudier la suite (w_n).
2. Soient $a, b > 0$. On définit deux suites (u_n) et (v_n) par les relations de récurrence :
$$u_{n+1} = \frac{u_n^2}{u_n + v_n} \quad \text{et} \quad v_{n+1} = \frac{v_n^2}{u_n + v_n}$$
   et $u_0 := a$, $v_0 := b$. Montrer qu'elles convergent et calculer leurs limites.
3. Comparer les questions 1 et 2.
4. Soient $c > 0$ et $f : x \mapsto \frac{x^2}{2x-c}$. montrer que $[c, +\infty[$ est invariant par f et étudier les suites récurrentes définies par f et $u_0 := a$, avec $a \geqslant c$. Comparer les questions 2 et 4.

IV.1.37 Étudier les suites récurrentes réelles définies par les fonctions suivantes.

1. On définit $f : \mathbb{R} \to \mathbb{R}$ par $f : x \mapsto x^2$.
2. On définit $f : \mathbb{R} \to \mathbb{R}$ par $f : x \mapsto x^2 + 1$.
3. On définit $f :]-\infty, 2] \to \mathbb{R}$ par $x \mapsto \sqrt{2-x}$
4. Soit $a > 0$. On définit $f : \mathbb{R}_+ \to \mathbb{R}$ par $x \mapsto \sqrt{a+x}$ et l'on suppose $u_0 = a$.
5. On définit $f : \mathbb{R} \to \mathbb{R}$ par $f : x \mapsto \frac{x^3+1}{3}$ (on montrera que l'équation $f(x) - x = 0$ admet trois racines réelles distinctes).

IV.1.38 Soient $a,\ b > 0$. Montrer que l'on peut définir par récurrence une suite (u_n) telle que $u_{n+2} = u_{n+1}^2/u_n$ et $u_0 := a$, $u_1 := b$. Étudier la convergence de cette suite.

IV.1.39 On définit par récurrence une suite u par la donnée de $u_0 > 0$ et par la condition $u_{n+1} := \sqrt{u_0 + u_1 + \cdots + u_n}$.

1. Montrer que la suite u tend vers $+\infty$.
2. On définit une suite $v := (v_n := u_{n+1} - u_n)$. Montrer que v converge et calculer sa limite.

IV.1.40 $*$ Soit $v := (v_n)$ une suite de réels positifs *donnée*.
On définit, par récurrence, une suite de réels par $u_0 := v_0$ et :

$$u_n := \sqrt{v_1 + \sqrt{v_2 + \cdots + \sqrt{v_n}}}.$$

Donner une condition nécessaire et suffisante sur la suite v pour que u converge (on pourra d'abord étudier le cas particulier où v_n est de la forme $v_n := ab^{2^n}$).

IV.1.41 $*$ Soit $v := (v_n)$ une suite de réels positifs *donnée*.

1. Soit $\ell \in \mathbb{R}_+$ donné. On définit, par récurrence, une suite w de réels par $w_0 = 1$ et :

$$w_{n+1} := \frac{1}{w_n + \ell + 1}.$$

   Montrer que w converge vers une limite strictement positive notée ℓ_1.

2. On suppose que la suite v est bornée. On définit, par récurrence, une suite de réels par $u_0 > 0$ et :

$$u_{n+1} := \frac{1}{u_n + v_n + 1}.$$

   Est-ce que u est une suite récurrente ? Montrer que la suite u converge si, et seulement si, la suite v converge. Si u_n converge, calculer sa limite en utilisant 1.

IV.1.42 Soient $r > 0$ et $\theta \in \,]-\pi, \pi[$. On considère la suite récurrente complexe (z_n) définie par $z_0 := re^{i\theta}$ et $f : \mathbb{C} \to \mathbb{C},\ f : z \mapsto \frac{z+|z|}{2}$.
Montrer qu'elle converge et calculer sa limite ℓ (on pourra écrire z_n comme un produit).

IV.1.43

1. a) Soit $\lambda \in [0, 1]$. Montrer que, pour tout $n \in \mathbb{N}$, $\lambda 5^n + (1-\lambda)6^n > 0$. On définit une suite réelle : $u := \left(u_n := \dfrac{\lambda 5^{n+1} + (1-\lambda)6^{n+1}}{\lambda 5^n + (1-\lambda)6^n}\right)_{n\in\mathbb{N}}$.
   Montrer qu'elle converge et calculer sa limite.

   b) Montrer que u est une suite récurrente homographique.

2. On définit des suites récurrentes à deux termes par la récurrence :

$$w_{n+2} = 11w_{n+1} - 30w_n \tag{15}$$

   et les conditions initiales w_0 et w_1. Montrer que l'ensemble E des suites solutions de la récurrence (15) est un sous-espace vectoriel *de dimension* 2 de l'espace des suites réelles. Calculer $\rho \in \mathbb{R}$ pour que la suite (ρ^n) appartienne à E et en déduire une base de E.

3. On considère la récurrence homographique :

$$v_{n+1} = \frac{11v_n - 30}{v_n}. \tag{16}$$

   Montrer que l'étude de la récurrence (16) est équivalente à celle de la récurrence (15).
   En déduire, quand la récurrence (16) est bien définie, une formule pour v_n.

4. On reprend les notations de 1. Soit (t_n) la suite réelle « définie » par la relation de récurrence :

$$t_{n+2} = 111 - \frac{1130}{t_{n+1}} + \frac{3000}{t_n t_{n+1}}. \tag{17}$$

et les conditions initiales $t_0 := u_0$ et $t_1 := u_1$. Montrer que, pour tout $n \in \mathbb{N}$, $t_n = u_n$ et, par suite, $t_n \neq 0$.

IV.1.44 *« Divergence numérique » d'une suite convergente.* On revient sur l'exercice 27 de la page 550. On considère la suite définie par la récurrence linéaire d'ordre 2 de Fibonacci : $w_{n+2} = w_n + w_{n+1}$ avec les conditions initiales $w_0 := 1$ et $w_1 := \frac{1-\sqrt{5}}{2} + \varepsilon$, où ε est un réel fixé. Calculer w_n pour tout $n \in \mathbb{N}$ et montrer que si $\varepsilon \neq 0$, alors la suite (w_n) est divergente. Comparer au cas $\varepsilon = 0$ et expliquer le phénomène numérique signalé dans l'exercice 27.

IV.1.45 Soient $a, b \in \mathbb{R}_+^*$. On considère les *deux* suites $u := (u_n)$ et $v := (v_n)$ définies par $(u_0, v_0) := (a, b)$ et la relation de récurrence : (R) $\begin{cases} u_{n+1} = \dfrac{u_n + v_n}{2} \\ v_{n+1} = \sqrt{u_{n+1} v_n}. \end{cases}$

1. Soit $\alpha \in]0, \pi/2[$. On suppose $(a, b) := (\cos\alpha, 1)$.
   a) Calculer u_n et v_n en utilisant des fonctions trigonométriques.
   b) Montrer que les suites (u_n) et (v_n) admettent une limite commune ℓ et la calculer en fonction de α.
2. On revient au cas général.
   a) Soit $\lambda \in \mathbb{R}_+^*$. On définit deux suites (u'_n) et (v'_n) par la relation de récurrence (R) et $(u'_0, v'_0) = (\lambda u_0, \lambda v_0)$. Comparer les suites (u_n) et (u'_n) et les suites (v_n) et (v'_n). Montrer que (u_n) (resp. (v_n)) converge si, et seulement si, (u'_n) (resp. (v'_n)) converge et comparer les limites.
   b) Montrer que les suites (u_n) et (v_n) admettent une limite commune et la calculer en fonction de a et b. On traitera d'abord le cas $a \leqslant b$. Pour le cas $a > b$, on utilisera des fonctions trigonométriques hyperboliques.

IV.1.46 *Moyenne arithmético-géométrique.* Soient deux réels $a, b > 0$. On considère les *deux* suites (a_n) et (b_n) définies par $(a_0, b_0) := (a, b)$ et la relation de récurrence :

$$a_{n+1} = \frac{a_n + b_n}{2} \quad \text{et} \quad b_{n+1} = \sqrt{a_n b_n}.$$

On suppose dans la suite $a \geqslant b > 0$.

1. Montrer que la suite (a_n) est décroissante, que la suite (b_n) est croissante et que ces deux suites sont adjacentes. On note $M(a, b)$ leur limite commune. Elle est appelée *moyenne arithmético-géométrique* de a et b.
2. On pose, pour tout $n \in \mathbb{N}$, $c_n := \sqrt{a_n^2 - b_n^2}$.

   Montrer que la suite (c_n) décroît vers 0 et que $c_{n+1} \leqslant \frac{c_n^2}{4M(a,b)}$.

   Montrer qu'il existe $n_0 \in \mathbb{N}^*$ tel que $a_n - b_n \leqslant 8M(a, b)10^{-2^{n+1-n_0}}$.

   Commenter la « vitesse de convergence » des suites (c_n), (a_n) et (b_n).

3. On suppose[30] $a := 1$ et $b := 1/\sqrt{2}$ et l'on note $M := M(a,b)$.

   Montrer que $c_1/4M \leqslant c_1/4b_1 < 1/10$. En déduire, que, pour tout $n \in \mathbb{N}^*$, on a $c_n \leqslant 4 \times 10^{-2^{n-1}}$. Montrer que a_{30} donne une approximation de M avec *un milliard* de décimales.

IV.1.47 On considère les conditions initiales $x_1, x_2 \in \mathbb{R}$ et la relation de récurrence $x_{n+2} = \frac{1+x_{n+1}}{x_n}$. Montrer que si $x_1, x_2 \neq 0$, $x_1 \neq -1$ et $x_1 + x_2 \neq -1$, cela définit une suite récurrente périodique.

IV.1.48 *Équation différentielle logistique (modélisation d'un problème écologique)*

1. On considère l'équation différentielle, dite logistique, $\mathrm{d}x/\mathrm{d}t = \rho x(1-x)$ (où $\rho \in \mathbb{R}$ est une constante fixée et où x est une fonction inconnue de la variable t). Trouver des solutions en utilisant le changement de fonction inconnue $y := 1/x$.

2. Expliquer pourquoi on peut utiliser l'équation différentielle logistique pour modéliser l'évolution d'une population (*cf.* la page 570).

IV.1.49 $*$ *Suites logistiques.* Soit $I := [0,1]$. Considérons, pour $\mu \in]0,4]$, l'application logistique $f_\mu : x \in I \mapsto \mu x(1-x) \in I$. Le but de l'exercice est de montrer que, si $0 < \mu < 3$, la suite logistique est convergente pour toute valeur initiale $u_0 \in I$.

(i) (a) On suppose $0 < \mu \leqslant 1$. Montrer que si $u_0 \in [0,1]$, la suite définie par f_μ et u_0 est décroissante et converge vers 0.

(b) On suppose $1 < \mu \leqslant 2$. Traiter les cas $u_0 = 0,\ \alpha_1,\ 1,\ 1/\mu$. Montrer que si $u_0 \in]0,\alpha_1]$ (resp. $u_0 \in]\alpha_1, 1/2[$), la suite définie par f_μ et u_0 est croissante (resp. décroissante) et converge vers α_1.

Que se passe-t-il pour $u_0 \in]1/2, 1[$?

(ii) On suppose $2 < \mu < 3$. On pose $\gamma := 1/\mu$. On a $1/3 < \gamma < 1/2 < \alpha_1 < 2/3$.

On pose $J := [\gamma, \alpha_1]$.

(a) Montrer que $f \circ f(J) \subset [1/2, \alpha_1] \subset J$. Montrer que si $u_0 \in J$, la suite récurrente définie par $f \circ f$ et u_0 converge vers α_1. En déduire que la suite récurrente définie par f et u_0 converge vers α_1.

(b) On suppose $0 < u_0 < \gamma$. Montrer qu'il existe $m \in \mathbb{N}^*$ tel que $u_m \in J$. En déduire que la suite récurrente définie par f et u_0 converge vers α_1.

(c) On suppose $\gamma < u_0 < 1$. Montrer que la suite récurrente définie par f et u_0 converge vers α_1.

---

[30] Ce cas est utilisé dans un calcul très rapide de π par une suite en utilisant une formule découverte par Richard Brent et Eugene Salamin (1975 et 1976) : on obtient plus de deux milliards de décimales de π au 30ème terme, avec un temps de calcul sur ordinateur très court ; *cf.* Jean-Pierre Demailly, *Analyse numérique et équations différentielles,* `www.grenoble-sciences.fr/pap-ebook`.

IV.1.50 Pour tous $\alpha \in \mathbb{K}^*$, $\beta \in \mathbb{K}$, on note $h_{\alpha,\beta} : x \in \mathbb{K} \mapsto \xi := \alpha x + \beta \in \mathbb{K}$; c'est une transformation affine.

1. Montrer que l'inverse de $h_{\alpha,\beta}$ est une transformation affine : $h_{\alpha,\beta}^{-1} = h_{\alpha',\beta'}$. Calculer α' et β'.
2. Soit $f : \mathbb{R} \to \mathbb{R}$ définie par $x \mapsto f(x) := ax^2+bx+c$, avec $a \in \mathbb{R}^*$, $b, c \in \mathbb{R}$. On considère $g := h \circ f \circ h^{-1}$, avec $h := h_{\alpha,\beta}$. On vérifie que g est encore un polynôme de degré 2 en x, rappelons que l'on dit que g est *conjuguée* de f par une transformation affine.
   a) Soient a, b, c, $\gamma \in \mathbb{K}$. Montrer qu'il existe une unique transformation affine $h_{\alpha,\beta}$ conjuguant f à $g_\gamma : x \mapsto x^2 - \gamma$ et qu'elle est donnée par $\alpha = a$, $\beta = \frac{b}{2}$, $\gamma = -ac - \frac{b}{2} + \frac{b^2}{4}$.
   b) Montrer que l'application logistique f_μ est conjuguée à g_γ si, et seulement si, $\alpha = -\mu, \beta = \frac{\mu}{2}$ et $\gamma = -\frac{\mu}{2} + \frac{\mu^2}{4}$. Montrer que f_4 est conjuguée à $g_2 : x \mapsto x^2 - 2$. Montrer que f_μ et $f_{2-\mu}$ sont conjuguées.

IV.1.51 On note, pour $x \in [0,1]$, $f_4(x) = 4x(1-x)$. Trouver une relation entre l'itération de f_4 et celle de l'application de doublement de l'angle (*cf.* l'exemple 3 de la page 576). En déduire un moyen d'étudier f_4 en utilisant l'écriture binaire des nombres réels.

IV.1.52 $*$ *Récurrences géométriques et arithmético-géométriques : les cas « difficiles ».*
Dans cet exercice, on utilise les résultats de l'exercice IV.1.11.
On note $U := \{z \in \mathbb{C} \mid |z| = 1\}$.
Soient $\xi \in \mathbb{R} \setminus \mathbb{Q}$ et $k := e^{2i\pi\xi}$. On note :

$$A := \mathbb{Z} + \mathbb{N}\xi = \{s + m\xi \mid s \in \mathbb{Z},\ m \in \mathbb{N}\} \quad \text{et} \quad B := \mathbb{Z} + \mathbb{Z}\xi = \{s + m\xi \mid s \in \mathbb{Z},\ m \in \mathbb{Z}\}.$$

On vérifie que A est stable pour l'addition et que B est un sous-groupe additif de $\mathbb{R}$.

1. Montrer que B est dense dans $\mathbb{R}$.
   En déduire qu'il existe une suite d'entiers (x_n) et une suite d'entiers *positifs* (y_n) telles que, pour tout $n \in \mathbb{N}$, $x_n + y_n\xi \neq 0$ et que $\lim\limits_{n\to+\infty}(x_n + y_n\xi) = 0$.
   Montrer que la suite (y_n) n'est pas bornée.
2. Montrer que A est dense dans $\mathbb{R}$.
3. On note $\mathcal{X} := \{k^n\}_{n\in\mathbb{N}}$ et $\mathcal{X}' := \{k^{-n}\}_{n\in\mathbb{N}}$. Montrer que $\mathcal{X}$ et $\mathcal{X}'$ sont denses dans U.
4. a) Soient $u_0 \in \mathbb{C}^*$ et $(u_n := u_0k^n)$. Montrer que l'image de la suite géométrique (u_n) est dense dans le cercle $\mathcal{C} \subset \mathbb{C}$ de centre 0 et de rayon $|u_0|$.
   b) Soient $b \in \mathbb{C}^*$, $u_0 \in \mathbb{C}^*$ et la suite (u_n) définie par la relation de récurrence $u_{n+1} := ku_n + b$. Montrer que l'image de la suite arithmético-géométrique (u_n) est dense dans le cercle $\mathcal{C} \subset \mathbb{C}$ de centre $\ell := \frac{b}{1-k}$ et de rayon $|u_0 - \ell|$.
5. Montrer que l'image de la suite $(u_n := \sin n)_{n\in\mathbb{N}}$ est dense dans $[-1,1]$.

IV.1.53 *Récurrences homographiques : les cas « difficiles ».*
Dans cet exercice, on utilise les résultats de l'exercice précédent IV.1.52.
On note $U := \{z \in \mathbb{C} \mid |z| = 1\}$.

1. (i) Soit $k \in U$, $k \neq 1$. Montrer qu'il existe une homographie *réelle* $f : x \mapsto \frac{ax+b}{cx+d}$ admettant $\pm i$ pour points fixes et telle que $k = \frac{d-ic}{d+ic}$.
   (ii) Montrer que, pour tout $p \in \mathbb{N}^*$, $p \geqslant 2$, il existe une homographie *réelle* f telle que toute suite homographique réelle définie par f soit périodique de plus petite période p.
2. Soient $\xi \in \mathbb{R} \setminus \mathbb{Q}$, $k := e^{2i\pi\xi}$, $\alpha \in \{\mathbb{C} \setminus \mathbb{R}\}$, et une homographie réelle $f : x \mapsto \frac{a\,x+b}{c\,x+d}$ admettant α et $\bar{\alpha}$ pour points fixes et telle que telle que $k = \frac{c\,\bar{\alpha}+d}{c\,\alpha+d}$.
   Montrer que, si la suite homographique $(u_{n+1} := \frac{a\,x+b}{c\,x+d})$ est définie, alors son image est dense dans $\mathbb{R}$. Montrer que l'ensemble des valeurs de u_0, pour lesquelles la suite (u_n) n'est pas définie, est dense dans $\mathbb{R}$.
3. Soient $\xi \in \mathbb{R} \setminus \mathbb{Q}$, $k := e^{2i\pi\xi}$ et $\alpha \in \mathbb{C} \setminus \mathbb{R}$. Montrer qu'il existe une homographie réelle $f : x \mapsto \frac{a\,x+b}{c\,x+d}$ admettant α et $\bar{\alpha}$ pour points fixes et telle que telle que $k = \frac{c\,\bar{\alpha}+d}{c\,\alpha+d}$.

Développement décimal

IV.1.54
Montrer qu'un nombre décimal inversible dans $\mathbb{D}$ est de la forme $\pm 2^m 5^n$.

IV.1.55

1. Montrer que pour tout $n \in \mathbb{N}^*$, la première décimale du nombre réel $\sqrt{n^2+n}$ est 4.
2. Montrer que pour tout $n \in \mathbb{N}^*$, $n \geqslant 5$, la première décimale du nombre réel $\sqrt{n^2+2n}$ est 9.

IV.1.56 *Nombres de Liouville.*
Soit $u := 0,\beta_1\beta_2\ldots\beta_h\ldots$ une écriture décimale. On suppose que les β_h ne sont pas nuls à partir d'un certain rang c'est-à-dire que u est un réel non décimal ou l'écriture impropre d'un décimal. On note alors $u \in \mathcal{U}$.

1. Montrer que la série $\sum\limits_{k\in\mathbb{N}^*} \frac{\beta_k}{10^{k!}}$ est convergente. On note $x(u) := 0,\alpha_1\alpha_2\ldots\alpha_h\ldots$ sa somme. Montrer que : $\forall n \in \mathbb{N}^*, \quad x_{n!} := 0,\alpha_1\alpha_2\ldots\alpha_{n!} = \sum\limits_{k=1}^{n} \frac{\beta_k}{10^{k!}}$. On dit que $x(u)$ est un nombre de Liouville.
2. Montrer que le nombre réel $x(u)$ est irrationnel.
3. Montrer que l'ensemble des nombres de Liouville $\{x(u)\}_{u\in\mathcal{U}}$ n'est pas dénombrable.

IV.1.57 Soit $k \in \mathbb{N}^*$ fixé. Pour tout $n \in \mathbb{N}$ tel que $n \geqslant k$, on note α_n le *chiffre des unités* dans l'écriture en base 10 de l'entier $\binom{n}{k}$ ($\binom{n}{k} = \alpha_n \mod 10$).
Si $k \geqslant 2$, on définit $\alpha_1 = \alpha_2 = \ldots = \alpha_{k-1} = 0$. Montrer que $x := 0,\alpha_1\alpha_2\ldots\alpha_h\ldots$ est l'écriture décimale d'un rationnel ou l'écriture décimale impropre d'un décimal.

IV.1.58 À toute partie $P \subset \mathbb{N}^*$ on associe le nombre réel $x(P) = 0,\alpha_1\alpha_2\ldots\alpha_h\ldots \in [0,1[$ défini par $\alpha_h := 1$ si $h \in P$ et $\alpha_h := 0$ si $h \notin P$. En utilisant l'application $f : \mathcal{P}(\mathbb{N}^*) \to [0,1[$ définie par $f(P) := x(P)$, montrer que l'ensemble $J := [0,1[$ n'est pas dénombrable. En déduire une nouvelle preuve de la non dénombrabilité de $\mathbb{R}$.

Quelques exercices plus difficiles

Les exercices (ou petits problèmes) suivants peuvent être pris comme points de départ pour un mémoire ou un travail personnel. Ils portent parfois sur *plusieurs* modules du livre et utilisent des approches « croisées ».

IV.1.59 $*$ *Polynômes réciproques et irrationalité de certaines valeurs de* $\cos$ *et* $\sin$.

1. On définit une suite de polynômes de $\mathbb{Q}[X]$ par la relation de récurrence :

$$S_{n+1} = XS_n - S_{n-1} \qquad \text{avec} \qquad S_0 := 2 \quad \text{et} \quad S_1 := X.$$

   a) Calculer S_n pour $n \in [\![2,7]\!]$. Montrer que, pour tout $n \in \mathbb{N}^*$, S_n est un polynôme unitaire à coefficients entiers.
   Montrer que, pour tout $n \in \mathbb{N}^*$ et tout $X \in \mathbb{R}^*$, $S_n\left(X + \frac{1}{X}\right) = X^n + \frac{1}{X^n}$.

   b) Montrer que, $\forall n \in \mathbb{N}^*$ et $\forall \theta \in \mathbb{R}$: $S_n(2\cos\theta) = 2\cos n\theta$.

2. Soient $P \in \mathbb{Q}[X]$ un polynôme unitaire à coefficients entiers et $\xi \in \mathbb{Q}$ une racine de P. Montrer que $\xi \in \mathbb{Z}$.

3. Soit $\lambda \in \mathbb{Q}$. Montrer que $\cos\lambda\pi$ est irrationnel sauf si $\lambda = 0 \mod \pi$ ou $\pm 1/3 \mod \pi$ ou $\pm 1/2 \mod \pi$. Montrer que $\sin\lambda\pi$ est irrationnel sauf si $\lambda = 0 \mod \pi$ ou $\pm 1/6 \mod \pi$ ou $\pm 1/2 \mod \pi$.

4. On note $\mathbb{K} = \mathbb{C}, \mathbb{R}$ ou $\mathbb{Q}$. Soit $P \in K[Z]$ un polynôme de degré d. On dit que c'est un *polynôme réciproque* si, pour tout $Z \in \mathbb{K}^*$, $Z^n P(1/Z) = P(Z)$. Donner une condition équivalente en utilisant les coefficients de P. Montrer que si d est pair : $d := 2\delta$, P est un polynôme réciproque si, et seulement s'il existe un polynôme Q de degré δ tel que, pour tout $Z \in \mathbb{K}^*$, $P(Z) = Z^\delta Q\left(Z + \frac{1}{Z}\right)$. Calculer Q en fonction de P en utilisant les polynômes $\{S_n\}$ de la question 1. Montrer que Q est à coefficients entiers si, et seulement si, P est à coefficients entiers. Appliquer au cas où $P := Z^4 + Z^3 + \cdots + Z + 1$ et au cas où $P := Z^6 + Z^5 + \cdots + Z + 1$.

5. Soit $n \in \mathbb{N}^*$ impair : $n = 2\nu + 1$. Montrer que, pour tout $n \in \mathbb{N}$ le polynôme $V_n := (X - 2\cos 2\pi/n)(X - 2\cos 4\pi/n)\ldots(X - 2\cos 2\nu\pi/n) \in \mathbb{R}[X]$ est à coefficients entiers. Comparer avec l'exercice IV.1.5.
   Trouver, pour n impair, toutes les racines dans $\mathbb{C}$ de $S_n - 2$. Comparer les polynômes $S_5 - 2$ et V_5 (resp. $S_7 - 2$ et V_7).

6. Comparer les polynômes $\{S_n\}$ et les polynômes de Tchebychef de première espèce $\{T_n\}$ (*cf.* la définition 11 de la page 311). Montrer que, $\forall n \in \mathbb{N}^*$, impair, on a $S_n - 2 = (X-2)V_n^2$.

7. Pour tout $n \in \mathbb{N}^*$, on note Φ_n le $n^{ème}$ polynôme cyclotomique (*cf.* définition 10 de la page 309).

   a) Calculer Φ_{15}.
   En déduire un polynôme P de degré 4 à coefficients entiers tel que $P(\cos 2\pi/15) = 0$.

   b) Montrer que, pour tout $n \in \mathbb{N}$ tel que $n \geqslant 2$, Φ_n est un polynôme réciproque.

IV.1.60 $*$ *Irrationalité de* π, *de* π^2 *et de* e^a *pour* $a \in \mathbb{Q}^*$.

Pour tout $n \in \mathbb{N}^*$, on définit $I_n := \frac{1}{2n!}\displaystyle\int_0^\pi x^n(\pi - x)^n \sin x \,\mathrm{d}x$. On note $\mathbb{Z}[\pi^2]$ l'ensemble des « polynômes en π^2 » c'est-à-dire des combinaisons linéaires finies à coefficients entiers de puissances entières positives de π^2.

1. Calculer I_n pour $n \in [\![2,5]\!]$. Montrer que $\forall n \in \mathbb{N}^*$: $I_n > 0$ et $I_n < \frac{\pi^{2n+1}}{n!}$.

2. Soit $n \in \mathbb{N}$. On note $f(x) := x^n(\pi - x)^n$. Montrer que :

$$\int_0^\pi f(x)\sin x\,\mathrm{d}x = 2f(0) - 2f''(0) + 2f^{(4)}(0) - \cdots + (-1)^n 2f^{2n}(0).$$

Montrer que, pour tout $k \in \mathbb{N}$, tel que $n \leqslant k \leqslant 2n$, $f^{(k)}(0) = k!\,a_{n,k}\,\pi^{2n-k}$, avec $a_{n,k} \in \mathbb{Z}$. En déduire que $I_n \in \mathbb{Z}[\pi^2]$ et que son « degré » est au plus $E[n/2]$.

3. Montrer que π et π^2 sont irrationnels.

4. Soit $a \in \mathbb{Q}^*$. En utilisant les intégrales $J_n := \frac{1}{n!}\displaystyle\int_0^a x^n(\mathrm{e}^a - x)^n \mathrm{e}^x\,\mathrm{d}x$, montrer que e^a est irrationnel. Montrer que si $b \in \mathbb{Q}_+^* \setminus \{1\}$, alors $\ln b$ est irrationnel.

IV.1.61 $**$ *Sommes de puissances d'entiers.*

Pour tout $k, n \in \mathbb{N}^*$, on note $S_k(n) := 1^k + 2^k + \cdots + n^k$. On se propose de calculer *une formule* pour $S_k(n)$ en utilisant des nombres rationnels dits de Bernoulli (*cf.* l'exercice 11 de la page 886). On utilise les résultats des deux exercices IV.1.24 et IV.1.25 des pages 602 et 603 .

1. Le but de cette question est de mettre en évidence un « ordre caché » dans les coefficients des polynômes (en n) : $(k+1)S_k(n)$ pour $k \in [\![0, 10]\!]$ en les comparant au triangle de Pascal.

   a) Écrire les 11 premières lignes du triangle de Pascal (*cf.* page 79).

   b) Calculer $(k+1)S_k(n)$ pour $k \in [\![0, 10]\!]$ en utilisant la formule (14) de l'exercice IV.1.24.
   Quels sont, pour $k \in [\![1, 10]\!]$, les coefficients de n^{k+1} et n^k dans le polynôme $(k+1)S_k(n)$?

   c) Écrire la liste, pour $k \in [\![2, 10]\!]$, des coefficients de n^{k-1} dans $(k+1)S_k(n)$. Réduire les termes de cette liste au même dénominateur (le plus petit) et écrire la liste des numérateurs. La comparer à la *troisième* colonne du triangle de Pascal et conclure.

   d) Quel est, pour $k \in [\![3, 10]\!]$, le coefficient de n^{k-2} dans $(k+1)S_k(n)$?
   Quel est, pour $k \in [\![5, 10]\!]$, le coefficient de n^{k-4} dans $(k+1)S_k(n)$.
   Procéder comme en c) pour les coefficients de n^{k-3} (pour $k \in [\![4, 10]\!]$), et n^{k-5} (pour $k \in [\![6, 10]\!]$).
   Écrire, pour $k \in [\![0, 10]\!]$, les polynômes $(k+1)S_k(n)$ en utilisant les nombres rationnels $\{f_i\}_{i \in [\![0,8]\!]}$ définis de la façon suivante : $f_0 := 1$, $f_1 := \frac{1}{2}$, $f_2 := \frac{1}{6}$, $f_3 := 0$, $f_4 := -\frac{1}{30}$, $f_5 := 0$, $f_6 := \frac{1}{42}$, $f_7 := 0$, $f_8 := -\frac{1}{30}$ et les observations qui précédent.
   Trouver des formules de récurrence liant les f_n pour $n \in [\![0, 8]\!]$.

2. On définit par récurrence des nombres rationnels f_n par $f_0 := 1$ et, pour tout $k \in \mathbb{N}$, la formule :

$$k + 1 = f_0\binom{k+1}{0} + f_1\binom{k+1}{1} + f_2\binom{k+1}{2} + \cdots f_k\binom{k+1}{k}. \tag{18}$$

   Vérifier la compatibilité avec les notations de la question précédente.
   Comparer les nombres f_n et les nombres b_n et B_p définis dans l'exercice 11 de la page 886 et sa solution. Montrer que, pour tout n impair, $n \geqslant 3$, on a $f_n = 0$.

3. Montrer que, pour tous $n \in \mathbb{N}^*$ et $a, b \in [\![0, n]\!]$, on a : $\binom{n}{a}\binom{n-a}{b} = \binom{n}{b}\binom{n-b}{a}$.
4. Soit $\Delta : \begin{array}{rcl} \mathbb{Q}[X] & \longrightarrow & \mathbb{Q}[X] \\ P & \longmapsto & \Delta(P) : x \mapsto P(x) - P(x-1) \end{array}$ l'application définie dans l'exercice IV.1.25. On note, pour $k \in \mathbb{N}$, $\Phi_k(x) := \sum_{p=0}^{k} \binom{k+1}{p} f_p x^{k-p+1}$.
   a) Justifier l'introduction des Φ_k.
   Montrer que, pour $k \in [\![1, 10]\!]$, on a $\Delta(\Phi_k) = (k+1)x^k$.
   b) Montrer, en utilisant la question 3, que, pour tout $k \in \mathbb{N}$, on a $\Delta(\Phi_k) = (k+1)x^k$. En utilisant l'exercice IV.1.25, en déduire $S_k(n) = \frac{\Phi_k(n)}{k+1}$, c'est-à-dire la formule (dite de Faulhaber) :
$$S_k(n) := \frac{1}{k+1} \sum_{p=0}^{k} \binom{k+1}{p} f_p n^{k-p+1}.$$

IV.1.62 *Fausses convergences numériques de suites et sensibilité aux conditions initiales.*
L'exercice suivant montre que des calculs *numériques* sur ordinateur, même s'ils sont faits avec une grande précision, peuvent conduire à des résultats « aberrants ».

1. Soit (u_n) la suite réelle « définie » par la relation de récurrence
$$u_{n+2} = 111 - \frac{1130}{u_{n+1}} + \frac{3000}{u_n u_{n+1}} \tag{19}$$
et les conditions initiales $u_0 := 11/2$ et $u_1 := 61/11$.
Nous verrons en 1.e) que u_n ne s'annule jamais et que la suite est bien définie. Nous l'admettons provisoirement.
   a) Montrer que tous les termes de la suite u_n sont rationnels. Quelles sont les limites (finies) possibles d'une suite récurrente définie par (19) ?
   b) * Calculer *numériquement* une vingtaine termes de la suite (u_n) en utilisant un ordinateur ou une calculette programmable et vérifier que cette suite « semble tendre vers 100 ».
      * Calculer *formellement* une trentaine de termes de la suite (u_n) sous forme de fraction (quotient d'entiers) en utilisant un logiciel de calcul formel, les évaluer sous forme décimale approchée et vérifier que cette suite « semble tendre vers 6 ».
      * Soit $(\tilde{u}_n)$ la suite réelle définie par la relation de récurrence (19) et $\tilde{u}_0 = u_0 = 5.5$, $\tilde{u}_1 := 55454545454545454545/10^{20}$. Procéder comme ci-dessus pour cette suite, en calculant *formellement* une vingtaine de termes et vérifier que cette suite « semble tendre vers 100 ».
   c) Vérifier qu'il se passe des phénomènes similaires si l'on remplace les « conditions initiales » par $u_0 := 6$ et $u_1 := 6$, puis par $\tilde{u}_0 := 6$ et $\tilde{u}_1 := 6{,}000000000003$.
   Vérifier que la situation est différente pour $u_0 := 5$ et $u_1 := 5$.
   d) Faire un bilan des expériences de calcul faites ci-dessus.
   e) On revient aux conditions initiales $u_0 := 11/2$ et $u_1 := 61/11$. Montrer que, pour tout $n \in \mathbb{N}$, on a $u_n = \frac{5^{n+1}+6^{n+1}}{5^n+6^n}$ et que la récurrence est bien définie. Calculer la limite de la suite (u_n).

2. a) On considère les suites réelles (w_n) dont tous les termes sont non nuls. Montrer que la suite définie par $(v_n := w_{n+1}/w_n)_{n\in\mathbb{N}}$ vérifie la récurrence (19) si, et seulement si, la suite (w_n) vérifie la récurrence *linéaire* à 3 termes :

$$w_{n+3} = 111w_{n+2} - 1130w_{n+1} + 3000w_n. \tag{20}$$

   b) Montrer que l'ensemble E des suites solutions de la récurrence (20) est un sous-espace vectoriel *de dimension* 3 de l'espace des suites réelles. Calculer $\rho \in \mathbb{R}$ pour que la suite (ρ^n) appartienne à E et en déduire une base de E.

   c) Donner une *formule* pour v_n quand la suite (v_n) vérifie la récurrence (19). Montrer qu'une telle suite est toujours convergente.

   d) Expliquer les résultats numériques de 1.b) et 1.c).

3. Construire une famille de suites récurrentes vérifiant des propriétés analogues à celles mises en évidence pour (u_n) en remplaçant $5, 6, 100$ par des *entiers* arbitraires p_1, p_2, p_3 tels que $0 < p_1 < p_2 < p_3$.

IV.1.63 $**$ *Cycles de l'application de Collatz et approximations rationnelles.* L'exercice porte sur l'application, dite de Collatz, $f : \mathbb{Z} \to \mathbb{Z}$ définie dans l'exercice IV.1.29 : $f(m) = m/2$ si m est pair et $f(m) = 3m + 1$ si m est impair.
On *suppose* que f possède un cycle positif non trivial (c'est-à-dire ne contenant pas 1) A. On note p (resp. q) le nombre d'éléments pairs (resp. impairs) de A et a le plus petit élément de A.

1. a) On note $c_1, \ldots, c_q$ les éléments impairs de A. Montrer que :

$$2^p = \left(3 + \frac{1}{c_1}\right)\left(3 + \frac{1}{c_2}\right)\cdots\left(3 + \frac{1}{c_q}\right).$$

   En déduire $3^q < 2^p \leqslant \left(3 + \frac{1}{a}\right)^q$, puis $\dfrac{\log_{10} 3}{\log_{10} 2} < \dfrac{p}{q} < \dfrac{\log_{10}\left(3 + \frac{1}{a}\right)}{\log_{10} 2}$.

   b) On suppose que $\operatorname{card} A = 18$. Montrer que $3^q < 2^{18-q} < 4^q$. Montrer que $q \geqslant 7$ et $q \leqslant 5$ sont impossibles. Montrer que $q = 6$ est impossible. Conclure.

   c) On suppose dans toute la suite que l'on sait que toute suite récurrente définie par f et $u_0 \in \mathbb{N}^*$ tel que $u_0 \leqslant 5 \times 10^{18}$ prend la valeur 1 (un résultat vérifié sur ordinateur par le mathématicien portugais Tomas Oliveira e Silva, *cf.* l'exercice IV.1.29).

   Montrer que : $\dfrac{\ln 3}{\ln 2} = \dfrac{\log_{10} 3}{\log_{10} 2} < \dfrac{p}{q} < \dfrac{\log_{10}\left(3 + 2 \times 10^{-19}\right)}{\log_{10} 2}$.

   En déduire, en utilisant un logiciel de calcul numérique que :

$$p/q \in J := \left]1, \underline{58496250072115618}45, 1, \underline{58496250072115618}55\right[. \tag{21}$$

2. *Cette question et la suivante sont indépendantes de l'application de Collatz. Elles concernent l'approximation rationnelle des nombres réels.*
   On appelle *paire de Farey* la donnée de 4 rationnels strictement positifs r_1, s_1, r_2, s_2 tels que $s_1 r_2 - s_2 r_1 = 1$. On note $\langle r_1/s_1, r_2/s_2\rangle$ ou $\left\langle \frac{r_1}{s_1}, \frac{r_2}{s_2}\right\rangle$ une telle paire (par exemple $\langle 3/2, 5/3\rangle$). Soit $\langle r_1/s_1, r_2/s_2\rangle$ une paire de Farey *fixée dans toute cette question*.

   a) Montrer que les fractions r_1/s_1 et r_2/s_2 sont *irréductibles* et que $r_1/s_1 < r_2/s_2$.

b) Montrer que, pour un rationnel sous forme irréductible r/s, les deux conditions suivantes sont équivalentes.

(i) On a $r_1/s_1 < r/s < r_2/s_2$.

(ii) Il existe des entiers strictement positifs premiers entre eux λ et μ tels que

$$r = \lambda r_1 + \mu r_2 \qquad \text{et} \qquad s = \lambda s_1 + \mu s_2.$$

c) Montrer que si p' et q' sont des nombres rationnels strictement positifs tels que $r_1/s_1 < p'/q' < r_2/s_2$, alors $p' \geqslant r_1 + r_2$ et $q' \geqslant s_1 + s_2$.

d) Montrer que $\left\langle \frac{r_1}{s_1}, \frac{r_1+r_2}{s_1+s_2} \right\rangle$ et $\left\langle \frac{r_1+r_2}{s_1+s_2}, \frac{r_2}{s_2} \right\rangle$ sont des paires de Farey.

3. Trouver une paire de Farey $\langle r_1/s_1, r_2/s_2 \rangle$ telle que $355 = r_1 + r_2$ et $113 = s_1 + s_2$ (on pourra utiliser l'algorithme d'Euclide). Montrer que si p' et q' sont des rationnels strictement positifs tels que $|p'/q' - \pi| < 8 \times 10^{-5}$, alors $p' \geqslant 355$ et $q' \geqslant 113$ (on admettra l'approximation : $|\pi - 3{,}14159265| < 10^{-8}$).

4. Pour cette question on utilisera un logiciel de calcul numérique ou une calculette « suffisamment performante ».
On note : $r_1 = 357638239,\ s_1 = 225644606,\ r_2 = 272500658,\ s_2 = 171928773.$

a) Vérifier que $\langle r_1/s_1, r_2/s_2 \rangle$ est une paire de Farey. Montrer que si p' et q' sont des rationnels strictement positifs tels que $r_1/s_1 < p'/q' < r_2/s_2$, alors $p' \geqslant 630138897$ et $q' \geqslant 397573379$.

b) Vérifier que l'intervalle J défini dans la question 1. c) (*cf.* (21)) est contenu dans l'intervalle ouvert $]r_1/s_1, r_2/s_2[$.

5. Montrer que si A est un cycle positif non trivial de f, alors $\operatorname{card}(A) \geqslant 1027712276$.

6. Cette question est indépendante des questions 2, 3, 4 et 5. On *suppose* que f possède une famille *infinie* de cycles positifs deux à deux distincts $\{A_n\}_{n\in\mathbb{N}}$. On note p_n (resp. q_n) le nombre d'éléments pairs (resp. impairs) de A_n et a_n le plus petit élément de A_n.

a) Montrer que l'on peut renuméroter les A_n de telle sorte que la suite (a_n) soit croissante. Montrer qu'alors $\lim\limits_{n\to+\infty} p_n/q_n = \ln 3/\ln 2$ et que $\lim\limits_{n\to+\infty} \operatorname{card} A_n = +\infty$.

b) Montrer que, pour tout $m \in \mathbb{N}^*$, $m \geqslant 2$, l'ensemble des m-cycles positifs de f est *fini*.

IV.1.64 $**$ *Un théorème de Sarkovski* Le but de l'exercice est de donner une preuve de la proposition 52 de la page 577. On utilisera l'exercice 29 de la page 554 et l'exercice 7 de la page 639.
Soient I un segment de $\mathbb{R}$ et $f : I \to I$ une fonction continue. On suppose que f admet un 3-cycle.

1. Montrer que f admet un point fixe

2. Montrer que f admet un 2-cycle.

3. Montrer que, pour tout $\tau \in \mathbb{N}^*$, $\tau \geqslant 4$, f admet un τ-cycle.

IV.1.65 $*$ *Nombres de Champernowne. Nombres univers.*

On appelle *suite univers* toute suite *infinie* de chiffres $0, 1, 2, \ldots, 9$: $a := (\alpha_1, \ldots, \alpha_h, \ldots)$, telle que toute suite *finie* de chiffres $S := (s_1, \ldots, s_m)$ y figure au moins une fois. C'est à dire qu'il existe $k \in \mathbb{N}^*$ tel que $(\alpha_k, \alpha_{k+1}, \ldots, \alpha_{k+m-1}) = (s_1, s_2, \ldots, s_m)$.

Étant données deux écritures décimales d'entiers (vues comme chaînes de caractères) : $\alpha_1 \ldots \alpha_k$ et $\beta_1 \ldots \beta_h$ on peut en fabriquer une nouvelle en les mettant bout à bout : $\alpha_1 \ldots \alpha_k \beta_1 \ldots \beta_h$; on dit que l'on a *concaténé* les deux écritures. Par exemple, la concaténation de 2585 et de 1098835 est 25851098835.

On définit une écriture décimale c en concaténant successivement, dans l'ordre croissant, les écritures décimales de *tous* les entiers naturels, puis en mettant une virgule après le premier 0 :

$$c = 0{,}1234567891011121314151517181920\ldots9899100101\ldots$$

1. a) Montrer que c définit un nombre réel. Ce nombre s'appelle nombre de Champernowne[31]
   b) Montrer que l'écriture décimale $0, S$, où S est une suite univers, définit un nombre réel. On dit que c'est un *nombre univers*. Donner des exemples de nombres univers.
   c) Montrer que la somme de deux nombres univers n'est pas nécessairement un nombre univers.
2. a) Montrer que la suite des décimales de c contient *une infinité de fois* toute suite finie donnée formée de chiffres $0, 1, 2, \ldots, 9$.
   b) Montrer que toute suite univers contient *une infinité de fois* toute suite finie donnée formée de chiffres $0, 1, 2, \ldots, 9$.
   c) Montrer que, si l'on change un nombre fini de décimales d'un nombre univers, on obtient encore un nombre univers.
   d) Montrer que tout nombre univers est irrationnel.
3. Pour tout entier $b \geqslant 2$, on peut définir des nombres de Champernowne c_b en utilisant l'écriture en base b au lieu de l'écriture décimale (ainsi $c_{10} = c$). Écrire les 25 premières décimales de c_2. Esquisser une preuve de l'irrationalité des nombre c_b.
4. On peut lire ici ou là « en surfant sur la toile » que tout nombre univers contient les oeuvres complètes de Shakespeare, celles de Victor Hugo, les symphonies de Mozart, un portrait du lecteur et la bibliothèque de Babel de la nouvelle de l'écrivain argentin Jorge Luis Borges... Peut-on justifier ces affirmations ?

IV.1.66 $**$ *Exemples d'applications « chaotiques ».*

1. Pour tout $t \in \mathbb{R}$, on note $E(t)$ sa partie entière. On définit une application $f : [0, 1[\to [0, 1[$ par $f(t) = 10t - E(10t)$.
   (i) a) Décrire f en utilisant l'écriture décimale de t. Décrire la $n^{ème}$ itérée $f^{\circ n} = f \circ \ldots \circ f$ de f et montrer que $f^{\circ n}(t) = 10^n t - E(10^n t)$.
   b) Soit $t_0 \in [0, 1[$.
   Montrer que, pour tout $u \in [0, 1[$ et tout $\varepsilon > 0$, il existe $v \in]t_0 - \varepsilon, t_0 + \varepsilon[\cap [0, 1[$ et $n \in \mathbb{N}^*$ tel que $f^{\circ n}(v) = u$. En déduire que le comportement asymptotique des suites récurrentes définies par f est « très sensible à la condition initiale ».

[31] Il a été introduit par D. G. Champernowne, alors étudiant à l'université de Cambridge, en 1933.

(ii) a) Montrer que la suite (u_n) définie par u_0 et f est constante et égale à 0 à partir d'un certain rang si, et seulement si, $u_0 \in \mathbb{D}$. En déduire que l'ensemble des $u_0 \in [0,1[$ tels que la suite (u_n) soit constante à partir d'un certain rang est dense dans $[0,1[$. Décrire cet ensemble.

b) Montrer que la suite (u_n) définie par u_0 et f est périodique à partir d'un certain rang si, et seulement si, $u_0 \in \mathbb{Q}$.

c) Soit $m \in \mathbb{N}^*$. Montrer qu'il existe $a \in [0,1[$ dont l'écriture décimale est périodique de *plus petite période* m. Montrer que la suite (u_n) définie par $u_0 := a$ et f est périodique de plus petite période m.

d) Soit $m \in \mathbb{N}^*$. Montrer que l'ensemble des $u_0 \in [0,1[$, tels que la suite (u_n) soit périodique, de plus petite période m, à partir d'un certain rang, est dense dans $[0,1[$.

e) Soit $t_0 \in [0,1[$. Montrer que, pour tout $\varepsilon > 0$, il existe $u_0 \in]t_0-\varepsilon, t_0+\varepsilon[\cap[0,1[$ tel que la suite définie par u_0 soit *périodique*.

(iii) Soit $c \in [0,1[$ le nombre de Champernowne (*cf.* l'exercice IV.1.65).
On définit une suite (u_n) par $u_0 := c$ et $u_{n+1} = f(u_n)$.

a) Soit $t = 0,\alpha_1\alpha_2\ldots\alpha_h\ldots$ une écriture décimale.
Montrer que, pour tout $k \in \mathbb{N}^*$, il existe un entier n_k tel que la troncature à l'ordre k de $f^{\circ n_k}(c)$ soit $(f^{\circ n_k}(c))_k = t_k = 0,\alpha_1\alpha_2\ldots\alpha_k$.

b) Montrer que l'image de la suite (u_n) est *dense* dans $[0,1[$.

(iv) Préciser la propriété de sensibilité aux conditions initiales pour les suites récurrentes définies par f mise en évidence en (i) b).

2. On note $U := \{z \in \mathbb{C} \mid |z| = 1\}$ et l'on définit une application $\varphi : \begin{array}{ccc} U & \longrightarrow & U \\ z & \longmapsto & z^{10}. \end{array}$

On considère les suites (z_n) de U définies par récurrence par $z_{n+1} = \varphi(z_n) = z_n^{10}$. Montrer que l'étude de ces suites se déduit de celle des suites étudiées dans la question 1. En déduire des propriétés de sensibilité aux conditions initiales, de périodicité et de densité pour les suites (z_n).

3. Soit $b \in \mathbb{N}^*$, $b \geqslant 2$. On remplace, dans la question précédente l'application φ par l'application $\varphi_b : U \to U$ définie par $z \in U \mapsto z^b$. Mettre en évidence des propriétés des suites récurrentes définies par φ_b.

Fonctions réelles IV.2

Les fonctions réelles sont l'une des briques fondamentales de l'analyse. Nous nous limiterons dans ce module introductif au cas le plus simple, celui des fonctions réelles *d'une seule variable*.
L'une des questions essentielles relatives aux fonctions réelles est l'étude de leur « régularité » : sont-elles continues, dérivables ? Ces propriétés sont définies et étudiées dans ce module. Elles ont de nombreuses conséquences présentées sous la forme de théorèmes généraux et d'applications.

1 Limites et continuité

1.1 Généralités

Définition 1. On appelle *fonction réelle de la variable réelle* toute application f d'un sous-ensemble non vide de $\mathbb{R}$, et à valeurs dans $\mathbb{R}$. L'ensemble de départ est appelé *ensemble de définition de* f, on le note D_f.

Dans tout ce module on dira aussi tout simplement *fonction*. Pour définir une application il faut au préalable définir son ensemble de départ (ici appelé ensemble de définition) et son ensemble d'arrivée (ici $\mathbb{R}$), puis donner la règle permettant d'associer à tout élément de l'ensemble de départ un élément de l'ensemble d'arrivée. En pratique, pour définir une fonction réelle f, on part souvent d'une « formule $f(x) =?$ » et l'on cherche un ensemble de réels x pour lesquels cette formule a un sens. Une fois obtenu un tel ensemble, on le note D_f et l'on *définit* $f : x \in D_f \mapsto f(x) \in \mathbb{R}$. Pour trouver D_f, on tient compte du fait que les divisions par zéro sont interdites, du fait que les racines carrées doivent avoir un argument positif ou nul, etc. Par exemple la formule $f(x) = \frac{\sqrt{1-x^2}}{x}$ permet de définir une fonction $f : x \mapsto \dfrac{\sqrt{1-x^2}}{x}$ sur l'ensemble $D_f := [-1, 0[\cup]0, +1]$.
On notera qu'une fonction n'est pas nécessairement définie par une formule. Ainsi la *fonction de Dirichlet* (fonction caractéristique de $\mathbb{Q}$) donnée par :

$$\mathrm{Dir}(x) := 1_{\mathbb{Q}} = \begin{cases} 1 & \text{si } x \in \mathbb{Q} \\ 0 & \text{si } x \notin \mathbb{Q} \end{cases} \tag{1}$$

est définie sur $\mathbb{R}$, mais ne l'est pas par une « formule ».

1.1.1 Parties paire et impaire d'une fonction

Soit $f : D_f \to \mathbb{R}$ une fonction réelle. Dans tout ce paragraphe on suppose que D_f est symétrique par rapport à 0, c'est-à-dire $x \in D_f$ si, et seulement si, $-x \in D_f$.
On rappelle que l'on dit que f est

- *paire* si, pour tout $x \in D_f$, on a $f(-x) = f(x)$;
- *impaire* si, pour tout $x \in D_f$, on a $f(-x) = -f(x)$.

Proposition et définition 1. Soit $f : D_f \to \mathbb{R}$ une fonction réelle. Il existe un couple de fonctions (f_0, f_1) déterminé de manière unique telle que : f_0 soit paire, f_1 soit impaire et $f = f_0 + f_1$. On dit que f_0 est la *partie paire* de f et que f_1 est sa *partie impaire*. On a alors, pour tout $x \in D_f$:

$$f_0(x) = \frac{1}{2}\left(f(x) + f(-x)\right) \qquad \text{et} \qquad f_1(x) = \frac{1}{2}\left(f(x) - f(-x)\right).$$

Démonstration. Supposons que f_0 soit paire, que f_1 soit impaire et que $f = f_0 + f_1$. On a $f(-x) = f_0(x) - f_1(x)$, donc :

$$f_0(x) = \frac{1}{2}\big(f(x) + f(-x)\big) \qquad \text{et} \qquad f_1(x) = \frac{1}{2}\big(f(x) - f(-x)\big).$$

D'où l'unicité. Inversement :

$$f(x) = \frac{1}{2}\big(f(x) + f(-x)\big) + \frac{1}{2}\big(f(x) - f(-x)\big);$$

la fonction $x \mapsto \frac{1}{2}\big(f(x) + f(-x)\big)$ est paire et la fonction $x \mapsto \frac{1}{2}\big(f(x) - f(-x)\big)$ est impaire. ∎

1.1.2 Parties positive et négative d'une fonction

Proposition et définition 2. Soit $f : D_f \to \mathbb{R}$ une fonction réelle. On définit, pour tout $x \in D_f$:

$$f^+(x) := \max\{f(x), 0\} \quad \text{et} \quad f^-(x) := \max\{-f(x), 0\} = -\min\{f(x), 0\}.$$

On dit que f^+ est *la partie positive* de f et f^- sa *partie négative*. On a :

$$f^+ \geqslant 0, \;\; f^- \geqslant 0, \quad f = f^+ - f^-, \quad |f| = f^+ + f^-.$$

Les vérifications sont faciles et laissées au lecteur. On peut obtenir les graphes de f^+ et f^- en utilisant le graphe de f, le lecteur le fera à titre d'exercice.

1.1.3 Fonctions périodiques

Soit $f : D_f \to \mathbb{R}$. S'il existe un nombre $\tau > 0$ tel que pour tout $x \in D_f$, on ait $x \pm \tau \in D_f$ et $f(x + \tau) = f(x)$, on dit que la fonction est *périodique*. Un tel réel τ est appelé *une période* de f et l'on dit aussi que f est τ-périodique. Si τ est une période de f, pour tout $n \in \mathbb{N}^*$, $n\tau$ est aussi une période. Lorsqu'il existe une *plus petite période* on l'appelle

la période de f. Une telle plus petite période n'existe pas toujours : par exemple, une fonction constante admet tous les réels strictement positifs comme période, donc n'en a pas de plus petite. Voici un autre exemple (moins trivial) :

Exercice 1.
Quelles sont les périodes de la fonction de Dirichlet (*cf.* ci-dessus page 619) ? Y en a-t-il une plus petite ?
Solution. Si τ est une période de la fonction Dir de Dirichlet, alors on doit avoir $\mathrm{Dir}(\tau) = \mathrm{Dir}(0) = 1$, donc $\tau \in \mathbb{Q}$. Réciproquement, si $\tau \in \mathbb{Q}$, alors pour $x \in \mathbb{R}$, les réels x et $x+\tau$ sont ou bien tous les deux rationnels, ou bien tous les deux irrationnels, donc $\mathrm{Dir}(x) = \mathrm{Dir}(x+\tau)$.
L'ensemble des périodes de la fonction de Dirichlet est donc $\mathbb{Q}_+^*$ et il n'a pas de plus petit élément.

1.2 Limite d'une fonction

1.2.1 Définitions et premières propriétés

Pour parler de limite d'une fonction en un point a, il faut que la fonction soit définie en des points « aussi proches de a que l'on veut ». On dira alors que « f est *définie au voisinage de* a » (on notera que la terminologie peut être différente dans d'autres ouvrages).

Définition 2. Soit $a \in \mathbb{R}$ ou $a = \pm\infty$. On dit qu'une fonction f est *définie au voisinage de* a s'il existe *une* suite à valeurs dans $D_f \setminus \{a\}$ admettant a pour limite[1].

C'est le cas, en particulier, si f est définie sur un intervalle I non réduit à un point (ou plus généralement sur un domaine contenant un tel intervalle) *et* si l'on a l'une des situations suivantes :

- a est un réel appartenant à I ;
- a est une borne réelle de I ;
- $a = +\infty$ et I est un intervalle non majoré ;
- $a = -\infty$ et I est un intervalle non minoré.

Dans la pratique, on se limitera le plus souvent à ces cas.
Par exemple la fonction définie par la formule $\frac{\sin x}{x}$ admet $D_f :=]-\infty, 0[\cup]0, +\infty[$ comme domaine de définition et la fonction $f : x \in D_f \mapsto \frac{\sin x}{x}$ est définie au voisinage de tout point de $\mathbb{R}$, y compris 0.

Remarque. On vérifie qu'une fonction $f : D_f \to \mathbb{R}$ est définie au voisinage de $a \in \mathbb{R}$ si, et seulement si, la condition suivante est satisfaite :

$$\forall \varepsilon > 0,\ \exists x \in D_f \setminus \{a\},\ \text{tel que } |x-a| \leqslant \varepsilon. \tag{2}$$

En effet, si f est définie au voisinage de a, la condition (2) est satisfaite d'après la définition de la limite d'une suite (*cf.* la définition 19). Inversement, si (2) est vraie,

[1] On dit aussi dans ce cas que a est *adhérent* à $D_f \setminus \{a\}$.

il existe une suite prenant ses valeurs dans $D_f \setminus \{a\}$ et convergeant vers a : en effet, pour tout $n \in \mathbb{N}$, il existe un élément de $D_f \setminus \{a\}$, que l'on peut noter x_n, tel que $|x_n - a| \leqslant \frac{1}{n+1}$. La suite (x_n) ainsi construite converge alors vers a.

Définition 3. Soient $a \in \mathbb{R} \cup \{\pm\infty\}$ et f définie au voisinage de a. On dit que f *admet une limite* $\ell \in \mathbb{R} \cup \{\pm\infty\}$ *au point* a si, *pour toute suite* $(x_n)_{n\in\mathbb{N}}$ prenant ses valeurs dans $D_f \setminus \{a\}$ et de limite a, la suite $\big(f(x_n)\big)_{n\in\mathbb{N}}$ tend vers ℓ.

Il est bon d'insister sur la condition « *pour toute suite* » dans la définition de la limite. Bien évidemment, il ne suffit pas de vérifier que la limite existe pour une suite particulière.
On dit, par un abus de langage traditionnel, que $f(x)$ tend vers ℓ (et l'on note $f(x) \to \ell$) lorsque x tend vers a.
Noter que, lorsque a est réel, l'on ne suppose pas $a \in D_f$, et que l'on ne considère que les suites telles que $x_n \in D_f$ et $x_n \neq a$. Avec notre définition, si $a \in D_f$, on n'a pas en général $\ell = f(a)$. Dans d'autres ouvrages ou cours polycopiés les conventions peuvent être différentes, le lecteur est donc invité à lire soigneusement les définitions. En particulier, certains auteurs donnent une notion de limite *plus restrictive* que la nôtre en considérant toutes les suites $(x_n)_{n\in\mathbb{N}}$ prenant leurs valeurs dans D_f et de limite a *sans exclure la possibilité de* $x_n = a$. Ainsi, si $a \notin D_f$ rien n'est changé, mais par contre si $a \in D_f$, on a nécessairement, avec cette autre notion de limite, $\ell = f(a)$ (prendre une suite constante égale à a).

Proposition 3. Lorsqu'une fonction f, définie au voisinage de a, admet une limite ℓ en a, celle-ci est unique. On l'appelle *la* limite de f en a et l'on écrit alors $\ell = \lim\limits_{a} f$ ou $\ell = \lim\limits_{x\to a} f(x)$ ou $f(x) \underset{x\to a}{\longrightarrow} \ell$.

Démonstration. Puisque f est définie au voisinage de a, il existe une suite (x_n) d'éléments de $D_f \setminus \{a\}$ tendant vers a. Fixons une telle suite. Si f admet deux limites ℓ et ℓ' en a, alors par définition la suite $\big(f(x_n)\big)$ doit tendre vers ℓ et vers ℓ'. L'unicité de la limite d'une suite donne alors $\ell = \ell'$. ∎

Point Méthode

Lorsque l'on veut montrer qu'une fonction n'a pas de limite en un point, il est souvent utile d'utiliser directement la définition. Par négation, il suffit de trouver deux suites (x_n) et (y_n), de limite a, telles que $\big(f(x_n)\big)$ et $\big(f(y_n)\big)$ aient des limites différentes.

Exemples.

1. On a $\lim\limits_{x\to 0} x \sin\frac{1}{x} = 0$. En effet, pour toute suite (x_n) de $\mathbb{R}\setminus\{0\}$, de limite 0, la suite $\sin\frac{1}{x_n}$ est bornée et la proposition 32 de la page 546 permet de conclure.

2. La fonction de Dirichlet vue plus haut n'a de limite en aucun point. En effet, tout réel a est la limite d'une suite de rationnels (x_n) par densité de $\mathbb{Q}$ dans $\mathbb{R}$. Mais il est aussi limite d'une suite d'irrationnels (y_n) (prendre par exemple $y_n = x_n + \frac{\sqrt{2}}{n+1}$). On a alors $\mathrm{Dir}(x_n) = 1$ et $\mathrm{Dir}(y_n) = 0$, donc les limites sont différentes.
3. Partant de la formule $f(x) = \sin\frac{1}{x}$, on définit $D_f := \mathbb{R}\backslash\{0\}$ et $f : x \in D_f \mapsto \sin\frac{1}{x}$. On considère les deux suites $\left(x_n := \frac{1}{(n+1)\pi}\right)$ et $\left(y_n := \frac{2}{\pi+4n\pi}\right)$ de D_f. Elles tendent vers 0 et l'on a $f(x_n) = \sin(n+1)\pi = 0$ et $f(y_n) = \sin(\pi/2 + 2n\pi) = 1$, donc la suite $(f(x_n))$ est constante et égale à 0 et la suite $(f(y_n))$ est constante et égale à 1. Par suite f n'a pas de limite en 0. À titre d'exercice le lecteur pourra vérifier que, pour tout $u \in [-1, 1]$, il existe une suite (z_n) de D_f telle que la suite $\big(f(z_n)\big)$ soit constante et égale à u et il essaiera de dessiner le graphe de f.

Proposition 4. Soient $a \in \mathbb{R}$, $\ell \in \mathbb{R} \cup \{\pm\infty\}$ et f, g deux fonctions définies au voisinage de a, avec $D_f = D_g = D$ et telles que $f = g$ au voisinage de a, c'est-à-dire qu'il existe $\varepsilon > 0$ tel que les restrictions de f et g à $D \cap [a-\varepsilon, a+\varepsilon]$ coïncident. Alors f admet ℓ pour limite en a si, et seulement si, g admet ℓ pour limite en a.

Démonstration. C'est une conséquence immédiate de la proposition 26 de la page 543. Si la suite (x_n) prend ses valeurs dans $D \setminus \{a\}$, alors, pour N assez grand, la suite tronquée en N prend ses valeurs dans $D \cap [a-\varepsilon, a+\varepsilon] \setminus \{a\}$. Par hypothèse, les images par f et par g de la suite tronquée coïncident et, par ailleurs, une suite et toutes ses suites tronquées ont la même limite, d'où le résultat. ∎

Il y a une variante de la proposition ci-dessus si $a = \pm\infty$. Le lecteur l'écrira et la prouvera à titre d'exercice.
On dit que la notion de limite d'une fonction ne dépend que du comportement de la fonction « au voisinage de a » et que cette notion est « locale ».

Proposition 5. Soient $a, \ell \in \mathbb{R}$ et f, g deux fonctions définies au voisinage de a, avec $D_f = D_g = D$. On suppose que g admet la limite 0 en a et que $|f - \ell| \leqslant g$. Alors f admet la limite ℓ en a.

Démonstration. Soit (x_n) une suite d'éléments de $D \setminus \{a\}$ tendant vers a. On a, par hypothèse, $\lim\limits_{n\to+\infty} g(x_n) = 0$ donc $\lim\limits_{n\to+\infty} \big|f(x_n) - \ell\big| = 0$ et $\lim\limits_{n\to+\infty} f(x_n) = \ell$. ∎

Exercice 2.

On suppose que f admet la limite ℓ en a. Montrer que $|f|$ admet la limite $|\ell|$ en a.

Solution. On a, pour tout $x \in D_f$, $\big||f(x)| - |\ell|\big| \leqslant \big|f(x) - \ell\big|$. On vérifie que $|f - \ell|$ tend vers 0 en a. Le résultat s'en déduit en utilisant la proposition ci-dessus.

Soit $f : D_f \to \mathbb{R}$ définie au voisinage de a. Il se peut que la limite de la suite $(f(x_n))$ n'existe que pour les suites (x_n) de limite a telles que $x_n > a$ (resp. $x_n < a$). On parle alors de *limite à droite* (resp. à gauche).

Définition 4. Soit $f : D_f \to \mathbb{R}$ définie au voisinage de a. On dit que f admet une limite à droite au point a si la *restriction* de f à $D_f \cap [a, +\infty[$ est définie au voisinage de a et admet une limite au point a. On note alors $\lim\limits_{x \overset{>}{\to} a} f(x)$ ou encore $\lim\limits_{x \to a^+} f(x)$.

On dit que f admet une limite à gauche au point a si la *restriction* de f à $D_f \cap]-\infty, a]$ est définie au voisinage de a et admet une limite au point a.
On note alors $\lim\limits_{x \overset{<}{\to} a} f(x)$ ou $\lim\limits_{x \to a^-} f(x)$.

On vérifie que la restriction de f à $D_f \cap [a, +\infty[$ (resp. $D_f \cap]-\infty, a]$) est définie au voisinage de a si, et seulement si, a est limite d'une suite (x_n) de D_f telle que $\forall n \in \mathbb{N}\,,\ x_n > a$ (resp. $\forall n \in \mathbb{N}\,,\ x_n < a$).
Une suite de $D_f \setminus \{a\}$ et de limite a peut toujours être « séparée » en termes strictement inférieurs à a et strictement supérieurs à a (avec éventuellement aucun terme dans l'une des catégories). On obtient ainsi une ou deux suites extraites de limite a. De ce fait, si $D_f \cap [a, +\infty[\neq \varnothing$ et $D_f \cap [-\infty, a[\neq \varnothing$, une fonction f a une limite en a si, et seulement si, elle a une limite à gauche et une limite à droite, et si ces deux limites sont les mêmes.
Nous avons vu, dans le module IV.1 que pour chaque réel x, il existait un unique entier (relatif) inférieur à x et plus grand que les autres, appelé *partie entière de* x et noté $\mathrm{E}(x)$ ou $\lfloor x \rfloor$; il vérifie $\lfloor x \rfloor \leqslant x < \lfloor x \rfloor + 1$ pour tout x. Cela définit sur $\mathbb{R}$ une fonction également appelée *partie entière*.

Exemple. La fonction partie entière $x \mapsto \lfloor x \rfloor$ a une limite à gauche et une limite à droite en tout point a, différentes si $a \in \mathbb{Z}$ puisqu'alors $\lim\limits_{x \to a^+} \lfloor x \rfloor = a$ et $\lim\limits_{x \to a^-} \lfloor x \rfloor = a - 1$.

Théorème 6 (Caractérisation de Weierstraß). Une fonction $f : D_f \to \mathbb{R}$ définie au voisinage de a admet une limite $\ell \in \mathbb{R}$ au point a si, et seulement si :

$$\forall \varepsilon > 0\,,\ \exists \eta > 0\,,\ \forall x \in D_f \setminus \{a\}\,,\ |x - a| \leqslant \eta \ \Rightarrow\ |f(x) - \ell| \leqslant \varepsilon. \tag{3}$$

La condition (3) est dite *caractérisation de Weierstraß* des limites.

Démonstration. Considérons d'abord une fonction satisfaisant (3), et montrons qu'elle admet ℓ pour limite au sens de la définition 3 de la page 622. On considère donc une suite (x_n) de limite a avec $x_n \in D_f \setminus \{a\}$, et l'on cherche à prouver que $f(x_n)$ tend vers ℓ.
Soit $\varepsilon > 0$ donné. Il existe un réel $\eta > 0$ satisfaisant à la relation (3). Comme x_n tend vers a, il existe $N \in \mathbb{N}$ tel que pour tout $n \geqslant N$, $|x_n - a| \leqslant \eta$. Mais dans ce cas, $|f(x_n) - \ell)| \leqslant \varepsilon$ d'après (3). Cela prouve la convergence de $f(x_n)$ vers ℓ.
Inversement, supposons que f admette ℓ pour limite au point a, au sens de la définition 3 de la page 622, et montrons qu'elle vérifie alors la propriété (3). On raisonne par l'absurde en supposant le contraire, c'est-à-dire :

$$\exists \varepsilon > 0,\ \forall \eta > 0,\ \exists x \in D_f \setminus \{a\},\ \text{tel que}\ |x - a| \leqslant \eta \quad \text{et} \quad |f(x) - \ell| > \varepsilon.$$

Pour chaque valeur de $\eta > 0$, on a donc un point $x \in D_f \setminus \{a\}$ correspondant. Notons x_n un tel x correspondant à $\eta = 1/n$ lorsque $n \in \mathbb{N}^*$.

Alors $|x_n - a| \leqslant 1/n$, donc la suite (x_n) tend vers a. Mais $|f(x_n) - \ell)| > \varepsilon > 0$, donc $f(x_n)$ ne peut tendre vers ℓ. Cela contredit l'hypothèse faite sur f. ■

On a un résultat analogue pour les limites infinies. Il suffit d'adapter la preuve du théorème 6 ; le lecteur le fera à titre d'exercice.

Théorème 7 (Caractérisation de Weierstraß).

(i) Une fonction $f : D_f \to \mathbb{R}$ définie au voisinage de a admet la limite $+\infty$ au point a si, et seulement si :

$$\forall A > 0\,,\ \exists \eta > 0\,,\ \forall x \in D_f \setminus \{a\}\,,\ |x - a| \leqslant \eta \ \Rightarrow\ f(x) \geqslant A. \tag{4}$$

(ii) Une fonction $f : D_f \to \mathbb{R}$ définie au voisinage de a admet la limite $-\infty$ au point a si, et seulement si :

$$\forall A < 0\,,\ \exists \eta > 0\,,\ \forall x \in D_f \setminus \{a\}\,,\ |x - a| \leqslant \eta \ \Rightarrow\ f(x) \leqslant A. \tag{5}$$

Point Méthode

Selon les problèmes à résoudre, on utilise la définition séquentielle des limites (définition 3) ou la caractérisation de Weierstraß. Pour montrer qu'une fonction « définie par une formule » admet une limite et la calculer, il est en général maladroit d'utiliser la caractérisation de Weierstrass. On peut utiliser la définition séquentielle mais il vaut mieux appliquer les règles d'opérations sur les limites de la proposition 11 énoncée plus loin.

1.2.2 Limites et propriétés au voisinage d'un point

Nous allons voir que la connaissance de la limite d'une fonction au point a permet d'en déduire des propriétés de cette fonction *au voisinage de* a.

Définition 5. On dit qu'une fonction f possède une propriété (P) au voisinage de $a \in \mathbb{R}$ si elle est définie au voisinage de a et s'il existe $\varepsilon > 0$ tel que la restriction de f à $D_f \cap [a - \varepsilon, a + \varepsilon]$ possède la propriété (P).

On a une définition analogue pour $a = +\infty$ (resp. $a = -\infty$) en remplaçant $\exists \varepsilon > 0$ par $\exists A \in \mathbb{R}$ et $D_f \cap [a - \varepsilon, a + \varepsilon]$ par $D_f \cap [A, +\infty[$ (resp. $D_f \cap]-\infty, A[$).
On pourra ainsi parler de fonction positive (resp. négative) au voisinage de a...

Proposition 8. Soient $\ell \in \mathbb{R}$, $a \in \mathbb{R} \cup \{\pm\infty\}$ et une fonction f tendant vers ℓ en a.

(i) Soit $m \in \mathbb{R}$, tel que $m < \ell$. Alors f est minorée par m au voisinage de a.

(ii) Soit $M \in \mathbb{R}$, tel que $\ell < M$. Alors f est majorée par M au voisinage de a.

(iii) Soient $m, M \in \mathbb{R}$ tels que $m < \ell < M$. Alors $m \leqslant f \leqslant M$ au voisinage de a.

(iv) Soit $m \in \mathbb{R}_+$ tel que $m < |\ell|$. Alors $|f|$ est minorée par m au voisinage de a.

Démonstration. On suppose $a \in \mathbb{R}$ (le lecteur traitera les autres cas de façon similaire).

(i) On applique la caractérisation de Weierstraß des limites en supposant $\varepsilon := \ell - m$. Il existe $\eta > 0$ tel que $|x - a| \leqslant \eta$ entraîne $\left|f(x) - \ell\right| \leqslant \ell - m$. Alors $\ell - f(x) \leqslant \ell - m$, donc $m \leqslant f(x)$. Ainsi f est minorée par m sur $D_f \cap [a - \eta, a + \eta]$, donc au voisinage de a.

(ii) On applique (i) à la fonction $-f$, avec $m := -M$.

(iii) L'assertion résulte immédiatement de (i) et (ii).

(iv) La fonction $|f|$ tend vers $|\ell|$ en a (*cf.* l'exercice 2), on utilise ensuite (i). ■

Corollaire 9. Une fonction admettant une limite $\ell > 0$ en $a \in \mathbb{R} \cup \{\pm\infty\}$ est minorée au voisinage de a par un réel *strictement positif*.

Démonstration. On applique la proposition ci-dessus avec $m := \ell/2$. ■

En utilisant le théorème 7 on montre facilement la variante suivante de la proposition 8 pour les limites infinies.

Proposition 10. Soient $a \in \mathbb{R} \cup \{\pm\infty\}\}$, $A \in \mathbb{R}$ et une fonction f tendant vers $+\infty$ (resp. $-\infty$) en a. Alors f est minorée (resp. majorée) par A au voisinage de a.

1.2.3 Opérations sur les limites

Opérations

Proposition 11. Soient f et g deux fonctions définies sur une partie D et admettant respectivement des limites ℓ et ℓ' au point $a \in \mathbb{R} \cup \{\pm\infty\}$. Alors :

(i) $f + g$, $f - g$, $f\,g$ admettent respectivement en a, les limites $\ell + \ell'$, $\ell - \ell'$, $\ell\ell'$.

(ii) Si $\ell' \neq 0$, f/g est définie au voisinage de a et admet ℓ/ℓ' pour limite au point a.

Démonstration.

(i) Soit (x_n) une suite de limite a. Par hypothèse, $f(x_n) \to \ell$ et $g(x_n) \to \ell'$. Donc $f(x_n) + g(x_n) \to \ell + \ell'$, et de même avec les autres opérations, par application du théorème 33 de la page 546.

(ii) Dans le cas du quotient, le corollaire 9 permet de montrer que $\frac{f}{g}(x)$ est bien défini pour $x \in D$ suffisamment proche de a.

On montre ensuite que $\frac{f}{g}(x_n) = f(x_n)/g(x_n) \to \ell/\ell'$ en utilisant le théorème 33 de la page 546. ■

Proposition 12. Soient f et g deux fonctions définies sur une partie D et admettant respectivement des limites $\ell \in \mathbb{R}$ et $\pm\infty$ au point $a \in \mathbb{R} \cup \{\pm\infty\}$. Alors :

(i) $f + g$ et $f - g$ admettent respectivement en a les limites $\pm\infty$ et $\mp\infty$;

(ii) si l'on suppose $\ell \neq 0$, fg admet la limite $\pm\infty$ au point a si $\ell > 0$ et la limite $\mp\infty$ au point a si $\ell < 0$.

Si $\ell = 0$, on ne peut a priori rien dire de la limite de fg. Nous reviendrons sur la question dans le module IV.8.

Démonstration.

(i) La preuve est facile et laissée au lecteur.

(ii) On suppose que $\ell > 0$ et que la limite de g est $+\infty$ (on en déduit les autres cas en remplaçant f par $-f$ et g par $-g$). On montre, en utilisant la proposition 8, qu'on a $\ell/2 \leqslant f$ au voisinage de a. Par suite, la fonction fg est minorée par $\frac{\ell}{2}g$ au voisinage de l'infini et sa limite au point a est $+\infty$. ■

On montre facilement le résultat suivant.

Proposition 13. Soient f et g deux fonctions définies sur une partie D et admettant des limites infinies au point $a \in \mathbb{R} \cup \{\pm\infty\}$. Alors :

(i) si les deux limites sont $+\infty$ (resp. $-\infty$), alors $f+g$ admet en a la limite $+\infty$ (resp. $-\infty$) ;

(ii) si les deux limites sont infinies avec le même signe (resp. des signes opposés), alors fg admet en a la limite $+\infty$ (resp. $-\infty$).

Si les deux limites sont infinies avec le même signe, on ne peut a priori rien dire de la limite de $f-g$. Nous reviendrons sur la question dans le module IV.8.

La notion de limite ne se comporte pas bien, en général, par composition, comme on le voit dans l'exercice suivant.

Exercice 3.
On considère la fonction $f = 0$ (qui vaut 0 pour tout x), et la fonction g définie par $g(0) = 1$ et $g(x) = 0$ si $x \neq 0$. Montrer que les fonctions f, g et $g \circ f$ ont toutes une limite en $0 = f(0)$ mais que $\lim\limits_{x\to 0} g \circ f(x) \neq \lim\limits_{x \to f(0)} g(x)$.

Solution. On a évidemment $\lim\limits_{x\to 0} f(x) = 0$ et $\lim\limits_{x\to f(0)} g(x) = \lim\limits_{x\to 0} g(x) = 0$. Par ailleurs la fonction $g\circ f(x) = g\left(f(x)\right)$ vaut 1 pour tout x, donc $\lim\limits_{x\to 0} g\circ f(x) = 1 \neq 0 = \lim\limits_{x\to f(0)} g(x)$.

Relations d'ordre

Proposition 14. Soit f une fonction positive admettant une limite finie ℓ quand x tend vers $a \in \mathbb{R} \cup \{\pm\infty\}$. Alors $\ell \geqslant 0$.

Démonstration. Soit $(x_n)_{n\in\mathbb{N}}$ une suite prenant ses valeurs dans $D_f \setminus \{a\}$ et de limite a. Par hypothèse, la suite $\big(f(x_n)\big)_{n\in\mathbb{N}}$ tend vers ℓ. On a, pour tout $n \in \mathbb{N}$, $f(x_n) \geqslant 0$, donc, en passant à la limite, $\ell \geqslant 0$. ■

En appliquant le résultat précédent à la fonction $g - f$, on obtient le résultat suivant.

Corollaire 15. Soient f et g deux fonctions définies sur un même voisinage D de a admettant des limites finies ℓ et ℓ' quand x tend vers a. Si $f \leqslant g$, alors $\ell \leqslant \ell'$.

Nous avons vu le théorème des gendarmes pour les suites (*cf.* le théorème 43 de la page 555). Il existe un résultat analogue pour les fonctions.

Théorème 16 (Théorème dit « des gendarmes »). Soient f, g et h trois fonctions définies sur un même voisinage D de a et telles que $f \leqslant g \leqslant h$. On suppose que f et h admettent une limite commune $\ell \in \mathbb{R}$ au point a. Alors g admet la limite ℓ au point a.

Démonstration. On a $g - f = |g - f| \leqslant h - f$ et $h - f$ admet la limite 0 en a, donc, en utilisant la proposition 5, $g - f$ admet la limite 0 en a. On a $g = (g - f) + f$ donc :

$$\lim_{x \to a} g(x) = \lim_{x \to a} (g - f)(x) + \lim_{x \to a} f(x) = 0 + \ell = \ell.$$

On peut aussi déduire le résultat du théorème des gendarmes pour les suites en utilisant la définition séquentielle des limites. Le lecteur pourra le faire à titre d'exercice. ■

1.2.4 Théorème de la limite monotone

Si a appartient à un intervalle $I \subset \mathbb{R}$, nous rappelons que l'on dit qu'il est *intérieur à* I s'il n'est pas une borne de I, ou, ce qui est équivalent, s'il appartient à l'intérieur de I (*cf.* la définition 13 de la page 519).

Dans tout ce paragraphe, on considère une fonction f définie sur un intervalle I de $\mathbb{R}$ contenant au moins deux points (ou, ce qui est équivalent, d'intérieur non vide).

Définition 6. On dit qu'une fonction f est :

- *croissante* sur I si $\forall (x, y) \in I^2$, $x \leqslant y \Rightarrow f(x) \leqslant f(y)$,
- *décroissante* sur I si $\forall (x, y) \in I^2$, $x \leqslant y \Rightarrow f(x) \geqslant f(y)$,
- *strictement croissante* sur I si $\forall (x, y) \in I^2$, $x < y \Rightarrow f(x) < f(y)$,
- *strictement décroissante* sur I si $\forall (x, y) \in I^2$, $x < y \Rightarrow f(x) > f(y)$.

Elle est *monotone* (respectivement *strictement monotone*) si elle est croissante ou décroissante (respectivement strictement croissante ou strictement décroissante).

On appelle *Théorème de la limite monotone* l'ensemble des résultats énoncés dans les propositions 18 et 19 ci-dessous.

Lemme 17. Soit f une fonction croissante définie sur l'intervalle I de borne supérieure b (avec $b = +\infty$ si I n'est pas majoré).

- Si f est majorée, alors f admet une limite finie en b.
- Sinon, f tend vers $+\infty$ en b.

Lorsque $b \in I$, on est dans le premier cas puisqu'alors, par croissance, f est majorée par $f(b)$.

Démonstration. Considérons l'ensemble $A := \{f(x) \mid x \in I,\ x < b\}$. C'est une partie non vide de $\mathbb{R}$. Soit (x_n) une suite à valeurs dans $I' := I \setminus \{b\}$ et tendant vers b.

- Si f est majorée, alors A est majorée et admet donc une borne supérieure ℓ. Montrons que $\big(f(x_n)\big)$ tend vers ℓ.

  Soit $\varepsilon > 0$. Comme $\ell - \varepsilon < \ell$ et que ℓ est le plus petit des majorants de A, $\ell - \varepsilon$ n'est pas un majorant de A et il existe donc $a \in I'$ tel que $f(a) > \ell - \varepsilon$. Or, (x_n) tend

vers $b > a$, donc il existe un rang N à partir duquel $x_n \geqslant a$. Par croissance de f, on en déduit :

$$\forall n \in \mathbb{N}\,,\ n \geqslant N \Longrightarrow x_n \geqslant a \Longrightarrow f(x_n) \geqslant f(a) \geqslant \ell - \varepsilon.$$

Puisque f est majorée par ℓ, on a $\forall n \in \mathbb{N}\,,\ f(x_n) \leqslant \ell$. On a ainsi montré :

$$\forall \varepsilon > 0\,,\ \exists N \in \mathbb{N}\,,\ \forall n \geqslant N\,,\ \big|f(x_n) - \ell\big| \leqslant \varepsilon$$

c'est-à-dire la convergence de $\big(f(x_n)\big)$ vers ℓ.

- Si f n'est pas majorée, on peut trouver, pour tout réel M un élément $a \in I'$ tel que $f(a) \geqslant M$. On démontre, de la même façon que dans le premier cas, qu'il existe un rang à partir duquel on a $f(x_n) \geqslant M$. Ainsi, $\big(f(x_n)\big)$ tend vers $+\infty$. ■

Proposition 18.
Toute fonction monotone sur I admet une limite, finie ou infinie, aux bornes de I.

Démonstration. Le lemme précédent donne le résultat en la borne supérieure de I lorsque f est croissante. En utilisant $-f$, on en déduit le résultat pour les fonctions décroissantes. Enfin, en considérant $x \mapsto f(-x)$, définie sur $-I$, on en déduit le résultat en la borne inférieure de I. ■

Proposition 19. Une fonction monotone sur I admet, en tout point a intérieur à I, une limite finie à droite, notée $f(a^+)$, et une limite finie à gauche, notée $f(a^-)$, avec $f(a)$ compris entre $f(a^-)$ et $f(a^+)$. Dans le cas où f est croissante , on a :

$$f(a^-) = \sup_{\substack{x \in I \\ x < a}} f(x) \leqslant f(a) \leqslant \inf_{\substack{x \in I \\ a < x}} = f(a^+). \tag{6}$$

De plus, pour tous $c,\ d \in I$, $c < d$, on a $f(c^+) \leqslant f(d^-)$.

Démonstration. Quitte à considérer $-f$, on peut supposer f croissante. La restriction de f à $I \cap \,]-\infty, a[$ est alors croissante et majorée par $f(a)$, d'où (en utilisant la caractérisation séquentielle de la borne supérieure, *cf.* 29) l'existence d'une limite à gauche en a vérifiant $f(a^-) = \sup\limits_{\substack{x \in I \\ x < a}} f(x) \leqslant f(a)$. De même, il existe une limite à droite vérifiant $f(a^+) = \inf\limits_{\substack{x \in I \\ a < x}} f(x) \geqslant f(a)$. D'où (6).

Pour tout $y > d$, on a $f(d) \leqslant f(y)$, donc $\inf\limits_{\substack{x \in I \\ c < x}} f(x) = \inf\limits_{\substack{x \in I \\ c < x < d}} f(x)$. Pour tout $y < c$, on a $f(y) \leqslant f(c)$, donc $\sup\limits_{\substack{x \in I \\ x < d}} f(x) = \sup\limits_{\substack{x \in I \\ c < x < d}} f(x)$. On en déduit :

$$f(c^+) = \inf_{\substack{x \in I \\ c < x}} f(x) = \inf_{\substack{x \in I \\ c < x < d}} f(x) \leqslant \sup_{\substack{x \in I \\ c < x < d}} f(x) = \sup_{\substack{x \in I \\ x < d}} f(x) = f(d^-).$$ ■

Exercice 4.

Soit $f : [a, b] \to \mathbb{R}$ une fonction croissante.

Pour tout $c \in]a, b[$, on note $\sigma(c) := f(c^+) - f(c^-)$. On dit que f fait un saut en c si $f(c^+) \neq f(c^-)$, c'est-à-dire si $\sigma(c) \neq 0$. On a alors $\sigma(c) > 0$. Nous admettrons le résultat suivant (on trouvera une preuve dans l'exercice IV.2.8 de la page 672) : Si $c_1,\ c_2, \ldots, c_p \in]a, b[$ sont des points deux à deux distincts, alors :

$$\sum_{i=1}^{p} \sigma(c_i) \leqslant f(b) - f(a). \tag{7}$$

1. Soient $c,\ d \in]a, b[$, tels que $c < d$. Montrer que :
$$]f(c_-), f(c_+)[\cap]f(d_-), f(d_+)[= \varnothing.$$
2. On note Σ l'ensemble des points de $]a, b[$ où f fait un saut.
   a) Montrer qu'il existe une application $\varphi : \Sigma \to \mathbb{Q}$ telle que, pour tout $c \in \Sigma$, $\varphi(c) \in]f(c_-), f(c_+)[$. Montrer qu'un telle application est injective. Montrer que Σ est au plus dénombrable (c'est-à-dire fini ou dénombrable).
   b) Donner une autre preuve du fait que Σ est au plus dénombrable en utilisant (7).

Solution.

1. D'après la proposition 19 ci-dessus, on a $f(c^+) \leqslant f(d^-)$. Soient $x \in]f(c_-), f(c_+)[$ et $y \in]f(d_-), f(d_+)[$. On a $x < f(c_+) \leqslant f(d_-) < y$, donc $x \neq y$.
   Par suite $]f(c_-), f(c_+)[\cap]f(d_-), f(d_+)[= \varnothing$.
2. a) Si $c \in \Sigma$, $f(c_-) < f(c_+)$, il existe donc une infinité de rationnels dans $]f(c_-), f(c_+)[$. On définit $\varphi : \Sigma \to \mathbb{Q}$ en choisissant, pour tout $c \in \Sigma$ l'un de ces rationnels. L'application φ est injective d'après 1. b), elle induit donc une bijection entre Σ et $\varphi(\Sigma) \subset \mathbb{Q}$. Donc Σ est au plus dénombrable.
   b) Pour tout $n \in \mathbb{N}$, on note :
$$\Sigma_n := \left\{ x \in]a, b[\ \middle|\ \sigma(x) \geqslant \frac{1}{n+1} \right\}.$$
   Soient $c_1,\ c_2, \ldots, c_p \in \Sigma_n$ des points deux à deux distincts. D'après (7) on a :
$$\frac{p}{n+1} \leqslant \sum_{i=1}^{p} \sigma(c_i) \leqslant f(b) - f(a),$$
   donc $p \leqslant (n+1)\,(f(b) - f(a))$, ce qui montre que l'ensemble Σ_n est *fini*.
   On a $\Sigma = \bigcup_{n \in \mathbb{N}} \Sigma_n$. Ainsi Σ est une réunion dénombrable d'ensembles finis, il est donc au plus dénombrable. ■

1.3 Continuité

1.3.1 Définitions et premières propriétés

Définition 7. On dit qu'une fonction f est *continue au point* a si $a \in D_f$ et si $f(x)$ a pour limite $f(a)$ lorsque x tend vers a.

Il faut donc deux choses : que la limite existe, et que f soit définie en a avec une valeur $f(a)$ égale à la limite en a. Dans le cas de la fonction de Dirichlet, $f(a)$ existe pour tout a, mais n'est jamais égal à la limite de $f(x)$, puisque celle-ci n'existe pas : une telle fonction n'est donc continue en aucun point.
Comme pour les limites, on dit aussi que f est continue à gauche (respectivement à droite) si la propriété n'est vraie que pour la limite à gauche (respectivement à droite). Ainsi la fonction partie entière est continue à droite en tout point, mais n'est pas continue à gauche aux points $a \in \mathbb{Z}$.

Définition 8. On dit que f est continue sur un intervalle I (ou plus généralement sur un sous-ensemble $I \subset \mathbb{R}$) si elle est continue en tout point $a \in I$.

Définition 9. Soient f une fonction, et $a \notin D_f$. Supposons que f ait une limite finie ℓ au point a. Considérons la nouvelle fonction $\tilde{f}$ définie par $\tilde{f}(x) = f(x)$ si $x \neq a$, et $\tilde{f}(a) = \ell$. On a alors $D_{\tilde{f}} = D_f \cup \{a\}$, et $\tilde{f}$ est continue au point a.

On dit que $\tilde{f}$ est le *prolongement par continuité* de f au point a.

Un tel procédé permet dans certains cas d'ajouter un point à l'ensemble de définition d'une fonction continue, en conservant la continuité. Ainsi la fonction $x \mapsto x \sin \frac{1}{x}$ peut se prolonger en $x = 0$ avec la valeur 0 (*cf.* l'exemple 1 de la page 622). On obtient une fonction définie sur $\mathbb{R}$ tout entier.

Théorème 20 (Caractérisation de Weierstraß). Une fonction f est continue en un point $a \in D_f$ si, et seulement si :

$$\forall \varepsilon > 0\,,\ \exists \eta > 0\,,\ \forall x \in D_f\,,\ |x - a| \leqslant \eta \ \Rightarrow\ |f(x) - f(a)| \leqslant \varepsilon. \tag{8}$$

Démonstration. Considérons d'abord une fonction satisfaisant (8). D'après le théorème 6, elle admet $f(a)$ pour limite en a, elle est donc continue au sens de la définition 7.
Inversement, supposons f continue en a au sens de la définition 7. Elle admet $f(a)$ pour limite en a, donc, d'après le théorème 6 :

$$\forall \varepsilon > 0\,,\ \exists \eta > 0\,,\ \forall x \in D_f \setminus \{a\}\,,\ |x - a| \leqslant \eta \ \Rightarrow\ |f(x) - f(a)| \leqslant \varepsilon.$$

Si $x = a$, $|f(x) - f(a)| = 0 \leqslant \varepsilon$. On a donc (8). ∎

Soient I un intervalle de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction continue sur I. Pour $\varepsilon > 0$ *fixé*, il existe, pour tout $a \in I$, $\eta > 0$, tel que :

$$|x - a| \leqslant \eta \ \Rightarrow\ |f(x) - f(a)| \leqslant \varepsilon. \tag{9}$$

En général on trouve un η *qui dépend de* a. Il peut être impossible de trouver un $\eta > 0$ tel que (9) soit satisfaite pour tout $a \in I$. Inversement, nous définirons plus loin la notion de fonction lipschitzienne qui permet dans certains cas d'obtenir un tel η.

Corollaire 21. Soit f une fonction définie sur un intervalle $]\alpha, \beta[$ contenant le point a, continue en a avec $f(a) > 0$. Alors, il existe $\varepsilon > 0$ tel que $]a - \varepsilon, a + \varepsilon[\subset]\alpha, \beta[$ et que, pour tout $x \in]a - \varepsilon, a + \varepsilon[$, l'on ait $f(x) > 0$.

Démonstration. Comme $\alpha < a < \beta$, le réel $\delta := \min(a - \alpha, \beta - a)$ est strictement positif et vérifie $]a - \delta, a + \delta[\subset]\alpha, \beta[$.
Soit $\varepsilon := f(a)/2$ (strictement positif); il existe $\eta > 0$, inférieur ou égal à δ, tel que $\big|f(x) - f(a)\big| < \varepsilon$ pour tout $x \in]a - \eta, a + \eta[$.
Or cela implique $f(x) > f(a) - \varepsilon = \varepsilon > 0$ ce qu'il fallait démontrer. ■

Exercice 5.
Montrer que la fonction $f : x \mapsto \sqrt{1 - x^2}$ est continue en tout point $a \in [-1, +1]$, en vérifiant directement la propriété (8) (c'est-à-dire en trouvant η pour a et $\varepsilon > 0$ donnés).
Solution. Supposons donné $\varepsilon > 0$.
Pour $a := 1$, l'inégalité $|f(x) - f(a)| < \varepsilon$ s'écrit $\sqrt{1 - x^2} < \varepsilon$ ce qui équivaut à $1 - x^2 < \varepsilon^2$ et est vérifié pour $|x - 1| = 1 - x < \varepsilon^2/2 =: \eta(1, \varepsilon)$ puisqu'alors :

$$1 - x^2 = (1 - x)\,(1 + x) \leqslant 2(1 - x) = 2|x - 1| < \varepsilon^2.$$

Il en va de même pour $a := -1$.
Soit donc $|a| < 1$. Pour $x \neq a$, on dispose des relations $|a + x| \leqslant |a| + |x| < 2$ et :

$$|f(x) - f(a)| = \big|\sqrt{1 - x^2} - \sqrt{1 - a^2}\big| = \frac{|x - a|\,|x + a|}{\sqrt{1 - x^2} + \sqrt{1 - a^2}} \leqslant 2\,\frac{|x - a|}{\sqrt{1 - a^2}}.$$

Par suite $\eta(a, \varepsilon) := \varepsilon\, f(a)/2$ convient.

On admet le résultat suivant (une preuve fait l'objet de l'exercice IV.2.7 de la page 672).

Proposition 22.
Une fonction continue périodique non constante admet une plus petite période.

Définition 10. Soient $I \subset \mathbb{R}$ un intervalle et $f : I \to \mathbb{R}$ et $k > 0$.
On dit que f est k-*lipschitzienne* (ou *lipschitzienne* de rapport k) sur I si, pour tous $x,\ y \in I$, on a $|f(x) - f(y)| \leqslant k|x - y|$.
Si $k < 1$, on dit que f est *contractante* sur I.

Proposition 23. Soient $I \subset \mathbb{R}$ et $k > 0$. Si $f : I \to \mathbb{R}$ est k-lipschitzienne sur I, elle est continue sur I.

Démonstration. Soit $\varepsilon > 0$, on choisit $\eta := \frac{\varepsilon}{k+1}$.

Alors, si $|y-x|<\eta$, on a $\big|f(y)-f(x)\big| \leqslant \frac{k}{k+1}\,\varepsilon < \varepsilon$.
Notons que η est *indépendant* de x et y. ■

1.3.2 Opérations sur les fonctions continues

Théorème 24. Soient f et g deux fonctions définies sur une partie D et continues au point a. Alors $f+g$, $f-g$, $f\,g$ sont continues en a, ainsi que f/g si $g(a)\neq 0$.

Démonstration. C'est une conséquence immédiate de la proposition 11 de la page 626. ■

Théorème 25. Soient f une fonction continue en a, et g une fonction continue en $f(a)$. On suppose $f(D_f)\subset D_g$. Alors $g\circ f$ est continue en a.

Démonstration. Soit (x_n) une suite de limite a.
Alors $y_n := f(x_n)$ est le terme général d'une suite de limite $f(a)$, par continuité de f. Par continuité de g, on doit avoir $\lim g(y_n) = g\big(f(a)\big)$. Mais cela revient exactement à dire que $\lim g\circ f(x_n) = g\circ f(a)$, donc que $g\circ f$ est continue en a. ■

En partant de fonctions très simples et visiblement continues (comme les constantes ou la fonction $x\mapsto x$), on en déduit, par application des quatre opérations élémentaires et de la composition, que bien d'autres fonctions sont continues.

Exemples.

1. Toute fonction polynomiale est continue sur $\mathbb{R}$. Toute fraction rationnelle est continue sur son ensemble de définition.
2. Soient $f : D_f\setminus\{a\}\to\mathbb{R}$ une fonction continue en a.
   Alors $|f|$ est définie sur $D_f\setminus\{a\}$ et elle est continue en a.
3. Soient $f\ g : D_f\setminus\{a\}\to\mathbb{R}$ deux fonction continues en a. On définit deux fonctions $\sup\{f,g\}$ et $\inf\{f,g\}$ sur $D_f\setminus\{a\}$ par :
   $$\sup\{f,g\} : x\mapsto \max\{f(x),g(x)\} \quad \text{et} \quad \inf\{f,g\} : x\mapsto \min\{f(x),g(x)\}\,.$$
   Ces fonctions sont continues en a. En effet :
   $$\sup\{f,g\} = \frac{1}{2}\,(f+g+|f-g|) \quad \text{et} \quad \inf\{f,g\} = \frac{1}{2}\,(f+g-|f-g|)\,.$$

1.4 Théorème des valeurs intermédiaires et image continue d'un segment

1.4.1 Théorème des valeurs intermédiaires

Le terme même de *continuité* implique l'absence de rupture dans un processus. Ainsi, un objet qui se déplace a intuitivement un *mouvement continu*, c'est-à-dire qu'il ne peut pas passer instantanément d'un point à un autre[2]. Une personne située dans une pièce ne peut pas sortir de la pièce sans passer par les bords de celle-ci (portes, fenêtres, voire murs[3]...). Le théorème suivant explicite cette notion intuitive.

Théorème 26 (des valeurs intermédiaires).

(i) Soit f une fonction continue sur un segment $[a,b]$. Alors pour toute valeur γ comprise entre $f(a)$ et $f(b)$, il existe un $c \in [a,b]$ tel que $f(c) = \gamma$.

(ii) Soit f une fonction continue sur un segment $[a,b]$ telle que $f(a)f(b) \leqslant 0$. Alors il existe un $c \in [a,b]$ tel que $f(c) = 0$.

L'assertion (i) est le théorème des valeurs intermédiaires proprement dit. Le réel γ est un réel « intermédiaire » entre $f(a)$ et $f(b)$ (en d'autres termes on a $\gamma \in [f(a), f(b)]$ si $f(a) \leqslant f(b)$, et $\gamma \in [f(b), f(a)]$ dans le cas contraire). Le théorème nous dit que ce nombre est une *valeur* de la fonction. Notons que le point c n'est pas nécessairement unique.

Démonstration. L'assertion (ii) est un cas particulier de (i) ($\gamma = 0$). Inversement, en appliquant (ii) à la fonction $g := f - \gamma$, on obtient (i). Il suffit donc de prouver (ii).

La démonstration suivante utilise un procédé appelé *dichotomie* (divisions successives par deux), très utile en mathématiques ou dans différentes sciences appliquées. Il y a aussi des variantes en informatique.

Supposons par exemple que $f(a) \leqslant 0 \leqslant f(b)$, de sorte que $0 \in [f(a), f(b)]$ (l'autre cas s'y ramène en considérant $-f$). On construit par récurrence deux suites (a_n) et (b_n) de la façon suivante. On part de $a_0 := a$, $b_0 := b$, et supposant construits a_n et b_n tels que $f(a_n) \leqslant 0 \leqslant f(b_n)$, on considère la valeur de f en $(a_n + b_n)/2$, milieu du segment $[a_n, b_n]$. On construit alors a_{n+1} et b_{n+1} ainsi :

- si $f\left(\frac{a_n+b_n}{2}\right) < 0$, on pose $a_{n+1} := \frac{a_n+b_n}{2}$ et $b_{n+1} := b_n$;
- sinon, c'est-à-dire si $f\left(\frac{a_n+b_n}{2}\right) \geqslant 0$, on pose $a_{n+1} := a_n$ et $b_{n+1} := \frac{a_n+b_n}{2}$.

En examinant les deux cas, on voit facilement que l'on a $a_n \leqslant a_{n+1} \leqslant b_{n+1} \leqslant b_n$ pour tout n, ainsi que :

$$0 \leqslant b_n - a_n = \frac{b-a}{2^n} \qquad \text{et} \qquad f(a_n) \leqslant 0 \leqslant f(b_n). \tag{$*$}$$

En particulier $\lim\limits_{n\to+\infty} |b_n - a_n| = 0$, et les deux suites (a_n) et (b_n) sont adjacentes. D'après le théorème 45 de la page 560, elles convergent donc vers une limite commune c. On a $c \in [a,b]$ car $a_n \in [a,b]$ pour tout n. Comme f est continue, on a $\lim f(a_n) = f(c) = \lim f(b_n)$ et la double inégalité de $(*)$ implique donc $f(c) = 0$. ■

[2] Tout au moins en mécanique classique...

[3] Comme le passe muraille de la nouvelle de Marcel Aymé...

Remarque. La démonstration ci-dessus donne un *algorithme* pour trouver la valeur approchée d'une solution de l'équation $f(x) = 0$. On cherche d'abord $a,\ b \in D_f$ tels que $a < b$ et $f(a)f(b) < 0$. On applique ensuite le procédé de dichotomie. On s'arrête lorsque $\frac{b-a}{2^n}$ est inférieur à la précision demandée. Il peut y avoir plusieurs solutions entre a et b et, dans la mesure du possible, on choisit a et b de telle sorte qu'il n'y en ait qu'une seule. C'est le cas si f est strictement monotone sur $[a, b]$.

Exemple. On se propose de trouver une valeur approchée de $\sqrt{10}$. On cherche une solution de l'équation $f(x) = x^2 - 10 = 0$. On sait qu'il existe une solution unique sur $\mathbb{R}_+$. On a $f(3) = -1 < 0$ et $f(4) = 6 > 0$. En appliquant le procédé de dichotomie, on trouve :

$$a_3 = 3,125 \quad b_3 = 3,250, \quad \text{et} \quad 3,125 < \sqrt{10} < 3,250$$

En allant à l'ordre 8, on obtient : $3,1601 < \sqrt{10} < 3,1641$. On obtient donc 2 décimales exactes : $\sqrt{10} = 3,16\ldots$.
Il est intéressant de comparer cette méthode de calcul à la méthode dite babylonienne (*cf.* l'exemple de la page 558). Pour calculer une valeur aprochée de $\sqrt{10}$ par la méthode babylonienne, on considère la suite récurrente (x_n) définie par la fonction $g : x \mapsto \frac{1}{2}\left(x + \frac{10}{x}\right)$ et la valeur initiale $x_0 := 4$. Cette suite converge vers $\sqrt{10}$ en décroissant. On a :

$$x_1 = \frac{13}{4}, \quad x_2 = \frac{329}{104} = 3,\underline{16}346\ldots, \quad x_3 = \frac{216401}{68432} = 3,\underline{16227}788\ldots.$$

On compare à $\sqrt{10} = 3,16227766\ldots$. Ainsi, on a 2 décimales exactes à la deuxième étape et 6 à la troisième. Cet algorithme est beaucoup plus rapide que l'algorithme de dichotomie.

Pour une application f quelconque définie sur un ensemble non vide $A \subset \mathbb{R}$, on appelle *infimum de f sur A* la borne inférieure de $f(A)$, si elle existe, ou le nombre $-\infty$ si $f(A)$ n'est pas minorée. On le note $\inf\limits_A f$ ou $\inf\limits_{x\in A} f(x)$. On définit de même le *supremum de f sur A*, qui est la borne supérieure de $f(A)$ ou $+\infty$; on le note $\sup\limits_A f$ ou $\sup\limits_{x\in A} f(x)$.

Corollaire 27. Soient I un intervalle de $\mathbb{R}$ et f une fonction continue sur I.
Alors $f(I)$ est un intervalle de $\mathbb{R}$, de bornes $\inf\limits_I f$ et $\sup\limits_I f$.

On dit que l'image continue d'un intervalle est un intervalle.

Démonstration. D'après la proposition 15 de la page 521, un ensemble $A \subset \mathbb{R}$ est un intervalle si, et seulement s'il est convexe, c'est-à-dire : pour tous $x, y \in A$, avec $x \leqslant y$, on a $[x, y] \subset A$. Pour montrer que $f(I)$ est un intervalle, nous allons montrer qu'il est convexe. Soient $\alpha, \beta \in f(I)$. Il existe $a,\ b \in I$ tels que $\alpha = f(a)$ et $\beta = f(b)$, et en appliquant le théorème des valeurs intermédiaires on obtient : $[\alpha, \beta] \subset f(I)$.

Le fait que les bornes soient exactement l'infimum et le supremum résulte de leur définition et de la remarque de la page 521. ∎

Noter qu'il importe peu ici que l'intervalle soit ouvert, fermé ou semi-ouvert. Par ailleurs $f(I)$ peut être fermé avec I ouvert, ou le contraire, et d'autres combinaisons sont possibles. Par exemple $\sin\big(]-\pi,\pi[\big) = [-1,+1]$, $\cos\big(]-\pi/2,\pi/2[\big) =]0,1]$, *etc.*

Exercice 6.

Soit $P \in \mathbb{R}[X]$ un polynôme de degré *impair*. Montrer qu'il admet au moins un zéro réel.

Solution. Soit $P(x) = a_n x^n + \cdots + a_1 x + a_0$ avec n impair et $a_n \neq 0$.

On pose, pour $x \in \mathbb{R}^*$, $g(x) = \frac{P(x)}{a_n x^n}$ et l'on vérifie que $g(x) = 1 + \cdots + \frac{a_1}{a_n x^{n-1}} + \frac{a_0}{a_n x^n}$ admet 1 pour limite en $\pm\infty$. On en déduit, en utilisant la proposition 12 de la page 626 pour $P = a_n x^n g$, que :

- si $a_n > 0$, P admet $+\infty$ pour limite en $+\infty$ et $-\infty$ pour limite en $-\infty$;
- si $a_n < 0$, P admet $-\infty$ pour limite en $+\infty$ et $+\infty$ pour limite en $-\infty$.

(Pour obtenir le résultat ci-dessus on peut aussi utiliser la notion d'équivalent introduite dans le module IV.8, *cf.* l'exercice 2 de la page 864.)

Notons J l'image de $\mathbb{R}$ par la fonction polynomiale P. D'après le corollaire 27, J est un intervalle. Montrons que $0 \in J$. On suppose $a_n > 0$ (le cas $a_n < 0$ est similaire). On a $\lim\limits_{x\to-\infty} P(x) = -\infty$ et $\lim\limits_{x\to+\infty} P(x) = +\infty$, donc il existe $a, b \in \mathbb{R}$ tels que $P(a) \leqslant -1 < 1 \leqslant P(b)$, et, puisque J est un intervalle, $0 \in [-1,1] \subset [P(a), P(b)] \subset J$. Il existe donc un réel c tel que $P(c) = 0$. Le lecteur vérifiera à tite d'exercice que $J = \mathbb{R}$.

On notera que l'image réciproque d'un intervalle par une fonction continue n'est pas toujours un intervalle. Par exemple l'image réciproque de $[0,1]$ par la fonction sinus est $[0,\pi] + 2\pi\mathbb{Z}$, qui n'est manifestement pas un intervalle.

Nous allons voir que le corollaire 27 ci-dessus possède une réciproque dans le cas des fonctions monotones.

Proposition 28. Soit f une fonction monotone sur un intervalle I telle que $f(I)$ soit aussi un intervalle. Alors f est continue.

Démonstration. Quitte à considérer $-f$, on peut supposer f croissante.

Soit a un élément de I qui n'en est pas sa borne supérieure. Le théorème de la limite monotone (voir la section 1.2.4 de la page 628) donne l'existence d'une limite à droite en a notée $f(a^+)$ et vérifiant $f(a^+) \geqslant f(a)$. On a alors :

$$\forall x \in I\,,\ x \leqslant a \Longrightarrow f(x) \leqslant f(a) \qquad \text{et} \qquad \forall x \in I\,,\ x > a \Longrightarrow f(x) \geqslant f(a^+).$$

La fonction f ne prend donc aucune valeur strictement comprise entre $f(a)$ et $f(a^+)$. Or, comme a n'est pas la borne supérieure de I, il existe $b \in I$ tel que $b > a$, de sorte que, par croissance de f : $f(a) \leqslant f(a^+) \leqslant f(b)$. Comme $f(I)$ est un intervalle, toutes les valeurs comprises entre $f(a)$ et $f(b)$ sont atteintes par f. On en déduit que l'inégalité $f(a) < f(a^+)$ serait absurde et donc que $f(a) = f(a^+)$. Ainsi, f est continue à droite en a.

On démontre de même la continuité à gauche de f en tout point qui n'est pas la borne inférieure de I. Ainsi f est bien continue sur I. ∎

On déduit de la proposition précédente le résultat important suivant. Il permet de compléter les opérations sur les fonctions continues : après la somme, le produit, le quotient et la composée, on a un résultat concernant la continuité d'une réciproque.

Théorème 29.
Une application continue strictement monotone f d'un intervalle I dans $\mathbb{R}$ induit une bijection de I sur l'intervalle $J := f(I)$ et sa réciproque est continue sur J.

Démonstration.

- La stricte monotonie de f donne son injectivité, puisque si x et y sont deux éléments de I, avec par exemple $x < y$, on a $f(x) < f(y)$ (lorsque f est croissante) ou $f(x) > f(y)$ (f décroissante). C'est donc une bijection de I sur $J = f(I)$.
- Le fait que J soit un intervalle est une conséquence du corollaire 27 de la page 635.
- Enfin, la bijection réciproque f^{-1} est monotone sur l'intervalle J et $f^{-1}(J) = I$ est un intervalle. D'après la proposition 28 ci-dessus, f^{-1} est continue sur J. ■

Soit $f : I \to J$ une bijection continue telle que $f^{-1} : J \to I$ soit continue. On dit que f est *bicontinue* ou qu'elle réalise un *homéomorphisme* de I sur J.

Proposition 30. Soit $f : I \to \mathbb{R}$ une fonction continue injective. Alors f est strictement monotone et réalise un homéomorphisme de I sur l'intervalle $J := f(I)$.

Démonstration. Soit $f : I \to \mathbb{R}$ une fonction continue injective. Montrons que f est monotone. On raisonne par l'absurde. Si f n'est pas monotone, il existe $x, x', y, y' \in I$ tels que $x < x'$, $y < y'$, $f(x) \geqslant f(x')$ et $f(y) \leqslant f(y')$. On définit la fonction

$$\begin{array}{rccl} g: & [0,1] & \longrightarrow & \mathbb{R} \\ & t & \longmapsto & f(tx + (1-t)y) - f(tx' + (1-t)y'). \end{array}$$

La fonction g est continue sur $[0,1]$ et l'on a $g(0) \leqslant 0$ et $g(1) \geqslant 0$. D'après le théorème des valeurs intermédiaires, il existe $t_0 \in [0,1]$ tel que $g(t_0) = 0$.
Posons $z := t_0x + (1-t_0)y$ et $z' := t_0x' + (1-t_0)y'$. On a $z < z'$ et $f(z) = f(z')$ ce qui contredit l'injectivité de f.
La fonction f étant monotone et injective, elle est strictement monotone. Elle réalise un homéomorphisme de I sur $f(I)$ d'après le théorème 29. ■

Remarque. Soient I et J deux intervalles réels et $f : I \to J$ une application *surjective*. On considère les trois conditions suivantes :

(i) f est *continue* ;

(ii) f est *strictement monotone* ;

(iii) f est *injective*.

En utilisant les propositions 28 et 30, on montre que si f vérifie deux des conditions, elle vérifie les trois. S'il en est ainsi f est une bijection bicontinue de I sur J.

1.4.2 Image continue d'un segment

Si I est un intervalle et f est continue, $f(I)$ est un intervalle. Il est possible que cet intervalle ne soit pas borné et cela peut arriver même si I est un intervalle borné. Par exemple si $f : x \mapsto 1/x$, alors $f\big(]0,1]\big) = [1,+\infty[$. Le théorème 31 ci-dessous dit que si l'intervalle I est *à la fois fermé et borné*, alors $f(I)$ l'est aussi ; en particulier f est alors aussi bornée.

On rappelle la définition d'un segment et la notation correspondante.

Notation. Lorsque a et b appartenant à $\mathbb{R}$ vérifient $a \leqslant b$, on note :

$$[a,b] := \big\{x \in \mathbb{R} \mid a \leqslant x \leqslant b\big\}.$$

Les ensembles $[a,b]$, avec $a \leqslant b$, sont ce que l'on appelle les *segments* de $\mathbb{R}$. Un segment peut être réduit à un point mais *n'est jamais vide*.

Théorème 31. Toute fonction continue sur un segment est bornée et atteint ses bornes.

Cet énoncé signifie que si f est une fonction continue sur un segment $[a,b]$, alors f admet une borne supérieure M, une borne inférieure m, et que ces bornes sont atteintes par la fonction f, c'est-à-dire qu'il existe $c,\ d \in [a,b]$ tels que $f(c) = M$ et $f(d) = m$.

Démonstration. Considérons $M := \sup\limits_{[a,b]} f$, avec la convention $M = +\infty$ si la fonction f n'est pas majorée (ce qui est a priori possible).

1. Si f n'est pas majorée, il existe, pour tout $n \in \mathbb{N}$, un élément $x_n \in [a,b]$ tel que $f(x_n) \geqslant n$ et donc $\lim f(x_n) = +\infty = M$.
2. Si f est majorée, d'après la caractérisation de la borne supérieure, il existe, pour tout $n \in \mathbb{N}$, $x_n \in [a,b]$ tel que $M \geqslant f(x_n) \geqslant M - 2^{-n}$ et donc $\lim f(x_n) = M$.

Dans les deux cas, on dispose d'une suite (x_n) d'éléments de $[a,b]$, donc d'une suite bornée. D'après le théorème de Bolzano-Weierstraß (théorème 47 de la page 562), cette suite admet une sous-suite (x_{n_k}) convergente vers $c \in [a,b]$. Par continuité de f en c, on en déduit $f(c) = \lim\limits_{k\to+\infty} f(x_{n_k})$. Dans le cas 1 on a une contradiction, donc f est majorée et l'on est nécessairement dans le cas 2. Dans ce cas, la suite $\big(f(x_{n_k})\big)$, extraite de $\big(f(x_n)\big)$ tend aussi vers M et $f(c) = M$.

Le cas de la borne inférieure se déduit de celui de la borne supérieure en considérant la fonction $-f$. ∎

En combinant ce théorème avec celui des valeurs intermédiaires, on obtient que non seulement les bornes m et M et f sont atteintes, mais aussi toute valeur comprise entre m et M. On en déduit :

Théorème 32. L'image d'un segment par une application continue est un segment.

On dit que l'image continue d'un segment est un segment.

Il est essentiel dans ce théorème que $[a, b]$ soit à la fois fermé et borné. On a, par exemple $\exp(\mathbb{R}_-) =]0, 1]$, qui n'est pas un segment (mais $\mathbb{R}_-$ n'est pas borné).

Exercice 7.

Soient $I \subset \mathbb{R}$ un segment, $f : I \to \mathbb{R}$ une fonction continue et $J \subset f(I)$ un segment. Montrer qu'il existe un segment $K \subset I$ tel que l'on ait $f(K) = J$.

Solution. Notons c et d les extrémités de J. On peut supposer que $c \neq d$, sinon le résultat est évident.

Soient $a,\ b \in I$ tels que $f(a) = c$ et $f(b) = d$. On a $a \neq b$, et, quitte à changer le nom des points c et d, on peut supposer $a < b$. D'après le corollaire 27, on a $J \subset f([a, b])$, mais il n'y a pas en général égalité. Nous allons voir que l'on peut (si nécessaire) « raboter la différence » $f([a, b]) \setminus J$ en modifiant convenablement les choix de a et b. L'idée de la preuve est de trouver dans le segment $[a, b]$ un antécédent a' de c et un antécédent b' de d tels que $a' < b'$ et qu'il n'y ait pas d'antécédent de c ni d'antécédent de d dans l'intervalle ouvert $]a', b'[$. On prouve ensuite que $K := [a', b']$ répond à la question, c'est-à-dire que $f(K) = J$.

À l'aide d'un dessin du graphe de f, le lecteur pourra visualiser « l'effet de rabot » obtenu en choisissant a plus à droite ou b plus à gauche. Les détails de la preuve, donnés ci-dessous, sont un peu délicats et peuvent être omis en première lecture.

On définit $A := \{x \in [a, b] \mid f(x) = c\}$. Cet ensemble n'est pas vide ($a \in A$) et il est majoré par b, il admet donc une borne supérieure $a' \leqslant b$. D'après la caractérisation séquentielle de la borne supérieure (*cf.* la proposition 29 de la page 544), il existe une suite (x_n) de points de A qui converge vers a' et, puisque f est continue, on a $f(a') = c$ et $a' < b$ ($f(b) = d \neq c$). On définit $B := \{x \in [a', b] \mid f(x) = d\}$. Cet ensemble n'est pas vide ($b \in B$) et il est minoré par a', il admet donc une borne inférieure $b' \geqslant a'$. D'après la caractérisation séquentielle de la borne inférieure, il existe une suite (y_n) de points de B qui converge vers b' et, puisque f est continue, on a $f(b') = d$ et $a' < b'$ ($f(a') = c \neq d$).

Le segment $[a', b']$ possède a propriété annoncée plus haut : il n'y a pas d'antécédent de c ni d'antécédent de d dans l'intervalle $]a', b'[$. Si l'on avait $x \in]a', b'[$ tel que $f(x) = c$, l'on aurait $x \in A$ et $x \leqslant a'$ donc une contradiction. De même, si l'on avait $x \in]a', b'[$ tel que $f(x) = d$, l'on aurait $x \in B$ et $x \geqslant b'$ donc une contradiction.

Soit $K := [a', b']$. Montrons que $f(K)) = J$.

On peut avoir $c < d$ ou $d < c$. Nous allons terminer la preuve en supposant $c < d$. Le cas $d < c$ s'en déduira (en remplaçant f par $-f$).

Il est clair que $J \subset f(K)$. Montrons que, pour tout $x \in K$, $f(x) \in J$, c'est-à-dire que l'on ne peut pas avoir $f(x) < c$ ni $f(x) > d$. On traite les deux cas en raisonnant par l'absurde.

- Si $x \in [a', b']$, avec $f(x) < c < d = f(b')$, d'après le théorème des valeurs intermédiaires, il existe $y \in]x, b'[\subset]a', b'[$, tel que $f(y) = c$. On a une contradiction.
- Si $x \in [a', b']$, avec $f(a') = c < d < f(x)$, d'après le théorème des valeurs intermédiaires, il existe $z \in]a', x[\subset]a', b'[$, tel que $f(z) = d$. On a une contradiction.

---

2 Dérivabilité

Dans toute cette partie, nous considérons des fonctions définies sur un intervalle I de $\mathbb{R}$.

2.1 Définitions, exemples

Définition 11. Soient f une fonction et $a \in D_f$. On dit que f est *dérivable au point* a si la fonction *taux d'accroissement* : $x \mapsto \dfrac{f(x)-f(a)}{x-a}$, définie sur $D_f \setminus \{a\}$, admet une limite lorsque x tend vers a.
Cette limite est appelée dérivée de f au point a, alors notée $f'(a)$, ou bien $\dfrac{\mathrm{d}f}{\mathrm{d}x}(a)$.

Une autre façon d'écrire la dérivabilité est de poser :

$$\forall x \in I \setminus \{a\}\,,\ \delta(x,a) := \frac{f(x)-f(a)}{x-a} - f'(a),$$

de sorte que :

$$f(x) = f(a) + (x-a)f'(a) + (x-a)\delta(x,a) \qquad \text{et} \qquad \lim_{x\to a} \delta(x,a) = 0. \tag{10}$$

Inversement l'existence d'un nombre ℓ et d'une fonction $\varepsilon(x,a)$ vérifiant :

$$f(x) = f(a) + (x-a)\,\ell + (x-a)\,\varepsilon(x,a) \qquad \text{et} \qquad \lim_{x\to a} \varepsilon(x,a) = 0$$

implique que f est dérivable en a avec $f'(a) = \ell$.
Dans cette écriture, chacun des trois termes du membre de droite est d'un ordre différent : les deux derniers « petits » par rapport au précédent, parce que $(x-a) \to 0$ et $\delta(x,a) \to 0$ lorsque $x \to a$. On a donc une approximation « vague » du terme $f(x)$ en le remplaçant par $f(a)$ (c'est la continuité, voir ci-dessous) ; une « plus précise » en prenant $f(a)+(x-a)f'(a)$ ($y = f(a)+(x-a)f'(a)$ est une équation de la droite tangente au graphe de f en a) ; une exacte avec le δ, qui est « très petit », mais habituellement inconnu.
Pour les fonctions les plus simples, on peut utiliser directement la définition pour démontrer la dérivabilité.

Exemples.

- Il est évident que les fonctions constantes ont une dérivée nulle puisque le taux d'accroissement est nul.
- La fonction $x \mapsto x$ est dérivable en tout point avec une dérivée égale à 1.
- Soient n un entier naturel non nul et a un réel non nul.
  On vérifie facilement la formule :
  $$\forall\, x,\ a \in \mathbb{R}, \quad x^n - a^n = (x-a)(a^{n-1} + xa^{n-2} + \cdots + x^{n-2}a + x^{n-1}).$$
  On en déduit, pour tout $x \neq a$:
  $$\frac{x^n - a^n}{x-a} = a^{n-1} + xa^{n-2} + \cdots + x^{n-2}a + x^{n-1}.$$
  Lorsque x tend vers a, avec $x \neq a$, le membre de droite tend vers $n\,a^{n-1}$, ce qui démontre que la fonction $x \mapsto x^n$ est dérivable en tout point $a \in \mathbb{R}$, avec $n\,a^{n-1}$ pour dérivée. Notons que cette formule vaut aussi pour $n = 0$, puisque la dérivée d'une constante est nulle.

Soient $I \subset \mathbb{R}$ un intervalle et a un point *intérieur* à I.
Il existe $\varepsilon > 0$ tel que $]a-\varepsilon, a+\varepsilon[\subset I$. On en déduit que $I \cap [a, +\infty[$ et $I \cap]-\infty, a]$ sont des intervalles d'intérieur non vide.

Définition 12. Soient $I \subset \mathbb{R}$ un intervalle et a un point *intérieur* à I.
On note $I_+ := I \cap [a, +\infty[$ (resp. $I_- := I \cap]-\infty, a]$). Soit $f : I \to \mathbb{R}$. On dit que f est *dérivable à droite* (resp. *à gauche*) au point a si la restriction de f à I_+ (resp. I_-) est dérivable au point a. La dérivée de la restriction de f à I_+ (resp. I_-) au point a s'appelle *dérivée à droite* (resp. *à gauche*) au point a et elle est notée $f'_d(a)$ (resp. $f'_g(a)$).

Quand elles existent, les dérivées à droite et à gauche sont données par :

$$\lim_{x \to a^+} = \frac{f(x) - f(a)}{x - a}, \qquad \lim_{x \to a^-} = \frac{f(x) - f(a)}{x - a}.$$

On vérifie immédiatement le résultat suivant.

Proposition 33. Soient $I \subset \mathbb{R}$ un intervalle et $a \in I$ un point intérieur. Pour que f soit dérivable au point a, il faut et il suffit que f soit dérivable à droite et à gauche au point a et que l'on ait $f'_d(a) = f'_g(a)$.

Exemple. La fonction valeur absolue $x \in \mathbb{R} \mapsto |x| \in \mathbb{R}$ est dérivable à droite et à gauche en 0. La dérivée à droite (resp. à gauche) est 1 (resp. -1). La fonction n'est donc pas dérivable en 0.

Proposition 34. Toute fonction dérivable au point a est continue en a.

Démonstration. Il suffit de noter que les deux derniers termes de (10) tendent vers zéro lorsque x tend vers a. ∎

Noter que la réciproque est fausse : la fonction $x \mapsto |x|$ par exemple est continue, mais non dérivable en $x = 0$ (le taux d'accroissement admet des limites à gauche et à droite différentes).

Exercice 8.
Soit f une fonction dérivable à gauche et à droite en a. Montrer qu'il existe une fonction f_1 dérivable en a, et un nombre réel γ tels que $f(x) = f_1(x) + \gamma\,|x-a|$ pour tout $x \in D_f$.
Solution. Nous venons de voir que la dérivée à droite de la valeur absolue est 1, alors que sa dérivée à gauche est -1. Pour tout $\gamma \in \mathbb{R}$, celles de $f - \gamma\,|x-a|$ sont donc $k - \gamma$ et $h + \gamma$ où $k := f'_d(a)$ et $h := f'_g(a)$. Choisissons $\gamma := \frac{k-h}{2}$ et notons $d := \frac{k+h}{2}$.
On a $k - \gamma = h + \gamma$, donc f_1, définie par l'égalité de l'énoncé, est dérivable en a et de dérivée d d'après la proposition 33.

Définition 13. On dit qu'une fonction f est dérivable sur un ensemble $I \subset \mathbb{R}$ si elle est dérivable en tout point $a \in I$.

Dans ce cas, on obtient une nouvelle fonction f' définie sur I, appelée la *fonction dérivée* de f.

Exemple. Comme nous l'avons vu précédemment, la fonction $\varphi_n : x \mapsto x^n$, avec $n \in \mathbb{N}^*$, est dérivable sur $\mathbb{R}$ et sa fonction dérivée est $n\,\varphi_{n-1}$.

On montre dans l'exercice suivant qu'une fonction dérivée n'est pas nécessairement continue.

Exercice 9.
Montrer que la fonction f définie par $f(0) = 0$ et $f(x) = x^2 \sin 1/x$ pour $x \neq 0$ est dérivable en 0 et que $f'(0) = 0$.
Calculer $f'(x)$ pour $x \neq 0$ (en admettant les règles décrites dans les paragraphes suivants), et montrer que $f'(x)$ n'a pas de limite lorsque x tend vers zéro.
Ainsi la fonction f' n'est pas continue en 0.

Solution. En effet : $\dfrac{f(x) - f(0)}{x - 0} = x \sin \dfrac{1}{x} \underset{x \to 0}{\longrightarrow} 0$.
Pour $x \neq 0$, le calcul est facile ; il donne :

$$f'(x) = 2x \sin \frac{1}{x} - \cos \frac{1}{x}$$

quantité dont le premier terme tend vers 0 avec x, mais dont le second n'a pas de limite comme on le voit en étudiant ses valeurs pour $x := 1/(n\pi)$, $(n \in \mathbb{N}^*)$.

Remarque. Les mathématiciens ont longtemps cru que les fonctions réelles continues étaient les seules à posséder la propriété des valeurs intermédiaires. Mais, en 1875, le mathématicien français Gaston Darboux a montré que les dérivées de fonctions réelles possédaient la propriété des valeurs intermédiaires (*cf.* l'exercice IV.2.17 de la page 673 pour une preuve). On a vu dans l'exercice 9 ci-dessus qu'une fonction dérivée n'est pas nécessairement continue, ce qui donne un exemple de fonction qui possède la propriété des valeurs intermédiaires sans être continue. Plus généralement, G. Darboux a aussi montré, en utilisant des séries de fonctions, qu'il existait des fonctions dérivables dont la dérivée n'est continue sur aucun intervalle.

2.2 Opérations sur les dérivées

En pratique, pour montrer la dérivabilité d'une fonction et calculer sa dérivée, plutôt que d'utiliser la définition, on se ramène au cas de quelques fonctions « classiques » en utilisant les opérations élémentaires habituelles sur les fonctions. En effet, lorsque l'on effectue de telles opérations sur des fonctions dérivables, les résultats obtenus sont généralement dérivables comme nous allons le voir.

Théorème 35. Soient f et g deux fonctions dérivables au point a. Alors $f + g$, $f - g$, fg sont dérivables en a, ainsi que f/g si $g(a) \neq 0$. On a de plus les formules :

$$(f+g)'(a) = f'(a) + g'(a), \qquad (f-g)'(a) = f'(a) - g'(a),$$

$$(fg)'(a) = f'(a)g(a) + f(a)g'(a), \qquad (f/g)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{\big(g(a)\big)^2}.$$

Ce théorème s'applique notamment aux fonctions constantes. On retrouve ainsi que la dérivée de λf est $\lambda f'$, ou que la dérivée de $1/g$ est $-g'/g^2$.

Démonstration. Pour $f+g$ et $f-g$, c'est évident parce que les taux d'accroissement eux-mêmes s'additionnent. Par exemple pour $f+g$, on a :

$$\begin{aligned}\frac{(f+g)(x)-(f+g)(a)}{x-a} &= \frac{f(x)+g(x)-f(a)-g(a)}{x-a}\\ &= \frac{f(x)-f(a)}{x-a}+\frac{g(x)-g(a)}{x-a}.\end{aligned}$$

Il suffit donc de faire tendre x vers a pour obtenir le résultat en utilisant les opérations sur les limites.

Pour le produit, on écrit (en ajoutant et retranchant le même terme au numérateur) :

$$\begin{aligned}\frac{(fg)(x)-(fg)(a)}{x-a} &= \frac{f(x)g(x)-f(a)g(x)+f(a)g(x)-f(a)g(a)}{x-a}\\ &= \frac{f(x)-f(a)}{x-a}\,g(x)+f(a)\,\frac{g(x)-g(a)}{x-a}.\end{aligned}$$

Lorsque x tend vers a, les taux d'accroissements tendent vers les dérivées, et $g(x)$ tend vers $g(a)$ d'après la proposition 34 de la page 641.

De même pour le quotient (qui est bien défini si $g(a)\neq 0$; en effet g étant continue en a, on obtient que $g(x)\neq 0$ pour x assez proche de a) :

$$\begin{aligned}\frac{(f/g)(x)-(f/g)(a)}{x-a} &= \frac{f(x)g(a)-f(a)g(x)}{(x-a)g(x)g(a)}\\ &= \frac{\dfrac{f(x)-f(a)}{x-a}\,g(a)-f(a)\,\dfrac{g(x)-g(a)}{x-a}}{g(x)g(a)}\end{aligned}$$

qui tend bien vers la valeur annoncée. ∎

Théorème 36. Soient J un intervalle de $\mathbb{R}$, $f\,:\,I\to J$ une fonction dérivable en $a\in I$, et $g\,:\,J\to\mathbb{R}$ une fonction dérivable en $b:=f(a)$.
Alors $g\circ f$ est dérivable en a, et l'on a :

$$(g\circ f)'(a)=g'(b)\cdot f'(a)=(g'\circ f)(a)\cdot f'(a). \tag{11}$$

Autrement dit, la dérivée d'une composée est le produit des dérivées (aux points *ad-hoc*, évidemment).

Démonstration. Soient α et β les deux fonctions définies par les égalités :

$$f(x)=f(a)+(x-a)\big(f'(a)+\alpha(x)\big),\qquad g(y)=g(b)+(y-b)\big(g'(b)+\beta(y)\big).$$

Elles ne sont pas définies pour $x:=a$ ou $y:=b$, mais par définition elles admettent en ces points respectivement des limites nulles. On peut donc les prolonger par continuité en posant $\alpha(a):=0$ et $\beta(b):=0$.

Posons $y:=f(x)$. On note que $\beta(y)$ tend alors vers 0 quand x tend vers a. De plus :

$$(g\circ f)(x)-(g\circ f)(a)=g\big(f(x)\big)-g\big(f(a)\big)=g(y)-g(b)$$

c'est-à-dire $(y-b)\big(g'(b)+\beta(y)\big)$. Ainsi :

$$\big(f(x)-f(a)\big)\big(g'(b)+\beta(y)\big)=(x-a)\big(f'(a)+\alpha(x)\big)\big(g'(b)+\beta(y)\big).$$

Après division par $x-a$, il reste donc à démontrer que :

$$\gamma : x \mapsto \big(f'(a)+\alpha(x)\big)\big(g'(b)+\beta\big(f(x)\big)\big)$$

tend vers $f'(a)\,g'(b)$ lorsque x tend vers a. Comme $\gamma(a)$ est justement égal à $f'(a)\,g'(b)$ (grâce au prolongement par continuité de α et β), cela provient de la continuité de γ en a, comme produit et composée de fonctions continues. ■

Exemples.

- La dérivée de $x \mapsto \sin(2x+1)$ en a est $2\cos(2a+1)$.
- La dérivée de $x \mapsto \mathrm{e}^{-x^2}$ en a est $-2a\,\mathrm{e}^{-a^2}$.

La formule (11) est souvent plus facile à manipuler en utilisant la notation $\dfrac{\mathrm{d}f}{\mathrm{d}x}$ pour f'.

C'est ainsi qu'elle est écrite traditionnellement dans les livres de physique ou de mécanique. Considérons, par exemple, un point mobile se déplaçant sur un axe et repéré par son abscisse x. Une grandeur physique V dépendant de la position du mobile est représentée par une fonction $f : x \mapsto V = f(x)$. D'autre part, la position du mobile en fonction du temps est représentée par une fonction $x : t \mapsto x(t)$, ce qui fait que les « variations » de V en fonction du temps sont données par la dérivée de la fonction $g = f \circ x : t \mapsto f\big(x(t)\big)$, c'est-à-dire :

$$g'(t) = f'\big(x(t)\big)\,x'(t) \qquad \text{ou encore} \qquad \frac{\mathrm{d}g}{\mathrm{d}t}(t) = \frac{\mathrm{d}f}{\mathrm{d}x}\big(x(t)\big)\,\frac{\mathrm{d}x}{\mathrm{d}t}(t)\cdot$$

En physique, on écrit plutôt $\dfrac{\mathrm{d}V}{\mathrm{d}t}(t) = \dfrac{\mathrm{d}V}{\mathrm{d}x}\big(x(t)\big)\,\dfrac{\mathrm{d}x}{\mathrm{d}t}(t)$, en notant abusivement de la même manière les fonctions $f : x \mapsto V = f(x)$ et $g : t \mapsto V = f\big(x(t)\big)$ et même, en supprimant les variables, $\dfrac{\mathrm{d}V}{\mathrm{d}t} = \dfrac{\mathrm{d}V}{\mathrm{d}x}\,\dfrac{\mathrm{d}x}{\mathrm{d}t}$, ce qui rend la formule particulièrement facile à retenir.

Il faut noter que le théorème 35 de la page 642 se généralise sans problème aux dérivées à gauche et à droite, mais *pas* le théorème 36 de la page précédente : le résultat reste vrai, mais la démonstration est bien plus délicate, car $x \overset{>}{\to} a$ n'entraîne pas que $f(x) \overset{>}{\to} f(a)$ par exemple.

2.3 Dérivées d'ordre n

On définit les dérivées successives d'une fonction réelle f par récurrence.

Définition 14. Soient $I \subset \mathbb{R}$ un intervalle et $f : I \to \mathbb{R}$.
On définit par récurrence sur $n \in \mathbb{N}$ la notion de *fonction* dérivée d'ordre n de f sur I. On note $f^{(0)} = f$. On suppose connue la fonction dérivée $(n-1)^{\text{ème}}$: $f^{(n-1)} : I \to \mathbb{R}$ et l'on dit que f admet une fonction dérivée d'ordre n sur I si $f^{(n-1)}$ est dérivable sur I ; on note $f^{(n)} = \big(f^{(n-1)}\big)'$. On dit aussi que f est n *fois dérivable sur* I.

On a ainsi $f^{(0)} = f$ et $f^{(1)} = f'$ lorsque f est dérivable. Lorsque f est deux fois dérivable, on note $f'' = f^{(2)}$.

La notion de dérivée d'ordre n *en un point* est plus délicate.

Définition 15. Soient, $n \in \mathbb{N}^*$ $(n \geqslant 2)$, $I \subset \mathbb{R}$ un intervalle, $a \in I$, et $f : I \to \mathbb{R}$. On dit que f admet une dérivée d'ordre n au point a, s'il existe $\varepsilon > 0$ tel que f admette une fonction dérivée d'ordre $(n-1)$ sur l'intervalle $I \cap]a-\varepsilon, a+\varepsilon[$ et que $f^{(n-1)}$ soit dérivable en a. On note alors $f^{(n)}(a)$ la *dérivée* $n^{ème}$ *au point* a.

On vérifie facilement que f est n fois dérivable sur I si, et seulement si, elle admet une dérivée $n^{ème}$ en tout point de I.

Définition 16. Une fonction f sur un intervalle I est dite de classe $\mathcal{C}^n$ (ou n *fois continûment dérivable*) sur I si elle admet une dérivée d'ordre n et si cette dérivée est continue.
Elle est de classe $\mathcal{C}^\infty$ si elle est de classe $\mathcal{C}^n$ pour tout n.

Une fonction de classe $\mathcal{C}^1$ est aussi dite *continûment dérivable*. Une fonction continue est aussi dite de classe $\mathcal{C}^0$.
Notons la proposition suivante, dont la démonstration est immédiate :

Proposition 37. Toute fonction admettant une dérivée d'ordre n sur un intervalle I y admet des dérivées d'ordre k pour tout $k \in [\![1, n]\!]$.

Il est évident que si f est n fois dérivable, alors pour $(p, q) \in \mathbb{N}^*$ tel que $p + q = n$, la fonction f est p fois dérivable et $f^{(p)}$ est q fois dérivable, et l'on a $\left(f^{(p)}\right)^q = f^{(n)}$.
En particulier, f est n fois dérivable, avec $n \geqslant 1$, si, et seulement si, f est dérivable avec une dérivée $n-1$ fois dérivable. On a alors $(f')^{(n-1)} = f^{(n)}$.
Les opérations sur les dérivées s'étendent aux dérivées successives. En particulier, la linéarité de la dérivation $D : f \mapsto f'$ donne, par composition, la linéarité de la dérivation $D^n : f \mapsto f^{(n)}$ d'ordre n.
On prouve facilement le résultat suivant.

Proposition 38. Soit $n \in \mathbb{N}$.
Si f et g sont n fois dérivables sur I, alors pour tout $(\lambda, \mu) \in \mathbb{R}^2$, la fonction $\lambda f + \mu g$ est n fois dérivable sur I et $(\lambda f + \mu g)^{(n)} = \lambda f^{(n)} + \mu g^{(n)}$.

Le résultat subsiste aussi pour le produit, mais la formule donnant la dérivée $n^{ème}$ (appelée formule de Leibniz) est plus compliquée.

Proposition 39 (Formule de Leibniz). Soit $n \in \mathbb{N}$. Si f et g sont n fois dérivables sur I, alors la fonction $f\,g$ est n fois dérivable sur I et l'on a :

$$(f\,g)^{(n)} = \sum_{p=0}^{n} \binom{n}{p} f^{(p)}\, g^{(n-p)}.$$

Démonstration. Démontrons ce résultat par récurrence sur n.

- La propriété est évidente pour $n = 0$.

- Supposons le résultat vrai pour un entier naturel n et considérons deux fonctions f et g supposées $n+1$ fois dérivables sur I. Elles sont donc n fois dérivables et l'hypothèse de récurrence permet d'écrire :

$$(f\,g)^{(n)} = \sum_{p=0}^{n} \binom{n}{p} f^{(p)}\, g^{(n-p)}.$$

La fonction $(f\,g)^{(n)}$ est une combinaison linéaire de produits de fonctions dérivables ; elle est donc dérivable et par suite $f\,g$ est $n+1$ fois dérivable. De plus, on a :

$$\begin{aligned}
(f\,g)^{(n+1)} &= \left(\sum_{p=0}^{n} \binom{n}{p} f^{(p)}\, g^{(n-p)}\right)' \\
&= \sum_{p=0}^{n} \binom{n}{p} \left(f^{(p)}\, g^{(n-p+1)} + f^{(p+1)}\, g^{(n-p)}\right) \\
&= \binom{n}{0} f\, g^{(n+1)} + \sum_{p=1}^{n} \binom{n}{p} f^{(p)}\, g^{(n-p+1)} + \sum_{p=0}^{n-1} \binom{n}{p} f^{(p+1)}\, g^{(n-p)} + \binom{n}{n} f^{(n+1)}\, g \\
&= \binom{n+1}{0} f\, g^{(n+1)} + \binom{n+1}{n+1} f^{(n+1)}\, g + \sum_{p=1}^{n} \left(\binom{n}{p} + \binom{n}{p-1}\right) f^{(p)}\, g^{(n-p+1)} \\
&\qquad\qquad \text{(changement d'indice et regroupement des deux sommes)} \\
&= \sum_{p=0}^{n+1} \binom{n+1}{p} f^{(p)}\, g^{(n+1-p)}
\end{aligned}$$

ce qui prouve le résultat pour $n+1$. ■

Pour la composition, le problème est encore plus compliqué car il n'y a pas de formule simple[4] donnant la dérivée $n^{ème}$ de $g \circ f$. Il faut aussi faire attention aux intervalles de définition qui sont, *a priori*, différents.

Proposition 40. Soient f une fonction n fois dérivable sur un intervalle I, à valeurs dans un intervalle J, et g une fonction n fois dérivable sur J. Alors $g \circ f$ est n fois dérivable sur I.

Démonstration. Démontrons ce résultat par récurrence sur n. Il n'y a rien à montrer pour $n = 0$. Supposons donc le résultat acquis pour un certain entier $n \in \mathbb{N}$. Soient f une fonction $(n+1)$ fois dérivable sur un intervalle I, à valeurs dans un intervalle J, et g une fonction $(n+1)$ fois dérivable sur J. Alors $g \circ f$ est dérivable et sa dérivée est $(g \circ f)' = \big(g' \circ f\big)\, f'$. Par l'hypothèse de récurrence, g' et f étant n fois dérivables respectivement sur J et sur I, la fonction $g' \circ f$ est n fois dérivable sur I. Par produit avec f', également n fois dérivable, on en déduit que $(g \circ f)'$ est n fois dérivable, ce qui prouve que $g \circ f$ est $(n+1)$ fois dérivable. ■

[4] Il existe une formule combinatoire un peu compliquée dite *formule de Fáa di Bruno*. Elle est vraisemblablement due, en 1800, au mathématicien alsacien Louis François Antoine Arbogast et elle a été retrouvée, en 1855, par le mathématicien italien Francesco Fáa di Bruno .

2.4 Sens de variation et extrema

Venons-en maintenant aux applications de la dérivation. Un des points les plus importants est de comprendre les relations entre le signe de la dérivée d'une fonction et le sens de ses variations. Pour les notions de croissance et décroissance, *cf.* la définition 6 de la page 628.

Proposition 41. Si f est croissante sur un intervalle I et dérivable en $a \in I$, alors $f'(a) \geqslant 0$. Si f est décroissante, alors $f'(a) \leqslant 0$.

Démonstration. Si f est croissante, alors $x < a \Rightarrow f(x) \leqslant f(a)$ et $x > a \Rightarrow f(x) \geqslant f(a)$. Dans les deux cas le taux d'accroissement $\frac{f(x)-f(a)}{(x-a)}$ est positif ou nul ; la limite est donc aussi positive ou nulle. ■

Remarque. La proposition 41 ci-dessus reste évidemment vraie si l'on suppose seulement que f est dérivable à droite (resp. à gauche). Par exemple, si f est croissante sur un intervalle I et dérivable *à droite* (resp. à gauche) en $a \in I$, alors $f'_d(a) \geqslant 0$ (resp. $f'_g(a) \geqslant 0$).

Une question naturelle est de savoir si la réciproque de la proposition 41 est vraie : si $f' \geqslant 0$ sur un intervalle, a-t-on f croissante ? La réponse est oui, mais ne peut se déduire des résultats démontrés jusqu'ici : voir plus loin le théorème 45 de la page 649.

Si la fonction est strictement croissante, il n'est pas vrai que $f'(a) > 0$ (considérer le cas de la fonction $x \mapsto x^3$: elle est strictement croissante, mais sa dérivée est nulle en $a = 0$).

Définition 17. Soit f une fonction définie sur un intervalle I. On dit que f admet *un maximum local* en $a \in I$ s'il existe $\varepsilon > 0$ tel que pour tout $x \in I \cap]a-\varepsilon, a+\varepsilon[$, on ait $f(x) \leqslant f(a)$. Si l'inégalité est vraie pour tous les $x \in I$, on dit que f a un *maximum global* en a. Si c'est l'inégalité inverse $f(x) \geqslant f(a)$ qui a lieu, on parle de *minimum* local ou global.
Dans l'un ou l'autre cas, on dit que a est un *extremum* (local ou global) de f.

Si les inégalités sont strictes ($f(x) > f(a)$, pour $x \neq a$, ou le contraire) on parle d'un *maximum ou minimum strict*.

Proposition 42. Soit f une fonction dérivable sur un intervalle *ouvert* I. Si elle admet un extremum local en $a \in I$, alors $f'(a) = 0$.

La réciproque de cette proposition est fausse : on peut avoir $f'(a) = 0$ sans que a soit un extremum : considérer encore le cas de $x \mapsto x^3$ et $a = 0$. Notons aussi que l'on se restreint à un intervalle ouvert, parce que si a est une extrémité de l'intervalle, la proposition devient fausse : par exemple la fonction $x \mapsto x$ a un extremum local en 0 sur l'intervalle fermé $[0, 1]$, mais sa dérivée ne s'annule pas en 0.

Démonstration. Considérons par exemple le cas d'un maximum local. Notons $I =]\alpha, \beta[$ Comme $\alpha < a < \beta$, il existe des $x \in I$ tels que $x < a$; s'ils sont assez proches de a, on a

alors $f(x) \leqslant f(a)$ par définition d'un maximum local. Dès lors $(f(x)-f(a))/(x-a) \geqslant 0$ et donc à la limite $f'(a) \geqslant 0$. Inversement si $x > a$ est proche de a, alors $f(x) \leqslant f(a)$ implique $\big(f(x)-f(a)\big)/(x-a) \leqslant 0$, et donc $f'(a) \leqslant 0$ à la limite. Donc $f'(a) = 0$. ■

2.5 Théorème de Rolle, accroissements finis

Le théorème de Rolle exprime qu'une fonction continue qui prend deux fois la même valeur doit admettre un extremum local ; si elle est dérivable, sa dérivée s'annule en ce point :

Théorème 43 (Rolle). Soit f une fonction continue sur un intervalle $[a,b]$, avec $a < b$, dérivable sur $]a,b[$, et telle que $f(a) = f(b)$.
Alors il existe un point $c \in]a,b[$ tel que $f'(c) = 0$.

Noter la subtilité que f est continue en a et b, mais n'a pas besoin d'être dérivable en a et b. Le graphe ci-contre donne un exemple d'une fonction vérifiant les hypothèses du théorème de Rolle. Noter qu'il peut y avoir plusieurs points où la dérivée s'annule ; le théorème affirme seulement qu'il y en a au moins un.

Démonstration. La fonction f étant continue sur le segment $[a,b]$, elle est bornée et atteint sa borne inférieure α et sa borne supérieure β (*cf.* le théorème 32 de la page 638). Prenons donc c et d dans $[a,b]$ tels que $f(c) = \alpha$ et $f(d) = \beta$.
Si $\alpha = \beta$, alors la fonction est en fait constante, et donc en tous les points $c \in]a,b[$, la dérivée s'annule.
Sinon, on a $\alpha < \beta$, et donc l'un des deux est différent de $f(a) = f(b)$. Disons par exemple que $f(c) = \alpha < f(a) = f(b)$. Donc $c \neq a$ et $c \neq b$, soit $c \in]a,b[$ et $f'(c) = 0$ par la proposition 42 de la page précédente. ■

Naturellement si $f(a) \neq f(b)$, il n'y a aucune raison que la dérivée s'annule (considérer le cas de la fonction $x \mapsto x$). Mais la dérivée doit, au moins en un point, être égale au taux d'accroissement entre a et b.

Théorème 44 (des accroissements finis).
Soit f une fonction continue sur un segment $[a,b]$, dérivable sur $]a,b[$. Alors il existe au moins un point $c \in]a,b[$ tel que :

$$f'(c) = \frac{f(b)-f(a)}{b-a}. \tag{12}$$

L'égalité ci-dessus s'appelle aussi *égalité des accroissements finis*. Le taux d'accroissement est le rapport entre l'accroissement de la fonction et celui de la variable. Les inventeurs du calcul infinitésimal parlaient d'accroissements finis par opposition à la notion « d'accroissement infinitésimal ».

Ce théorème signifie graphiquement que si Γ est le graphe de la fonction f dans le plan $\mathbb{R}^2$ et si A et B sont les points de Γ d'abscisses respectives a et b, alors il existe au moins un point de Γ en lequel la tangente à Γ est parallèle à la droite (AB).
Ce théorème est plus général que celui de Rolle (si $f(a) = f(b)$ on retrouve bien $f'(c) = 0$), mais nous le démontrons à partir de ce dernier.

Démonstration. Notons $\gamma := \dfrac{f(b) - f(a)}{b - a}$ le taux d'accroissement, et considérons la fonction $g : x \mapsto f(x) - \gamma(x - a)$. Alors g est continue sur $[a, b]$, et dérivable sur $]a, b[$, comme f. De plus $g(a) = f(a)$ et :

$$g(b) = f(b) - \frac{f(b) - f(a)}{b - a}(b - a) = f(b) - \big(f(b) - f(a)\big) = f(a) = g(a).$$

On peut donc appliquer le théorème de Rolle à g, et en déduire qu'il existe $c \in]a, b[$ tel que $0 = g'(c) = f'(c) - \gamma$. C'est ce qu'il fallait démontrer. ■

Malgré son air « anodin », le théorème des accroissements finis a de nombreuses conséquences. D'abord, il permet de démontrer une réciproque de la proposition 41 de la page 647.
L'énoncé ci-dessous utilise la notion d'intérieur d'un intervalle : *cf.* la définition 9 de la page 520.

Théorème 45. Soit f une fonction continue sur un intervalle I et dérivable sur l'intérieur I' de I. Alors f est croissante sur I si, et seulement si, $f' \geqslant 0$ sur I'.
De même f est décroissante sur I si, et seulement si, $f' \leqslant 0$ sur I'.

Démonstration. On sait déjà que si f est croissante, alors $f' \geqslant 0$ (proposition 41 de la page 647). Supposons maintenant que $f' \geqslant 0$ par exemple. Soient x et y deux éléments de I, avec $x < y$. Alors le segment $[x, y]$ est contenu dans I et l'intervalle ouvert $]x, y[$ dans I', donc f est continue sur $[x, y]$ et dérivable sur $]x, y[$ et le théorème des accroissements finis implique qu'il existe $c \in]x, y[\subset I$ tel que $f(y) - f(x) = f'(c)(y - x)$. Cette quantité est un produit de deux nombres positifs, puisque $f' \geqslant 0$ par hypothèse, donc elle est positive ; ainsi $f(y) \geqslant f(x)$, ce qui montre que f est croissante. Le cas $f' \leqslant 0$ se montre de la même façon, ou en appliquant la propriété précédente à $-f$. ■

Corollaire 46. Soient f et g deux fonction continues sur un segment $[a, b]$, dérivables sur $]a, b[$ et telles que, pour tout $x \in]a, b[$ $f'(x) \leqslant g'(x)$.
Alors $f(b) - f(a) \leqslant g(b) - g(a)$.

Démonstration. On pose $h := f - g$.
On a $h' \leqslant 0$ sur $]a, b[$, donc $h(b) \leqslant h(a)$, c'est-à-dire $f(b) - f(a) \leqslant g(b) - g(a)$. ■

Comme nous l'avons vu précédemment, le théorème 45 ne s'étend pas aux fonctions strictement croissantes (ou décroissantes) : pour celles-ci on peut seulement affirmer que $f' \geqslant 0$, mais pas que $f' > 0$. Mais la réciproque subsiste, la démonstration étant identique à celle du théorème 45 :

Proposition 47. Toute fonction f continue sur l'intervalle I, dérivable sur l'intérieur I' de I avec une dérivée vérifiant $\forall x \in I'$, $f'(x) > 0$, est strictement croissante sur I.

Démonstration. La démonstration précédente convient, puisque le théorème des accroissements finis affirme l'existence d'un nombre c appartenant à l'ouvert $]a, b[$ et non pas seulement à $[a, b]$. ∎

Enfin, une fonction étant constante si, et seulement si, elle est à la fois croissante et décroissante, on déduit du théorème 45 de la page précédente une caractérisation des fonctions constantes sur un intervalle :

Corollaire 48. Une fonction continue sur un intervalle I et dérivable sur l'intérieur I' de I est constante sur I si, et seulement si, $f' = 0$ sur I'.

Exercice 10.

Soit f une fonction continue sur un intervalle I, dérivable à dérivée strictement positive en tout point de I sauf un nombre fini. Montrer que f est strictement croissante sur I.

Solution. Soit $a < b$ dans I. Notons $x_1 < x_2 < \cdots < x_{n-1}$ les points de $]a, b[$ en lesquels f n'est pas dérivable ou admet une dérivée nulle et posons $x_0 := a$ et $x_n := b$. La proposition 47 nous donne la stricte croissance de f sur chacun des segments $[x_{k-1}, x_k]$. En particulier, $f(a) = f(x_0) < f(x_1) < \cdots < f(x_n) = f(b)$.

Exercice 11.

On suppose que la fonction exponentielle est définie « comme en terminale » (*cf.* le corollaire 6 de la page 681) : c'est l'unique fonction $f : \mathbb{R} \to \mathbb{R}$, dérivable sur $\mathbb{R}$, égale à sa dérivée et telle que $f(0) = 1$. Montrer que l'exponentielle est strictement positive. Montrer que l'exponentielle est strictement croissante.

Solution. On considère la fonction $f : x \mapsto f(x) := e^x e^{-x}$. On a, en utilisant la formule de dérivation des fonctions composées, $f'(x) = e^x e^{-x} - e^x e^{-x} = 0$ et $f(0) = 1$, donc $f = 1$. Par suite $e^x \neq 0$ pour tout $x \in \mathbb{R}$. S'il existait $a \in \mathbb{R}$ tels que $e^a < 0$, d'après le théorème des valeurs intermédiaires, il existerait un point b de l'intervalle d'extrémités 0 et a tel que $e^b = 0$, ce qui est impossible. Par suite $e^x > 0$ pour tout $x \in \mathbb{R}$. On en déduit que l'exponentielle est strictement croissante sur $\mathbb{R}$ en utilisant la proposition 47.

Voici un autre corollaire du théorème des accroissements finis, très important dans la pratique.

Corollaire 49 (Inégalité des accroissements finis). Soit f une fonction dérivable sur un intervalle I et telle qu'il existe $m \in \mathbb{R}_+$ avec $|f'(x)| \leqslant m$ pour tout $x \in I$. Alors :

$$\forall x \in I,\ \forall y \in I, \qquad |f(y) - f(x)| \leqslant m|y - x|. \tag{13}$$

Autrement dit f est m-lipschitzienne sur I.

Démonstration. La preuve est immédiate : il suffit d'appliquer l'égalité des accroissements finis et de prendre la valeur absolue. ∎

L'intérêt de ce corollaire est que l'on a souvent des bornes sur les dérivées (voir l'exemple ci-dessous), et qu'en outre il se généralise aux fonctions à valeurs vectorielles, *contrairement* au théorème des accroissements finis.

Exemple. Puisque, pour tout $x \in \mathbb{R}$, $|\cos x| \leqslant 1$ et $|-\sin x| \leqslant 1$, les fonctions sinus et cosinus sont 1-lipschitziennes sur $\mathbb{R}$.

Le résultat suivant est appelé formule des accroissements finis généralisés ou formule de la moyenne de Cauchy.

Proposition 50 (formule de la moyenne de Cauchy). Soient $f,\ g : [a,b] :\to \mathbb{R}$ des fonctions continues sur $[a,b]$, dérivables sur $]a,b[$. Alors il existe $c \in]a,b[$ tel que :

$$\begin{vmatrix} f(b)-f(a) & f'(c) \\ g(b)-g(a) & g'(c) \end{vmatrix} = 0. \tag{14}$$

Démonstration. Si $f(a) = f(b)$ et $g(a) = g(b)$, le résultat est trivial. Sinon, en supposant par exemple $g(a) \neq g(b)$, on peut appliquer le théorème de Rolle à la fonction :

$$x \mapsto f(x) - f(a) - \frac{f(b)-f(a)}{g(b)-g(a)}\,(g(x)-g(a))\,. \qquad \blacksquare$$

Le résultat suivant est appelé règle de l'Hôpital[5].

Proposition 51 (règle de l'Hôpital). Soient $I \subset \mathbb{R}$ un intervalle, $a \in \mathbb{R}$ et $f,\ g : I \to \mathbb{R}$ deux applications continues sur I, dérivables sur $I \setminus \{a\}$ et vérifiant :

$$f(a) = g(a) = 0 \quad \text{et} \quad \forall x \in I \setminus \{a\},\ \ g'(x) \neq 0.$$

Alors, $\forall x \in I \setminus \{a\},\ \ g(x) \neq 0$ et, si $\lim\limits_{x\to a} \dfrac{f'(x)}{g'(x)} = \ell$, alors $\lim\limits_{x\to a} \dfrac{f(x)}{g(x)} = \ell$.

Démonstration. Le résultat se déduit de la proposition 50 ci-dessus.
D'après le théorème de Rolle l'hypothèse entraîne $g(x) \neq 0$, pour tout $x \in I \setminus \{a\}$.
Par suite, à tout $x \in I \setminus \{a\}$, on peut associer $\xi(x)$, compris entre a et x, tel que :

$$\frac{f(x)}{g(x)} = \frac{f\left(\xi(x)\right)}{g\left(\xi(x)\right)}.$$

Pour tout $\varepsilon > 0$, il existe $\alpha > 0$ tel que $0 < |u-a| < \alpha$ entraîne $\left|\frac{f'(u)}{g'(u)} - \ell\right| < \varepsilon$.

Donc, si $0 < |x-a| < \alpha$, l'on a $0 < |\xi(x)-a| < \alpha$ et $\left|\frac{f(x)}{g(x)} - \ell\right| = \left|\frac{f'(\xi(x))}{g'(\xi(x))} - \ell\right| < \varepsilon$. $\blacksquare$

La règle de l'Hôpital sert à « lever des indéterminations du type $\frac{0}{0}$ ».
Nous verrons plus tard, dans le module IV.8, une technique plus efficace basée sur la notion de développement limité, *cf.* le point méthode de la page 872.

Exemple. En appliquant deux fois la règle de l'Hôpital, on obtient :

$$\lim_{x\to 0} \frac{x^2}{1-\cos 4x} = \lim_{x\to 0} \frac{2x}{4\sin 4x} = \lim_{x\to 0} \frac{2}{16\cos 4x} = \frac{1}{8}.$$

[5] Il figure dans un traité de 1696 du mathématicien français Guillaume François Antoine marquis de l'Hôpital.

2.6 Formules de Taylor

Les formules de Taylor sont des généralisations de la formule des accroissements finis. Nous reviendrons sur les formules de Taylor dans le module IV.8 en utilisant la notion d'intégrale.

2.6.1 Formule et inégalité de Taylor-Lagrange

Théorème de Taylor-Lagrange

Le théorème de Taylor-Lagrange est une généralisation du théorème des accroissements finis. On le démontre de façon analogue en utilisant le théorème de Rolle. Comme le théorème des accroissements finis, le théorème de Taylor-Lagrange n'est pas vrai, en général, pour des fonctions à valeurs complexes ou à valeurs vectorielles.

Théorème 52 (Théorème de Taylor-Lagrange). Soient $n \in \mathbb{N}$ et f une fonction n fois continûment dérivable sur un segment $[a, b]$ et dont la dérivée $(n+1)^{ème}$ existe sur $]a, b[$. Alors, il existe $c \in]a, b[$ tel que :

$$f(b) = f(a) + \sum_{k=1}^{n} \frac{(b-a)^k}{k!} f^{(k)}(a) + \frac{(b-a)^{n+1}}{(n+1)!} f^{(n+1)}(c). \tag{15}$$

Si $a := 0$ et $b := x$, la *formule de Taylor-Lagrange* (15) est appelée *formule de Mac-Laurin*. Si l'on note $b := a + h$, on peut écrire $c = a + \theta h$, avec $\theta \in]0, 1[$.

Démonstration. Le cas $n = 0$ est le théorème des accroissements finis. Supposons $n \geqslant 1$. On définit une fonction $g : [a, b] \to \mathbb{R}$ par :

$$g(t) := f(b) - f(t) - \sum_{k=1}^{n} \frac{(b-t)^k}{k!} f^{(k)}(t) - \gamma \frac{(b-t)^{n+1}}{(n+1)!},$$

en choisissant $\gamma \in \mathbb{R}$ de telle sorte que $g(a) = 0$ (puisque $b - a \neq 0$, cette condition détermine γ de façon unique).

La fonction g est continue sur $[a, b]$, dérivable sur $]a, b[$ et l'on a $g(a) = g(b) = 0$. On peut donc lui appliquer le théorème de Rolle et l'on obtient $c \in]a, b[$ tel que $g'(c) = 0$. L'égalité $g'(c) = 0$ s'écrit :

$$-f'(c) - \sum_{k=1}^{n} \frac{(b-c)^k}{k!} f^{(k+1)}(c) + \sum_{k=1}^{n} \frac{(b-c)^{k-1}}{(k-1)!} f^{(k)}(c) + \gamma \frac{(b-c)^n}{n!} = 0.$$

Après simplification et en utilisant $b - c \neq 0$, on obtient $\gamma = f^{(n+1)}(c)$. ∎

Soient f une fonction $(n+1)$ fois dérivable sur un intervalle I et $a \in I$. Pour tout $x \in I$, en appliquant le théorème ci-dessus à la restriction de f à $[a, x]$ (ou $[x, a]$), on montre qu'il existe $c_x \in I$ tel que :

$$f(x) = f(a) + \sum_{k=1}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a) + \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(c_x) \tag{16}$$

Le second membre de (16) se décompose en deux parties ; la première est :

$$\begin{aligned} P_{a,n}(x) &:= f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{n!}f^n(a) \\ &= \sum_{k=0}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a), \end{aligned}$$

qui est un *polynôme* en x, de degré au plus n, pour chaque valeur donnée de a. On l'appelle le *polynôme de Taylor*, ou le *polynôme d'approximation*, ou la *partie régulière* de f au point a à l'ordre n. La seconde partie est appelée *le reste de Taylor* : elle mesure la différence entre la fonction f et le polynôme d'approximation. les différentes formules de Taylor diffèrent par l'écriture de ce reste. Ici il est écrit sous forme dite de Lagrange.

Exercice 12.

Montrer le théorème 52 en utilisant la formule de la moyenne de Cauchy (proposition 50 de la page 651).

Solution. On note, pour tout $x \in [0, b-a]$:

$$R_n(x) := f(a+x) - f(a) - \sum_{k=1}^{n} f^{(k)}(a)\frac{x^k}{k!} \qquad \text{et} \qquad S_n(x) := \frac{x^{n+1}}{(n+1)!}\cdot$$

On a :

$$R_n(0) = R_n'(0) = \cdots = R_n^{(n)}(0) = 0 \quad \text{et} \quad S_n(0) = S_n'(0) = \cdots = S_n^{(n)}(0) = 0.$$

En appliquant $(n+1)$ fois la formule des accroissements finis généralisés sur le segment $[0, b-a]$, on obtient des réels $d_1, \ldots, d_{n+1}$ tels que $0 < d_{n+1} < \cdots < d_1$ et :

$$\begin{aligned} \frac{R_n(b-a)}{S_n(b-a)} = \frac{R_n(b-a) - R_n(0)}{S_n(b-a) - S_n(0)} = \frac{R_n'(d_1)}{S_n'(d_1)} = \frac{R_n'(d_1) - R_n'(0)}{S_n'(d_1) - S_n'(0)} = \frac{R_n''(d_2)}{S_n''(d_2)} = \cdots \\ = \frac{R_n^{(n+1)}(d_{n+1})}{S_n^{(n+1)}(d_{n+1})}\cdot \end{aligned}$$

On a $S_n^{(n+1)}(x) = 1$ et $R_n^{(n+1)}(x) = f^{n+1}(a+x)$, donc $R_n(b-a) = f^{n+1}(a+d_{n+1})\frac{(b-a)^{n+1}}{(n+1)!}\cdot$

On a, en posant $c := a + d_{n+1}$:

$$R_n(b-a) = f(b) - f(a) - \sum_{k=1}^{n} \frac{(b-a)^k}{k!} f^{(k)}(a) = f^{n+1}(c)\frac{(b-a)^{n+1}}{(n+1)!}\cdot$$

Exercice 13.

Soient I un intervalle ouvert de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction deux fois dérivable sur I, à dérivée seconde positive. Montrer que le graphe de f est « au-dessus de ses tangentes ».

Solution. Soit $a \in I$. Une équation de la tangente au graphe au point $\big(a, f(a)\big)$ est :

$$g(x) - f(a) - f'(a)(x-a) = 0.$$

Pour tout $x \in I$, on a $f(x) - g(x) = f(x) - f(a) - f'(a)(x-a)$.

D'après le théorème de Taylor-Lagrange, il existe c_x, compris entre a et x, tel que :

$$f(x) = f(a) + f'(a)(x-a) + f''(c_x)\frac{(x-a)^2}{2}.$$

Donc : $f(x) - g(x) = f''(c_x)\dfrac{(x-a)^2}{2} \geqslant 0$.

Inégalité de Taylor-Lagrange

Le résultat suivant généralise l'inégalité des accroissements finis.

Théorème 53 (Inégalité de Taylor-Lagrange).
Soient $n \in \mathbb{N}$ et f une fonction $n+1$ fois continûment dérivable sur un intervalle I, dont la $(n+1)^{ème}$ dérivée est bornée par $M \geqslant 0$. Alors pour tout point $a \in I$, on a :

$$\forall x \in I\,,\ \Big|f(x) - \sum_{k=0}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a)\Big| \leqslant M\,\frac{|x-a|^{n+1}}{(n+1)!}\cdot \tag{17}$$

Si une majoration du module de la $n^{ème}$ dérivée de f est connue, on a une majoration de l'erreur commise en remplaçant f par son polynôme d'approximation.

Démonstration. L'inégalité de Taylor-Lagrange est une conséquence immédiate du théorème de Taylor-Lagrange. D'après (15) on a :

$$\left|f(x) - f(a) - \sum_{k=1}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a)\right| = \big|f^{(n+1)}(c)\big|\frac{|x-a|^{n+1}}{(n+1)!} \leqslant M\,\frac{|x-a|^{n+1}}{(n+1)!}\cdot \qquad \blacksquare$$

L'inégalité (17) est très utile pour obtenir ce que l'on appelle le développement en série d'une fonction en un point (quand ce développement existe...). Voici un exemple[6].

Exemple. On suppose que l'on ne dispose que de la définition de l'exponentielle « de terminale » : c'est l'unique fonction f dérivable sur $\mathbb{R}$ telle que $f' = f$ et $f(0) = 1$. On va montrer que $\exp x$ est la somme de la série entière convergente $\sum\limits_{k\geqslant 0} \frac{x^k}{k!}$ sur $\mathbb{R}$ (*cf.* le théorème et définition 21 de la page 272).
On montre par récurrence que $\exp$ est continûment dérivable à tout ordre et que, pour tout $n \in \mathbb{N}$, $\exp^{(n)} = \exp$ et $\exp^{(n)}(0) = 1$. En utilisant la définition « de terminale », on montre que l'exponentielle est strictement positive et qu'elle est strictement croissante (*cf.* l'exercice 11 de la page 650).
Soit $\alpha > 0$ fixé. Si $|x| \leqslant \alpha$, $0 < \exp x \leqslant \exp \alpha$ et l'inégalité de Taylor-Lagrange permet de majorer le reste par $\exp \alpha\, \frac{\alpha^{n+1}}{(n+1)!}$. La suite $\left(\exp \alpha\, \frac{\alpha^{n+1}}{(n+1)!}\right)_{n\in\mathbb{N}}$ tend vers 0, donc, pour tout $x \in \mathbb{R}$, $|x| \leqslant \alpha$, l'on a :

$$\exp x = \sum_{k\geqslant 0} \frac{x^k}{k!}\cdot$$

Puisque α peut être choisi arbitrairement, le résultat est vrai pour tout $x \in \mathbb{R}$.

À titre d'exercice le lecteur pourra retrouver de façon similaire les développements en série des fonctions $\sin$ et $\cos$ en utilisant seulement :

$$\sin' = \cos, \quad \cos' = -\sin, \qquad \text{et} \quad \forall x \in \mathbb{R}, \quad |\sin x| \leqslant 1, \quad |\cos x| \leqslant 1.$$

[6]Nous reviendrons sur cette question de « développement en série entière » dans le livre L2.

Exercice 14 (Inégalité de Kolmogorov).
Soient $a \in \mathbb{R}$ et $f :]a, +\infty[\to \mathbb{R}$ une fonction deux fois dérivable. On suppose que $|f|$ et $|f''|$ sont bornées respectivement par M_0 et M_2 sur $]a, +\infty[$.
Montrer que $|f'|$ est bornée par $2\sqrt{M_0 M_2}$.
Solution. Soient $x > a$ et $h > 0$.
D'après l'inégalité de Taylor-Lagrange, on a :

$$\left|f(x+h) - f(x) - h\, f'(x)\right| \leqslant M_2 \frac{h^2}{2}$$

et donc :

$$h\left|f'(x)\right| \leqslant \left|h\, f'(x) - \big(f(x+h) - f(x)\big)\right| + |f(x+h) - f(x)| \leqslant M\, \frac{h^2}{2} + 2M_0$$

ce qui donne :

$$\left|f'(x)\right| \leqslant \frac{2M_0}{h} + \frac{hM_2}{2} \quad \text{puis} \quad M_1 \leqslant \frac{2M_0}{h} + \frac{hM_2}{2}.$$

Si $M_2 = 0$, on montre, en faisant tendre h vers $+\infty$, que $f'(x) = 0$.
Supposons $M_2 \neq 0$. En étudiant les variations de la fonction $\varphi : t \in \mathbb{R}_+^* \mapsto \frac{2M_0}{t} + \frac{tM_2}{2}$, on vérifie que le membre de droite est minimal, et égal à $2\sqrt{M_0 M_2}$, pour $t := 2\sqrt{\frac{M_0}{M_2}}$, donc $M_1 \leqslant 2\sqrt{M_0 M_2}$.

2.6.2 Formules de Taylor-Young

La formule de Taylor-Young est une formule de Taylor de nature *locale*. Elle est basée sur des informations sur le comportement d'une fonction au voisinage d'un point.
Le théorème de Taylor-Lagrange permet de prouver très facilement une première version du théorème de Taylor-Young.

Théorème 54.
Soient $n \in \mathbb{N}$ et f une fonction n fois continûment dérivable sur un intervalle I. Alors, pour tout point $a \in I$, on a :

$$f(x) = \sum_{k=0}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a) + (x-a)^n \varepsilon_n(x) \quad \text{où} \quad \lim_{x \to a} \varepsilon_n(x) = 0. \tag{18}$$

Démonstration. Soit $x \in I$, on écrit la formule de Taylor-Lagrange à l'ordre $(n-1)$ sur $[a, x]$, si $a \leqslant x$ (ou $[x, a]$, si $x < a$). Il existe $c_x \in [a, x]$ (ou $c_x \in [x, a]$) tel que :

$$\begin{aligned} f(x) &= \sum_{k=0}^{n-1} \frac{(x-a)^k}{k!} f^{(k)}(a) + \frac{(x-a)^n}{n!} f^{(n)}(c_x) \\ &= \sum_{k=0}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a) + \frac{(x-a)^n}{n!} \left(f^{(n)}(c_x) - f^{(n)}(a)\right). \end{aligned}$$

On pose, pour $x \neq a$:

$$\varepsilon_n(x) := \frac{1}{(x-a)^n}\left(f(x) - \sum_{k=0}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a)\right).$$

On a $\varepsilon_n(x) = \frac{1}{n!}\left(f^{(n)}(c_x) - f^{(n)}(a)\right)$, donc, par continuité de $f^{(n)}$ au point a, on en déduit $\lim\limits_{x\to a} \varepsilon_n(x) = 0$ puisque $|c_x - a| \leqslant |x-a|$. ∎

On a une forme plus forte du théorème de Taylor-Young. La preuve est un peu délicate et peut être omise en première lecture.

Théorème 55 (Théorème de Taylor-Young). Soient $I \subset \mathbb{R}$, $a \in \mathbb{R}$, $n \in \mathbb{N}$ et $f : I \to \mathbb{R}$ une fonction admettant une dérivée d'ordre n au point a. Alors :

$$f(x) = \sum_{k=0}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a) + (x-a)^n \varepsilon_n(x) \quad \text{où} \quad \lim_{x\to a} \varepsilon_n(x) = 0. \tag{19}$$

Rappelons que si f admet une dérivée d'ordre n au point a, il existe un intervalle $I' \subset I$ contenant a sur lequel f est définie et admet des dérivées jusqu'à l'ordre $(n-1)$, la fonction $f^{(n-1)}$ étant dérivable au point a. On vérifie qu'il suffit de prouver le résultat pour $I = I'$.

Démonstration. On démontre le résultat par récurrence sur $p \leqslant n$. Pour tout $p \in [\![1, n]\!]$ et toute fonction g sur I, p fois dérivable au point a, on définit, pour tout $x \in I \setminus \{a\}$:

$$\varepsilon_p(g)(x) := \frac{1}{(x-a)^p}\left(g(x) - \sum_{k=0}^{p} \frac{(x-a)^k}{k!} g^{(k)}(a)\right),$$

Notons (H_p) l'assertion :

« pour toute fonction g sur I, p fois dérivable au point a, on a $\lim\limits_{x\to a} \varepsilon_p(g)(x) = 0$. »

L'assertion (H_1) est vraie d'après la définition de la dérivabilité au point a. Montrons (H_p) par récurrence.

Supposons (H_{p-1}) vraie (pour $p \in [\![2, n-1]\!]$) et montrons (H_p). Soit $f : I \to \mathbb{R}$, p fois dérivable au point a. Sa dérivée f' est $(p-1)$ fois dérivable au point a, elle vérifie donc H_{p-1}. Soit $\varepsilon > 0$, il existe $\eta > 0$ tel que, pour tout $t \in J := I \cap]a-\eta, a+\eta[$, l'on ait $\varepsilon_{p-1}(f') \leqslant \varepsilon$, c'est-à-dire :

$$\left|f'(t) - f'(a) - \sum_{k=1}^{p-1} \frac{(t-a)^k}{k!} f^{(k+1)}(a)\right| \leqslant \varepsilon |t-a|^{p-1}. \tag{20}$$

On définit, sur J, la fonction $R : \begin{array}{ccc} J & \longrightarrow & \mathbb{R} \\ t & \longmapsto & f(t) - f(a) - \sum\limits_{k=1}^{p} \frac{(t-a)^k}{k!} f^{(k)}(a). \end{array}$

D'après (20), on a $\left|R'(t)\right| \leqslant \varepsilon |t-a|^{p-1}$, donc, pour tout $x \in J$, d'après l'inégalité des accroissements finis (sur $[a, x]$ ou $[x, a]$) : $\left|R(x) - R(a)\right| \leqslant \varepsilon |x-a|^p$ et, puisque $R(a) = 0$, cela donne $\left|R(x)\right| \leqslant \varepsilon |x-a|^p$ c'est-à-dire :

$$\left|f(x) - f(a) - \sum_{k=1}^{p} \frac{(x-a)^k}{k!} f^{(k)}(a)\right| \leqslant \varepsilon |x-a|^p.$$

On en déduit $\varepsilon_p(f)(x) \leqslant \varepsilon$. On a ainsi prouvé l'assertion (H_p). ∎

2.7 Dérivée de la réciproque

Théorème 56. Soit f une fonction dérivable sur un intervalle I, et telle que $f' > 0$ sur I (ou bien $f' < 0$ sur I). Alors f est bijective de I sur $J := f(I)$ et sa réciproque f^{-1} est dérivable sur J, et vérifie :

$$\forall y \in J\,,\ \left(f^{-1}\right)'(y) = \frac{1}{f'\left(f^{-1}(y)\right)}\cdot$$

Démonstration. La fonction f est continue (puisque dérivable) et strictement monotone d'après la proposition 47 de la page 650. Le théorème 29 de la page 637 nous dit alors que f est une bijection de I sur J et que f^{-1} est continue sur J.

Soit $y \in J$. Pour tout $z \in J$ tel que $z \neq y$, on a $f^{-1}(z) \neq f^{-1}(y)$ et :

$$\frac{f^{-1}(z) - f^{-1}(y)}{z - y} = \frac{1}{\frac{f\left(f^{-1}(z)\right) - f\left(f^{-1}(y)\right)}{f^{-1}(z) - f^{-1}(y)}}\cdot$$

La continuité de f^{-1} nous donne $\lim\limits_{z \to y} f^{-1}(z) = f^{-1}(y)$, ce qui permet d'obtenir, par passage à la limite :

$$\lim_{z \to y} \frac{f^{-1}(z) - f^{-1}(y)}{z - y} = \frac{1}{f'\left(f^{-1}(y)\right)}\cdot$$

On peut retrouver autrement la formule de dérivée de la réciproque lorsque l'on sait déjà que f^{-1} est dérivable. On a $f \circ f^{-1}(y) = y$ par définition ; en dérivant par rapport à y à l'aide du théorème 36 de la page 643, on trouve $\left(f^{-1}\right)'(y) \cdot \left(f' \circ f^{-1}(y)\right) = 1$, soit le résultat recherché.

Des exemples de fonctions réciproques sont donnés au module suivant.

2.8 Fonctions convexes

Nous allons introduire la notion de *fonction réelle convexe* et établir les principales propriétés d'une telle fonction. Nous reviendrons plus en détail sur les notions de convexité et les applications dans le livre L2, en relation avec la notion *d'ensemble convexe*.

La notion de fonction convexe permet de prouver des inégalités importantes pour l'analyse. Nous en donnons ci-dessous des exemples (*cf.* aussi les inégalités de Hölder et de Minkowski qui font l'objet des exercices IV.2.28 et IV.2.29). On peut étendre la notion de convexité aux fonctions de plusieurs variables. C'est un outil important pour résoudre des *problèmes d'optimisation*.

Dans toute cette partie, I désigne un intervalle réel contenant au moins deux points.

2.8.1 Généralités

Remarque. Soient $x, y \in \mathbb{R}$, $x < y$. On vérifie facilement que l'application de $\mathbb{R}$ dans $\mathbb{R}$, définie par $u \mapsto \lambda := \frac{y-u}{y-x}$ est bijective, d'inverse $\lambda \mapsto u := \lambda x + (1-\lambda)y$, et qu'elle induit une *bijection* du segment $[x, y]$ sur le segment $[0, 1]$. On en déduit la description suivante des segments de $\mathbb{R}$:

$$[x, y] = \{\lambda x + (1-\lambda)y \mid \lambda \in [0, 1]\}.$$

La remarque ci-dessus justifie la définition suivante.

Définition 18. Soient $n \in \mathbb{N}^*$ et $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^n$.
Le *segment* reliant $\mathbf{a}$ à $\mathbf{b}$ noté $[\mathbf{a}, \mathbf{b}]$ est :

$$[\mathbf{a}, \mathbf{b}] = \{\lambda \mathbf{a} + (1-\lambda)\mathbf{b} \mid \lambda \in [0,1]\}.$$

Définition 19. Soit $f : I \to \mathbb{R}$. La fonction f est dite *convexe* sur I si :

$$\forall x,\ y \in I,\ \ \forall \lambda \in [0,1],\ \ f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y). \qquad (21)$$

Si l'inégalité (21) est stricte quand $x \neq y$ et $\lambda \neq 0, 1$, la fonction f est dite *strictement convexe*.
La fonction f est dite *concave* (resp. *strictement concave*) sur I si la fonction $-f$ est convexe (resp. *strictement convexe*).

La définition ci-dessus s'interprète graphiquement de la façon suivante.

Notons, pour $x,\ y \in I$, $A := (x, f(x)) \in \mathbb{R}^2$ et $B := (y, f(y)) \in \mathbb{R}^2$. Les points A et B appartiennent à la courbe représentative (ou graphe) de f et l'on dit que le segment AB est une *corde* de cette courbe. Pour tout $\lambda \in [0,1]$, le point

$$M := \big(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)\big)$$

appartient à la corde AB et le point

$$P := \big(\lambda x + (1-\lambda)y, f(\lambda x + (1-\lambda)y)\big)$$

appartient à la courbe représentative de f.
L'inégalité (21) signifie que la point M est « au-dessus » de P. Ainsi, la fonction f est convexe (resp. concave) si, et seulement si, sa courbe représentative est « au-dessous de chacune de ses cordes » (resp. « au-dessus de chacune de ses cordes »).

Exemples.

1. Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction affine. Les fonctions f et $-f$ sont convexes sur $\mathbb{R}$.
2. La fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $x \in \mathbb{R} \mapsto f(x) := |x|$ est convexe. En effet, $\forall x,\ y \in \mathbb{R}$, $\forall \lambda \in [0,1]$, on a :
   $\big|\lambda x + (1-\lambda)y\big| \leqslant \big|\lambda f(x)\big| + \big|(1-\lambda) f(y)\big| = |\lambda x| + \big|(1-\lambda)y\big| = \lambda|x| + (1-\lambda)|y|.$
3. Montrons que la fonction $f : [0,1] \to \mathbb{R}$ définie par $f(x) := 0$ si $0 < x < 1$ et $f(0) = f(1) := 1$ est convexe.
   Si $\lambda \in]0,1[$, $\lambda x + (1-\lambda)y \in]0,1[$ et $f(\lambda x + (1-\lambda)y) = 0$. L'inégalité (21) est vérifiée, puisque $f \geqslant 0$.
   Si $\lambda = 0$, $f(\lambda x + (1-\lambda)y) = f(y)$ et $\lambda f(x) + (1-\lambda) f(y) = f(y)$. Ainsi l'inégalité (21) est vérifiée. Il en est de même si $\lambda = 1$.
   On aurait pu choisir pour $f(0)$ et $f(1)$ d'autres valeurs pourvu qu'elles soient *positives*.

Remarque. Soit $f : I \to \mathbb{R}$ une fonction *continue* sur I. Si elle est convexe sur l'intérieur J de I, elle est convexe sur I. En effet, on peut passer à la limite en x ou y, à λ fixé, dans l'inégalité (21).

Supposons par exemple $I :=]a, b]$, alors $J =]a, b[$. Dans l'inégalité (21), fixons $x \in J$ et faisons tendre $y \in J$ vers b. L'inégalité est conservée à la limite car f est continue à gauche en b. Les autres cas se traitent de façon analogue.

Exercice 15.

On suppose que I est ouvert. Soit $f : I \to \mathbb{R}$ une fonction convexe et concave sur I. Montrer que f est la restriction à I d'une fonction affine.

Solution. En utilisant (21) pour f et $-f$, on obtient :

$$\forall x,\ y \in I,\ \ \forall \lambda \in [0,1],\ \ f(\lambda x + (1-\lambda)y) = \lambda f(x) + (1-\lambda) f(y).$$

On suppose $a,\ b \in I$, $a < b$. Pour tout $\lambda \in [0,1]$, $(\lambda a + (1-\lambda)b, \lambda f(a) + (1-\lambda) f(b))$ est le barycentre des points $A := (a, f(b)) \in \mathbb{R}^2$ et $B := (b, f(b))$ munis respectivement des coefficients λ et $1-\lambda$. Par suite le graphe de la restriction de f au segment $[a, b]$ est le segment AB. Le résultat s'en déduit.

On montre facilement le résultat suivant.

Proposition 57. Soient $f,\ g : I \to \mathbb{R}$ deux fonctions convexes sur I et $\mu \in \mathbb{R}_+$. Alors les fonctions $f + g$ et μf sont convexes sur I.

Nous allons utiliser la notion de *barycentre* (voir le module III.1 page 428) ainsi que la propriété d'associativité du barycentre (*cf.* le théorème 3 de la page 429).

Lemme 58. Soient $n \in \mathbb{N}^*$, $x_1, \ldots, x_n \in I$ et $\lambda_1, \ldots, \lambda_n \in [0,1]$ des réels positifs non tous nuls. Alors le barycentre $x := \frac{\sum \lambda_i x_i}{\sum \lambda_i}$ de la famille $\{(x_1, \lambda_1) \ldots, (x_n, \lambda_n)\}$ appartient à I.

Démonstration. Il suffit de prouver le résultat en supposant que, pour tout $i \in [\![1, n]\!]$, $\lambda_i \neq 0$. On fait une récurrence sur n. Le résultat est vrai pour $n = 2$ en utilisant la convexité de I (*cf.* la proposition 15 de la page 521). Supposons $n > 2$ et le résultat vrai pour $n - 1$. Notons y le barycentre de la famille $\{(x_2, \lambda_2) \ldots, (x_n, \lambda_n)\}$ $(\lambda_2 + \cdots + \lambda_n \neq 0)$. D'après l'hypothèse de récurrence $y \in I$. D'après la propriété d'associativité du barycentre (*cf.* le théorème 3 de la page 429), x est le barycentre de la famille $\{(x_1, \lambda_1), (\lambda_2 + \cdots + \lambda_n, y)\}$ et, d'après le résultat pour $n = 2$, $x \in I$. ■

Proposition 59 (Inégalité de convexité ou de Jensen). Soit $f : I \to \mathbb{R}$ une fonction convexe. Soient $x_1, \ldots, x_n \in I$ et $\lambda_1, \ldots, \lambda_n \in [0,1]$ des réels positifs de somme 1. On a :

$$f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leqslant \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n).$$

Le point $x := \lambda_1 x_1 + \cdots + \lambda_n x_n$ est barycentre à coefficients positifs non tous nuls de points de I donc, d'après le lemme ci-dessus, $x \in I$.

Démonstration. On note H_n la propriété à démontrer et l'on procède par récurrence. Par hypothèse H_2 est vraie. On suppose que H_{n-1} est vraie ($n > 2$). Soient $x_1, \dots, x_n \in I$ et $\lambda_1, \dots, \lambda_n \in [0,1]$ des réels de somme 1. On distingue deux cas.

1. On suppose $\lambda_n = 1$. Alors :
$$\lambda_1 = \cdots = \lambda_{n-1} = 0 \quad \text{et} \quad f(\lambda_1 x_1 + \cdots + \lambda_n x_n) = f(x_n) = \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n).$$
2. On suppose $\lambda_n \neq 1$.
On note, pour tout $i \in [\![1, n-1]\!]$, $\mu_i := \frac{\lambda_i}{1-\lambda_n}$ et $y := \mu_1 x_1 + \cdots + \mu_{n-1} x_{n-1}$. On a $\mu_i \in [0,1]$ et $\mu_1 + \cdots + \mu_{n-1} = 1$. D'après H_{n-1}, on a :
$$f(y) = f(\mu_1 x_1 + \cdots + \mu_{n-1} x_{n-1}) \leqslant \mu_1 f(x_1) + \cdots + \mu_{n-1} f(x_{n-1}).$$
d'après la convexité de f, on a :
$$\begin{aligned} f(\lambda_1 x_1 + \cdots + \lambda_n x_n) &= f\left(\lambda_n x_n + (1-\lambda_n) y\right) \leqslant \lambda_n f(x_n) + (1-\lambda_n) f(y) \\ &\leqslant \lambda_n f(x_n) + (1-\lambda_n)\left(\mu_1 f(x_1) + \cdots + \mu_{n-1} f(x_{n-1})\right) \\ &= \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n), \end{aligned}$$
ce qui prouve H_n. ∎

La proposition suivante permet de caractériser la convexité d'une fonction $f : I \to \mathbb{R}$ en termes de pentes de cordes associées au graphe de f. Elle est la clé de la preuve des principaux résultats sur les fonctions convexes.

Proposition 60. Soit $f : I \to \mathbb{R}$. Les conditions suivantes sont équivalentes :

(i) f est convexe ;

(ii) pour tous $x,\ y,\ z \in I$ tels que $x < y < z$, on a :
$$\frac{f(y) - f(x)}{y - x} \leqslant \frac{f(z) - f(x)}{z - x}. \tag{22}$$

Démonstration. Soient $x,\ y,\ z \in I$ tels que $x < y < z$.
Il existe un (unique) réel $\lambda \in]0,1[$ tel que $y = \lambda x + (1-\lambda) z$, et l'on a $\lambda = \frac{z-y}{z-x}$.
Montrons (i) $\Rightarrow$ (ii). Supposons f convexe. Alors :
$$f(y) \leqslant \lambda f(x) + (1-\lambda) f(z) = \frac{z-y}{z-x} f(x) + \left(1 - \frac{z-y}{z-x}\right) f(z)$$
et
$$f(y) - f(x) \leqslant \left(1 - \frac{z-y}{z-x}\right)\left(f(z) - f(x)\right) = (y-x)\frac{f(z) - f(x)}{z-x}.$$
d'où (ii), en utilisant $y - x \geqslant 0$.
Montrons (ii) $\Rightarrow$ (i). On a $\frac{f(y)-f(x)}{y-x} \leqslant \frac{f(z)-f(x)}{z-x}$, soit :
$$\frac{f\left(\lambda x + (1-\lambda) z\right) - f(x)}{(1-\lambda)(z-x)} \leqslant \frac{f(z) - f(x)}{z - x}$$
et $f\left(\lambda x + (1-\lambda) z\right) \leqslant \lambda f(x) + (1-\lambda) f(z)$, en utilisant $1 - \lambda \geqslant 0$ et $z - x \geqslant 0$. ∎

Interprétation géométrique.

Notons $A := (x, f(x))$, $B := (y, f(y))$, $C := (z, f(z))$ ($A,\ B,\ C \in \mathbb{R}^2$). L'inégalité (22) signifie que la pente de la corde AC majore la pente de la corde AB.

Le lemme suivant donne des versions équivalentes de l'inégalité (22), donc d'autres caractérisations de la convexité.

Lemme 61. Soit $f : I \to \mathbb{R}$ une fonction convexe.
Pour tous $x,\ y,\ z \in I$ tels que $x < y < z$, les conditions suivantes sont équivalentes :

(i) $(z-x)f(y) + (y-z)f(x) + (x-y)f(z) \leqslant 0$;

(ii) $\dfrac{f(y)-f(x)}{y-x} \leqslant \dfrac{f(z)-f(x)}{z-x}$;

(iii) $\dfrac{f(y)-f(x)}{y-x} \leqslant \dfrac{f(z)-f(y)}{z-y}$;

(iv) $\dfrac{f(z)-f(x)}{z-x} \leqslant \dfrac{f(z)-f(y)}{z-y}$.

Démonstration. On a $y - x > 0$ et $z - x > 0$. La condition (ii) est donc équivalente à :

$$(f(y) - f(x))\,(z-x) \leqslant (f(z) - f(x))\,(y-x).$$

Un calcul simple montre que cette dernière inégalité est équivalente à (i). On prouverait de façon analogue les équivalences (iii) $\Leftrightarrow$ (i) et (iv) $\Leftrightarrow$ (i). ∎

En utilisant le lemme ci-dessus et la proposition 60 on obtient le résultat suivant[7].

Corollaire 62.
Soit $f : I \to \mathbb{R}$ une fonction convexe. Pour tous $x,\ y,\ z \in I$ tels que $x < y < z$, on a :

$$\frac{f(y)-f(x)}{y-x} \leqslant \frac{f(z)-f(x)}{z-x} \leqslant \frac{f(z)-f(y)}{z-y}. \tag{23}$$

Interprétation géométrique

La figure ci-dessus (dont on reprend les notations) permet de mémoriser les inégalités (23) qui signifient que la pente de AC majore la pente de AB et que la pente de BC majore la pente de AC.

[7] Appelé par certains auteurs le lemme des trois cordes

Exercice 16.

Montrer que si une fonction f convexe sur $\mathbb{R}$ n'est pas constante, on a $\lim\limits_{x\to+\infty} f = +\infty$ *ou* $\lim\limits_{x\to-\infty} f = +\infty$. En particulier f n'est pas majorée.

Solution. Puisque f n'est pas constante, il existe $x,\ y \in \mathbb{R}$ tels que $f(x) \neq f(y)$. On peut supposer $x < y$. On distingue deux cas.

1. Supposons $f(x) < f(y)$. Notons $m := \frac{f(y)-f(x)}{y-x}$. On a $m > 0$ et, d'après (60), pour tout $z > x$, $f(z) \geqslant f(y) + m(z-y) = mz + f(y) - my$.
   On a $\lim\limits_{z\to+\infty}(mz + f(y) - my) = +\infty$, donc, *a fortiori*, $\lim\limits_{z\to+\infty} f(z) = +\infty$.
2. Supposons $f(x) > f(y)$. La fonction $g : x \in \mathbb{R} \mapsto f(-x)$ est convexe sur $\mathbb{R}$. On peut lui appliquer le résultat de 1. On en déduit $\lim\limits_{x\to+\infty} g = +\infty$ et $\lim\limits_{x\to-\infty} f = +\infty$.

Proposition 63. Soit $f : I \to \mathbb{R}$ une fonction. La fonction f est convexe si, et seulement si, pour tout $a \in I$, la fonction $\varphi_a : I \setminus \{a\} \to \mathbb{R}$ définie par $x \mapsto \frac{f(x)-f(a)}{x-a}$ est croissante.

Démonstration. On suppose f convexe. Soient $x, y \in I$ tels que $x < y$. Nous allons montrer l'inégalité :

$$\varphi_a(x) = \frac{f(x)-f(a)}{x-a} \leqslant \frac{f(y)-f(a)}{y-a} = \varphi_a(y). \tag{24}$$

On distingue trois cas.

1. On suppose $a < x < y$. On obtient (24) en utilisant (22) pour le triplet (a, x, y).
2. On suppose $x < a < y$. En utilisant (23) pour le triplet (x, a, y), on obtient :
$$\frac{f(x)-f(a)}{x-a} = \frac{f(a)-f(x)}{a-x} \leqslant \frac{f(y)-f(a)}{y-a}.$$
3. On suppose $x < y < a$. En utilisant (23) pour le triplet (x, y, a), on obtient :
$$\frac{f(x)-f(a)}{x-a} = \frac{f(a)-f(x)}{a-x} \leqslant \frac{f(a)-f(y)}{a-y} = \frac{f(y)-f(a)}{y-a}.$$

Ainsi φ_a est croissante.

Inversement, supposons φ_x croissante pour tout $x \in I$. Soient $x\ , y\ , z \in I$ tels que $x < y < z$. On a $\varphi_x(y) \leqslant \varphi_x(z)$, c'est-à-dire :

$$\frac{f(y)-f(x)}{y-x} \leqslant \frac{f(z)-f(x)}{z-x}.$$

On en déduit la convexité de f en utilisant la proposition 60. ∎

Exercice 17.
Soit $f : I \to \mathbb{R}$ une fonction convexe.
Montrer que, pour tous $x,\ y,\ z \in I$ tels que $x < y < z$, l'on a :

$$\begin{vmatrix} 1 & 1 & 1 \\ x & y & z \\ f(x) & f(y) & f(z) \end{vmatrix} \geqslant 0.$$

Solution. En retranchant la première colonne du déterminant aux deux autres, on obtient :

$$\begin{aligned} \begin{vmatrix} 1 & 1 & 1 \\ x & y & z \\ f(x) & f(y) & f(z) \end{vmatrix} &= \begin{vmatrix} y-x & z-x \\ f(y)-f(x) & f(z)-f(x) \end{vmatrix} \\ &= (y-x)(z-x)\left(\frac{f(z)-f(x)}{z-x} - \frac{f(y)-f(x)}{y-x}\right) \end{aligned}$$

et le dernier terme des égalités ci-dessus est positif d'après la croissance de φ_x.

Proposition 64. Soit $f : I \to \mathbb{R}$ une fonction convexe.

(i) Si f admet en $a \in I$ un minimum *relatif*, elle admet en a un minimum *absolu*.

(ii) Si f est *strictement convexe*, elle admet un minimum en *au plus* un point. Si elle admet un minimum celui-ci est strict.

Démonstration.

(i) On suppose que f admet un minimum relatif en $a \in I$. Il existe $\alpha > 0$ tel que, pour tout $y \in I \cap]a-\alpha, a+\alpha[$, l'on ait $f(a) \leqslant f(y)$. On distingue deux cas (dont l'un ne se produit pas si a est une extrémité de I).

(a) Soit $x \in I$ tel que $x > a$ (s'il en existe).
Il existe $y \in I$ tel que $a < y < x$ et $|y-a| < \alpha$. On a $a < y < x$, donc, d'après (22) : $\frac{f(y)-f(a)}{y-a} \leqslant \frac{f(x)-f(a)}{x-a}$.
Puisque $f(y) - f(a) \geqslant 0$, on a $f(x) - f(a) \geqslant 0$.

(b) Soit $x \in I$ tel que $x < a$ (s'il en existe).
Il existe $y \in I$ tel que $a < y < x$ et $|y-a| < \alpha$. On a $x < y < a$, donc, d'après (23) : $\frac{f(x)-f(a)}{a-x} \geqslant \frac{f(y)-f(a)}{a-y}$.
Puisque $f(y) - f(a) \geqslant 0$, on a $f(x) - f(a) \geqslant 0$.

(ii) Supposons f strictement convexe sur I. Soit a un minimum (absolu) de f sur I. On distingue, comme en (i) les cas (a) et (b).

(a) Soit $x \in I$ tel que $x > a$ (s'il en existe).
Soit $y \in I$ tel que $a < y < x$. On a $y = \lambda x + (1-\lambda)a$, avec $\lambda \in]0,1[$, donc $f(y) < \lambda f(x) + (1-\lambda)f(a)$. Puisque $f(a) \leqslant f(x)$, on a :

$$f(y) < \lambda f(x) + (1-\lambda)f(a) \leqslant \lambda f(x) + (1-\lambda)f(x) = f(x).$$

Ainsi $f(a) \leqslant f(y) < f(x)$.

(b) Soit $x \in I$ tel que $x < a$ (s'il en existe).
On prouve de façon similaire que $f(a) < f(x)$.

Ainsi a est un minimum absolu strict et il est évidemment unique. ■

2.8.2 Convexité et dérivabilité

Théorème 65. On suppose que l'intervalle I est *ouvert*. Si la fonction $f : I \to \mathbb{R}$ est convexe, elle est *continue* et possède en chaque point $x \in I$ une *dérivée à gauche* $f'_g(x)$ et une *dérivée à droite* $f'_d(x)$. De plus, pour tous $x,\ y \in I$, tels que $x < y$, on a :

$$f'_g(x) \leqslant f'_d(x) \leqslant \frac{f(y) - f(x)}{y - x} \leqslant f'_g(y) \leqslant f'_d(y). \tag{25}$$

Démonstration.
Soient $u,\ v,\ x \in I$ tels que $u < x < v$. La fonction $\varphi_x : t \in]u, v[\setminus \{x\} \mapsto \frac{f(t)-f(x)}{t-x}$ est croissante sur les intervalles $]u, x[$ et $]x, v[$. Elle est minorée par $\varphi_x(u)$ et majorée par $\varphi_x(v)$, elle admet donc une limite quand t tend vers x par valeurs inférieures et une limite quand t tend vers x par valeurs supérieures. D'après la définition des dérivées à droite et à gauche, f est dérivable à droite et à gauche en x et l'on a :

$$\lim_{t \overset{>}{\to} x} \varphi_x(t) = f'_d(x) \quad \text{et} \quad \lim_{t \overset{<}{\to} x} \varphi_x(t) = f'_g(x).$$

De plus $f'_g(x) \leqslant f'_d(x)$, $\varphi_x(t) \leqslant f'_g(x)$ si $t \in I$, $t < x$ et $f'_d(x) \leqslant \varphi_x(t)$ si $t \in I$, $x < t$.
Soit y tel que $u < x < y < v$. On a $\varphi_x(y) = \varphi_y(x) = \frac{f(y)-f(x)}{y-x}$, donc :

$$f'_d(x) \leqslant \varphi_x(y) = \varphi_y(x) \leqslant f'_g(y) \quad \text{et} \quad f'_d(x) \leqslant \frac{f(y) - f(x)}{y - x} \leqslant f'_g(y).$$ ■

Nous invitons le lecteur à illustrer les inégalités (25) par un dessin.

Interprétation géométrique

Le graphe d'une fonction convexe est « au-dessus » de chacune de ses tangentes.

• Une fonction convexe n'est pas nécessairement dérivable. Par exemple la fonction $x \in \mathbb{R} \mapsto |x|$ n'est pas dérivable en $x = 0$.
• Une fonction convexe sur un segment n'est pas nécessairement continue aux extrémités de ce segment. Par exemple la fonction $f : [0, 1] \to \mathbb{R}$ définie par $f(x) := 0$ si $0 < x < 1$ et $f(0) = f(1) := 1$ est convexe sur $[0, 1]$, mais elle n'est pas continue en $x = 0$ et $x = 1$.
• Une fonction convexe continue sur un intervalle $[a, b[$ (resp. $]a, b]$) n'est pas nécessairement dérivable à droite (resp. à gauche) en l'extrémité a (resp. b). Par exemple la fonction $f : [0, 1[$ définie par $x \mapsto f(x) := \sqrt{x}$ est continue sur $[0, 1[$ mais n'est pas dérivable à droite en 0 : $\lim\limits_{x \overset{>}{\to} 0} \sqrt{x}/x = \lim\limits_{x \overset{>}{\to} 0} 1/\sqrt{x} = +\infty$.

Le critère *différentiel* suivant est utile pour la recherche des minima d'une fonction convexe.

Corollaire 66. Soient I un intervalle ouvert, $f : I \to \mathbb{R}$ une fonction convexe et $a \in I$. Les deux conditions suivantes sont équivalentes :

(i) a est un minimum de f sur I ;

(ii) $f'_g(a) \leqslant 0 \leqslant f'_d(a)$.

Si f est dérivable, la condition (ii) signifie : $f'(a) = 0$. On retrouve la condition nécessaire classique pour un minimum. Dans le cas convexe elle est aussi suffisante.

Démonstration. Montrons (i)$\Rightarrow$ (ii). Si a est un minimum, on a :

- pour tout $x \in I$, $x > a$, $f(x) - f(a) \geqslant 0$, donc $\frac{f(x)-f(a)}{x-a} \geqslant 0$, et, en passant à la limite, $f'_d(a) \geqslant 0$;
- pour tout $x \in I$, $x < a$, $f(x) - f(a) \geqslant 0$, donc $\frac{f(x)-f(a)}{x-a} \leqslant 0$, et, en passant à la limite, $f'_g(a) \leqslant 0$.

Montrons (ii)$\Rightarrow$ (i). Supposons $f'_g(a) \leqslant 0 \leqslant f'_d(a)$.

- Si $x > a$, $\frac{f(x)-f(a)}{x-a} \geqslant f'_d(a) \geqslant 0$, donc $f(x) \geqslant f(a)$.
- Si $x < a$, $\frac{f(x)-f(a)}{x-a} \leqslant f'_g(a) \leqslant 0$, donc $f(x) \geqslant f(a)$.

ainsi a est un minimum. ∎

Exemple. Soit $f : \mathbb{R} \to \mathbb{R}$ définie par $x \mapsto f(x) := |x|$. Le point 0 est le minimum de f sur $\mathbb{R}$ et $f'_d(0) = 1$, $f'_g(0) = -1$. On a bien $-1 \leqslant 0 \leqslant 1$.

La proposition suivante permet de caractériser les fonctions dérivables convexes.

Proposition 67. On suppose que I est ouvert. Soit $f : I \to \mathbb{R}$.

(i) La fonction f est convexe si, et seulement si, f'_d est croissante.

(ii) La fonction f est convexe si, et seulement si, f'_g est croissante.

(ii) On suppose que f est dérivable sur I. Elle est convexe si, et seulement si, f' est croissante.

Démonstration.

(i) Supposons f convexe.
Soient x, $y \in I$, tels que $x < y$. D'après (25), on a $f'_d(x) \leqslant f'_d(y)$, donc f' est croissante.
Supposons inversement que f'_d est croissante. Soient x, y, $y \in I$ tes que $x < y < z$. D'après le théorème des accroissements finis, il existe $c \in \,]x, y[$ et $d \in \,]y, z[$ tels que :

$$\frac{f(y) - f(x)}{y - x} = f'(c) \quad \text{et} \quad \frac{f(z) - f(y)}{z - y} = f'(d).$$

On a $x < c < y < d < z$ et f'_d est croissante, donc $f'(c) \leqslant f'(d)$ et

$$\frac{f(y) - f(x)}{y - x} \leqslant \frac{f(z) - f(y)}{z - y}.$$

En utilisant le lemme 61 et la proposition 60, on en déduit que f est convexe.

(ii) La preuve est similaire, *mutatis mutandis*, à celle de (i).

(iii) L'assertion résulte de (i) puisque, si f est dérivable, $f'(x) = f'_d(x)$. ■

Corollaire 68. On suppose que I est ouvert. Si f est une fonction convexe et dérivable sur I, on a :

$$\forall a,\ x \in I,\ \ f(a) + (x - a)f'(a) \leqslant f(x).$$

Démonstration. On distingue deux cas.
Supposons $a < x$. D'après la formule des accroissements finis, il existe $c \in \,]a, x[$ tel que $f(x) = f(a) + (x - a)f'(c)$. D'après la croissance de f' : $f'(a) \leqslant f'(c) = \frac{f(x)-f(a)}{x-a}$, donc $f(a) + (x - a)f'(a) \leqslant f(x)$.
Supposons $x < a$. D'après la formule des accroissements finis, il existe $c \in \,]x, a[$ tel que $f(x) = f(a) + (x - a)f'(c)$ et l'on termine la preuve comme ci-dessus. ■

Exercice 18.

On supose I ouvert. Soit $f : I \to \mathbb{R}$ une fonction convexe sur I.

1. Montrer que :
$$\forall a,\ x \in I,\ \ f(a) + (x - a)f'_d(a) \leqslant f(x), \quad \text{et} \quad f(a) + (x - a)f'_g(a) \leqslant f(x).$$
2. Soit $m \in \mathbb{R}$. Montrer que la droite affine d'équation $y = f(a) + m(x - a)$ est « au dessous » de la courbe représentative de f si, et seulement si, $f'_g(a) \leqslant m \leqslant f'_d(a)$.

Solution.

1. On raisonne comme dans la preuve du corollaire 68 en remplaçant la dérivée par la dérivée à droite (resp. à gauche).
2. D'après 1., les demi-droites :
$$T_d := \{a + t\left(1, f'_d(a)\right) \mid t \in \mathbb{R}_+\} \quad \text{et} \quad T_g := \{a - t\left(1, f'_g(a)\right) \mid t \in \mathbb{R}_+\}$$
sont au dessous de la courbe représentative de f. Si $f'_g(a) \leqslant m \leqslant f'_d(a)$, la droite affine d'équation $y = f(a) + m(x - a)$ est au-dessous de T_d et de T_g, donc, *a fortiori*, au-dessous de la courbe représentative de f.
Inversement, si la droite affine D d'équation $y = f(a) + m(x - a)$ est au dessous de la courbe représentative de f, les cordes issues de a sont au dessus de D et, en passant à la limite sur les pentes de ces cordes, on en déduit $f'_g(a) \leqslant m \leqslant f'_d(a)$.

Un point $a \in I$ est minimum pour f si, et seulement si, la droite horizontale $y = f(a)$ est en dessous de la courbe représentative de f.
On retrouve donc le corollaire 66 de la page précédente.

Corollaire 69. Une fonction $f : I \to \mathbb{R}$ deux fois dérivable sur I est convexe si, et seulement si, $f'' \geqslant 0$.

Démonstration.

Si I est ouvert le résultat est une conséquence immédiate de la proposition 67.

Supposons, par exemple, que $I =]a, b]$ et notons $J :=]a, b[$. Par hypothèse f est deux fois dérivable sur I donc f et f' sont continues sur I.

Supposons que f est convexe sur I. Alors, sur J, la fonction f' est croissante. Puisqu'elle est continue sur I, elle est croissante sur I et $f''(b) = (f')'_g(b) \geqslant 0$.

Inversement, supposons $f'' \geqslant 0$ sur I. Alors f est convexe sur J et, puisqu'elle est continue sur I, elle est convexe sur I (*cf.* la remarque de la page 659).

Si $I = [a, b[$, $I = [a, b]$, $I = [a, +\infty[$ ou $I =]-\infty, b]$, on utilise des variantes du raisonnement précédent. ∎

Exemples.

1. On a, pour tout $x \in \mathbb{R}$, $(e^x)'' = e^x$ et $e^x > 0$ pour tout $x \in \mathbb{R}$, donc la fonction exponentielle est convexe sur $\mathbb{R}$.
2. On a, pour tout $x \in \mathbb{R}_+^*$, $(\ln x)'' = -1/x^2$, donc la fonction logarithme est concave sur $\mathbb{R}_+^*$.
3. Soient $a,\ b,\ c \in \mathbb{R}$. On définit $f : \mathbb{R} \to \mathbb{R}$ par $x \in \mathbb{R} \mapsto f(x) := ax^2 + bx + c$. On a $f''(x) = a$, donc f est convexe si $a \geqslant 0$ et concave si $a \leqslant 0$.
4. Soient $p > 1$ et $f : \mathbb{R}_+ \to \mathbb{R}$ définie par $x \mapsto f(x) := x^p$.

   On a $f''(x) = p(p-1)x^{p-2} \geqslant 0$, donc f est convexe sur $\mathbb{R}_+$.

Exercice 19 (Inégalité de Young).

Soient p et q deux réels strictement positifs tels que $\frac{1}{p} + \frac{1}{q} = 1$. Montrer que, si x et y sont des réels positifs, on a :

$$xy \leqslant \frac{x^p}{p} + \frac{y^q}{q}. \tag{26}$$

Solution. Si $x = 0$ ou $y = 0$ l'inégalité est évidente. Supposons $xy \neq 0$. La fonction logarithme étant concave, on a :

$$\ln\left(\frac{x^p}{p} + \frac{y^q}{q}\right) \geqslant \frac{1}{p}\ln x^p + \frac{1}{q}\ln y^q = \ln x + \ln y = \ln xy,$$

d'où, en utilisant la croissance de la fonction exponentielle : $\frac{x^p}{p} + \frac{y^q}{q} \geqslant xy$.

Si $p = 2$, alors $q = 2$ et l'inégalité de Young est l'inégalité de Cauchy : $xy \leqslant \frac{x^2+y^2}{2}$.

---

3 Étude d'une fonction

On étudie une fonction $f : D_f \to \mathbb{R}$ pour mettre en évidence ses propriétés importantes. Il est très souvent utile de tracer le *graphe* de f, c'est-à-dire l'ensemble $\Gamma := \{(x, f(x)) \mid x \in D_f\}$ dans le plan $\mathbb{R}^2$.
Notons que l'étude d'une fonction réciproque peut être faite de deux façons : soit par la méthode usuelle, soit en étudiant la fonction dont elle est réciproque. Les deux graphes se déduisent l'un de l'autre par échange des coordonnées x et y, autrement dit par symétrie par rapport à la « première bissectrice ».
L'étude d'une fonction (souvent « définie » par une ou plusieurs formules) se décompose en un certain nombre d'étapes « standardisées ».

3.1 Définition et variations

En premier lieu, il faut définir précisément la fonction f à partir de la (ou des) formule(s) de départ, donc essentiellement définir D_f. (C'est ce que l'on appelle traditionnellement, mais *improprement*, rechercher *l'ensemble de définition* de la fonction.) Il faut exclure les valeurs qui donnent des divisions par zéro, des racines de nombres strictement négatifs, etc. Si la fonction n'est pas définie en certaines valeurs $a \in \mathbb{R}$, mais est définie au voisinage de a (par exemple sur un intervalle $]a, b[$), alors on recherche une limite éventuelle au point a. Si cette limite existe et est finie, alors la fonction peut être prolongée par continuité au point a, et celui-ci peut être ajouté à l'ensemble de définition.

On examine si la fonction est paire ou impaire. Dans chacun de ces cas, il suffit de l'étudier pour $x \geqslant 0$.

Si la fonction f est τ-périodique, on peut se contenter de faire l'étude sur une période, par exemple sur $[0, \tau]$ ou $[-\tau/2, \tau/2]$. Si la fonction est, de plus, paire ou impaire on choisit $[-\tau/2, \tau/2]$, ce qui pemet de réduire l'intervalle d'étude à $[0, \tau/2]$.

L'étape suivante est l'étude des *variations* de f. Pour cela on examine d'abord si elle est dérivable. On calcule la dérivée, et on cherche son signe. Sur chaque intervalle où ce signe reste constant, la fonction est monotone. On peut alors établir un *tableau de variations*. À ce stade, il est important, pour repèrer une erreur possible, de vérifier que les variations sont compatibles aux valeurs de la fonction, calculées aux points limites, ou aux points d'annulation de la dérivée.

3.2 Branches infinies

Reste l'étude des *branches infinies* du graphe, c'est-à-dire des morceaux de courbes qui « partent vers l'infini ». Dans ce cas, on cherche une *courbe asymptote* simple (en général une droite) pour mieux spécifier l'allure de la branche infinie. Le mot *asymptote* indique que les deux courbes se rapprochent indéfiniment.
Si f a une « valeur interdite », c'est-à-dire un point $a \in \mathbb{R}$ où la limite de f est infinie, alors il y a une (ou deux) branches infinies en ce point. Si $\lim\limits_{x \to a} f(x) = +\infty$ par exemple, le graphe a une droite asymptote verticale (d'équation $x = a$), la courbe partant vers le haut. Si les limites à gauche et à droite sont toutes deux infinies, mais de signes opposés, alors la droite est doublement asymptote, à la fois vers le haut et le bas.

Les autres branches infinies proviennent des cas $x \to \pm\infty$, lorsque la fonction est définie sur un intervalle non borné. Si f admet une limite finie α en $+\infty$ ou en $-\infty$, alors le graphe admet la droite horizontale d'équation $y = \alpha$ pour asymptote. La courbe et la droite peuvent se croiser, éventuellement un nombre infini de fois (considérer par exemple $f(x) = \frac{\sin x}{x}$). Étudier le signe de $f(x) - \alpha$ permet de placer plus précisément la courbe par rapport à la droite.

Le cas le plus difficile arrive lorsque f a une limite infinie lorsque x tend vers $\pm\infty$. En effet la branche infinie peut avoir de nombreux comportements possibles (comparer par exemple les graphes des fonctions exponentielles et logarithmes, donnés au module suivant).

Dans la pratique, on cherche alors une courbe asymptote simple. Par exemple la fonction $x \mapsto x^2 + e^{-x}\sin x$ admet visiblement la parabole d'équation $y = x^2$ comme asymptote, puisque la différence tend vers zéro.

Lorsqu'une telle courbe asymptote ne saute pas si clairement aux yeux, on cherche une droite asymptote d'équation $y = ax + b$. La question est alors : comment trouver a et b ? On les cherche l'un après l'autre, en partant du constat suivant : si la droite $y = ax + b$ est asymptote au graphe de f, alors la limite (quand x tend vers l'infini considéré) de $f(x) - (ax + b)$ doit être nulle. En particulier, la limite de $(f(x)/x) - a$ doit être nulle (puisque x tend vers $\pm\infty$ et donc $\lim b/x = 0$), c'est-à-dire que l'on doit avoir :

$$a = \lim_{x\to\infty} \frac{f(x)}{x} \qquad \text{et} \qquad b = \lim_{x\to\infty} \big(f(x) - ax\big).$$

Parfois il est nécessaire de distinguer la recherche des limites en $+\infty$ et en $-\infty$, ce qu'indique le contexte.

On cherche donc la limite éventuelle de $f(x)/x$. Si celle-ci est infinie, alors f croît vers l'infini plus vite que toute droite, ce qui correspond à une *branche parabolique orientée verticalement* : c'est le cas de la fonction $x \mapsto x^2$ (qui donne une vraie parabole) mais aussi de l'exponentielle (dont le graphe n'est pas une parabole, mais a juste un air de ressemblance). Si cette limite est nulle, alors f croît vers l'infini moins lentement que toute droite, ce qui donne une *branche parabolique orientée horizontalement* : c'est le cas de la fonction $x \mapsto \sqrt{x}$ (dont le graphe est une demi parabole) mais aussi du logarithme.

Si la limite de $f(x)/x$ est un nombre fini $a \neq 0$, alors on cherche $b := \lim\big(f(x) - ax\big)$. Si cette limite est infinie, la courbe s'oriente comme les droites d'équation $y = ax + c$, mais s'éloigne indéfiniment de chacune d'elles : on parle de branche parabolique orientée selon la pente a. C'est le cas de la figure IV.2.1 de la page suivante, où l'on a représenté à gauche la fonction $x \mapsto \frac{x}{2} + \sqrt[3]{x} + 1$ et la droite $y = \frac{x}{2}$: elles ne sont pas asymptotes, mais ont la même « orientation ».

Finalement si $b := \lim\big(f(x) - ax\big)$ est un nombre fini, alors la droite $y = ax + b$ est asymptote à la courbe. On la place alors aussi sur le graphe (en pointillés), de façon à mettre en valeur le rapprochement de la courbe. On peut aussi étudier le signe de la différence $f(x) - (ax + b)$, de façon à savoir de quel côté de la droite la courbe se situe. Ce côté n'est pas forcément constant. La figure IV.2.1 donne un exemple à droite : on a représenté la fonction $x \mapsto \dfrac{x}{2} + 1 + x^{-4/3}\sin(x^2)$, et sa droite asymptote $y = \dfrac{x}{2} + 1$; les deux courbes se croisent un nombre infini de fois, car la différence a le signe de $\sin(x^2)$.

FIGURE IV.2.1 – Branche parabolique (à gauche) et asymptote oblique (à droite).

Exercice type corrigé

IV.2.0 Dans l'exercice suivant, on se propose de retrouver certains résultats *sans utiliser le théorème des accroissements finis* et de les généraliser. On note I un intervalle réel contenant au moins deux points.

1. Montrer que, pour tous $a, b \in I$ vérifiant $a < b$ et toute fonction f continue sur $[a, b]$, dérivable à droite sur $[a, b[$ et vérifiant $f(b) > f(a)$, il existe au moins un réel $\ell \in [a, b[$ vérifiant $f'_d(\ell) \geqslant 0$.
2. Soit $g : I \to \mathbb{R}$ continue sur I, dérivable à droite sur l'intérieur J de I et telle que $g'_d = 0$ sur J. Montrer que g est constante sur I.
3. Soit $g : I \to \mathbb{R}$ continue sur $[a, b]$, dérivable à droite sur $[a, b[$ et telle que $g'_d \leqslant 0$ (resp. $g'_d \geqslant 0$) sur $[a, b[$. Montrer que g est décroissante (resp. croissante) sur I.

Solution.

1. Soit A l'ensemble des $x \in [a, b]$ tels que $f(x) \leqslant f(a)$. Comme $f(b) > f(a)$, par continuité de f en b il existe $\eta > 0$ tel que $\forall x \in [b - \eta, b]\,,\ f(x) > f(a)$. Donc $A \subset [a, b - \eta[$. D'autre part, $a \in A$, donc A est une partie non vide de $\mathbb{R}$ majorée par $b - \eta$ et elle admet par conséquent une borne supérieure $\ell \leqslant b - \eta < b$. On ne peut pas avoir $f(\ell) > f(a)$, sinon, en appliquant ce qui précède à $[a, \ell]$ (au lieu de $[a, b]$), on pourrait trouver un réel $\varepsilon > 0$ tel que $A \subset [a, \ell - \varepsilon[$, ce qui contredirait le fait que ℓ est le plus petit des majorants de A. Donc $f(\ell) \leqslant f(a)$ et $\ell \in A$ (en fait, on a $f(\ell) = f(a)$, mais ce n'est pas nécessaire pour ce qui suit).
   Soit $x \in]\ell, b]$. Par définition de ℓ, on a $f(x) > f(a)$. Donc :
   $$\frac{f(x) - f(\ell)}{x - \ell} \geqslant \frac{f(x) - f(a)}{x - \ell} > 0$$
   et par passage à la limite $f'_d(\ell) \geqslant 0$.
2. Soit $g : I \to \mathbb{R}$ continue sur I, dérivable à droite sur l'intérieur J de I et telle que $g'_d = 0$ sur J. On raisonne par l'absurde. Si g n'est pas constante sur I, elle n'est pas non plus constante sur J (d'après la continuité sur I) et il existe un segment $[a, b] \subset J$ tel que $g(b) \neq g(a)$.

Posons $m := \dfrac{g(b)-g(a)}{b-a} \neq 0$ et :

$$\forall x \in I, \quad f(x) := -x + \frac{2}{m}\, g(x).$$

La fonction f est dérivable à droite sur J, donc sur $[a, b[$, et l'on vérifie :

$$f(b) - f(a) = b - a > 0 \qquad \text{et} \qquad f'_d = -1 \quad \text{sur} \quad J,$$

ce qui contredit l'existence d'un $\ell \in [a, b[$ comme ci-dessus en 1.

3. Soit $g : I \to \mathbb{R}$ continue sur I, dérivable à droite sur l'intérieur J de I et telle que $g'_d \leqslant 0$ sur J. Pour montrer que g est décroissante, il suffit, par continuité de montrer qu'elle est décroissante sur J. On raisonne par l'absurde. Si g est non décroissante, il existe un intervalle $[a, b] \subset J$ tel que $g(b) > g(a)$. Posons encore $m := \dfrac{g(b)-g(a)}{b-a} \neq 0$ et :

$$\forall x \in I, \quad f(x) := -x + \frac{2}{m}\, g(x).$$

On a a $m > 0$. On vérifie l'égalité $2g'_d = m\,(f'_d + 1)$ (sur J), ce qui contredit l'existence d'un $\ell \in [a, b[$ comme ci-dessus en 1 ; puisque $f'_d(\ell) \geqslant 0$ impliquerait $g'_d(\ell) \geqslant m/2 > 0$. Le résultat analogue pour les fonctions croissantes s'obtient en changeant f en $-f$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Limites et continuité

IV.2.1 Étudier la limite de la fonction f aux points proposés :

1. $f(x) := x^2 - 5x^4 + 3x$ en $+\infty$ et en $-\infty$.
2. $f(x) := (1-3x)(2x-1)^2$ en $+\infty$ et en $-\infty$.
3. $f(x) := \dfrac{x-1}{2} + \dfrac{2}{x-1}$ en $+\infty$, en $-\infty$ et en 1.
4. $f(x) := \dfrac{3x-1}{(x+2)(3-5x)}$ en -2.
5. $f(x) := \dfrac{5x^2+1}{2x-1}$ en $+\infty$, en $-\infty$ et en $1/2$.
6. $f(x) := \dfrac{3x^3-7x^2+5x-1}{3x^2-4x+1}$ en 1 et en $\dfrac{1}{3}$.
7. $f(x) := \sqrt{x^2+x+1} - x$ en $+\infty$.
8. $f(x) := \dfrac{\sqrt{x}-2}{x-4}$ en 4.
9. $f(x) := \dfrac{x^3+5x^2+7x+3}{2-\sqrt{x^2-x+2}}$ en -1.
10. $f(x) := x^3 \cos^4\left(\dfrac{1}{x^2}\right)$ en 0.

IV.2.2 Montrer que la fonction $x \mapsto \cos \frac{1}{x}$ n'a pas de limite en 0.

IV.2.3 Dans chacun des cas suivants, dire si l'application f se prolonge par continuité en a et donner le prolongement par continuité le cas échéant :

1. $f : [-3, +\infty[\setminus \{6\} \to \mathbb{R}$ définie par $f(x) := \dfrac{\sqrt{x+3} - 3}{x - 6}$ et $a = 6$.
2. $f : \mathbb{R} \setminus \{-1, 1\} \to \mathbb{R}$ définie par $f(x) := \dfrac{1}{1-x} - \dfrac{2}{1-x^2}$ et $a \in \{-1, 1\}$.

Dans le cas de fonctions prolongeables, discuter de la dérivabilité au point de prolongement.

IV.2.4 On se donne une application continue $f : [0,1] \to \mathbb{R}$ telle que $f([0,1]) \subset [0,1]$. Montrer qu'il existe $a \in [0,1]$ vérifiant $f(a) = a$.
Que se passe-t-il sans l'hypothèse de continuité ? sans la condition $f([0,1]) \subset [0,1]$?

IV.2.5 Un randonneur part du village de La Gurraz en Haute-Tarentaise pour se rendre au refuge de Turia près du Mont Pourri. Après avoir passé la nuit au refuge, il redescend au village le lendemain en empruntant le même itinéraire que la veille à l'aller.
Montrer que si les heures de départ et d'arrivée sont identiques pour les deux jours, il existe au moins un endroit du sentier où le randonneur est passé exactement à la même heure.

IV.2.6

1. Soit f une fonction continue de $\mathbb{R}$ dans $\mathbb{R}$, périodique et admettant des périodes arbitrairement petites (c'est-à-dire que l'ensemble des périodes de f admet 0 pour borne inférieure). Montrer que f est constante.
2. Trouver toutes les applications continues de $\mathbb{R}$ dans $\mathbb{R}$ admettant 1 et $\sqrt{2}$ pour périodes.

IV.2.7 Soit f une fonction de $\mathbb{R}$ dans $\mathbb{R}$ périodique.

1. Montrer que l'ensemble T des réels τ tels que $\forall x \in \mathbb{R},\ f(x+\tau) = f(x)$ est un sous-groupe de $\mathbb{R}$.
2. Montrer que si f est continue non constante, alors il existe $a > 0$ tel que $T = a\mathbb{Z}$ et donc que a est la plus petite période de f.
   On pourra utiliser l'exercice IV.1.11 de la page 601.

IV.2.8
Soit $f : [a,b] \to \mathbb{R}$ une fonction croissante. Montrer, pour tous $c_1, c_2, \ldots, c_p \in [a,b]$, deux à deux distincts, l'inégalité (7) de la page 630 : $\displaystyle\sum_{i=1}^{p} \sigma(c_i) \leqslant f(b) - f(a)$.

IV.2.9

1. Caractériser les bijections bicontinues *décroissantes* $f : \mathbb{R} \to \mathbb{R}$ telles que $f(0) = 0$ et $f \circ f = id_{\mathbb{R}}$ par leurs graphes.
2. Donner une infinité d'exemples de bijections bicontinues décroissantes $f = \mathbb{R} \to \mathbb{R}$ telles que $f \circ f = id_{\mathbb{R}}$.

IV.2.10 $*$

1. Montrer que l'ensemble des homéomorphismes de $\mathbb{R}$ (c'est-à-dire des bijections bicontinues de $\mathbb{R}$ sur $\mathbb{R}$) forme un groupe G pour la loi de composition $\circ$.
2. Trouver un sous-groupe *infini* de G.
3. Trouver les sous-groupes *finis* de G. Montrer qu'il y en a une infinité.

Dérivabilité

IV.2.11 On considère la fonction réelle f définie sur $\mathbb{R}$ par $f(x) := |x|/\big(1+|1-x^2|\big)$.
Étudier sa continuité et sa dérivabilité.

IV.2.12 Étudier la continuité et la dérivabilité de la fonction f définie par $f(x) := e^{-x}+x+1$ si $x \leqslant 0$ et $f(x) := 2+x^2\ln x$, si $x>0$. Même question sur sa dérivée première.

IV.2.13 Calculer les « dérivées » des fonctions suivantes :

$$f_1(x) := 3^{2x}, \qquad f_2(x) := (\sin x)^{\tan x},$$

$$f_3(x) := \ln\left(\tan\left(\frac{x}{2}+\frac{\pi}{4}\right)\right), \qquad f_4(x) := \frac{\sqrt{a+x}}{\sqrt{a}+\sqrt{x}},$$

en précisant les domaines de définition et de dérivabilité.

IV.2.14 Déterminer les dérivées $n^{\text{èmes}}$ des fonctions :

$$f : x \mapsto \frac{1-x}{1+x} \qquad \text{et} \qquad g : x \mapsto \sin^2 x.$$

IV.2.15 Soit $f : [a,+\infty[\,\to \mathbb{R}$ une fonction dérivable, telle que $\lim\limits_{t\to+\infty} f(t) = f(a)$.
Montrer qu'il existe $c \in]a,+\infty[$ tel que $f'(c)=0$.

IV.2.16 $*$ Soit $f : [a,b] \to \mathbb{R}$ une fonction continue admettant en tout point de $[a,b[$ une dérivée à droite.

1. On suppose que, pour tout $x \in [a,b[$, $f'_d(x) > 0$.
   a) Soit $c \in [a,b[$. Montrer qu'il existe $h>0$ tel que $c+h \leqslant b$ et que, pour tout $u \in]c,c+h]$, $f(c) < f(u)$.
   b) Soit $x \in [a,b[$. On définit :
   $$E_x := \{v \in [x,b] \mid \forall u \in [x,v], f(x) \leqslant f(u)\}.$$
      – Montrer que E_x est un intervalle non vide.
      – Montrer que E_x admet une borne supérieure $\gamma \leqslant b$ et que $\gamma \in E_x$.
      – Montrer que $\gamma = b$. En déduire que $E_x = [x,b]$, puis que f est croissante sur $[a,b]$.
   c) Montrer que f est strictement croissante sur $[a,b]$.
2. On suppose que, pour tout $x \in [a,b[$, $f'_d(x) \geqslant 0$. Montrer que f est croissante sur $[a,b]$.

IV.2.17 $*$ *Théorème de Darboux*
Soient I un intervalle de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction dérivable sur I. Montrer que $f'(I)$ est un intervalle.

Fonctions réciproques, accroissements finis, formules de Taylor

IV.2.18 Montrer que les fonctions suivantes admettent une fonction réciproque et l'expliciter :

$$f : \begin{array}{rcl} [1,+\infty[& \longrightarrow & \mathbb{R} \\ x & \longmapsto & \dfrac{2x^2+x+2}{x^2+1} \end{array} \qquad \text{et} \qquad g : \begin{array}{rcl} [-\frac{1}{2},+\infty[& \longrightarrow & \mathbb{R} \\ x & \longmapsto & \dfrac{1}{x^2+x+1} \end{array}.$$

IV.2.19 Montrer que :

$$\forall x \in \mathbb{R}\,,\ |\sin x| \leqslant |x| \qquad \forall x \in [0\,,\frac{\pi}{2}],\ 1-\cos x \leqslant x\sin x$$

$$\forall x \in \,]-1,1[\,,\ |\arcsin x| \leqslant \frac{|x|}{\sqrt{1-x^2}}.$$

IV.2.20 Soit f une fonction dérivable sur $\mathbb{R}$, vérifiant $f(0)=0$ et $\forall x \geqslant 0\,,\ f(x) \geqslant x$.

1. Montrer que $f'(0) \geqslant 1$. En appliquant la formule des accroissements finis à la fonction f sur l'intervalle $\left[0,\dfrac{1}{n}\right]$, $\forall n \in \mathbb{N}^\star$, montrer qu'il existe une suite (u_n) telle que $\lim u_n = 0$ et $f'(u_n) \geqslant 1$ pour tout n.
2. On suppose en outre que $f''(0)$ existe et $f'(0)=1$. Montrer que l'on a alors $f''(0) \geqslant 0$.
3. Soit g la fonction définie par :

$$g(x) := x\left[1+x+\sin x \sin\frac{1}{x}\right] \quad \text{pour} \quad x \neq 0 \quad \text{et} \quad g(0)=0.$$

   (a) Montrer que g est dérivable sur $\mathbb{R}$ et étudier la continuité de g'.

   (b) Montrer que $g(x) \geqslant x$ pour tout $x \geqslant 0$.

IV.2.21 Montrer que $\forall x > -1,\ \dfrac{x}{x+1} \leqslant \ln(1+x) \leqslant x$.

IV.2.22 Soient $a,\ h \in \mathbb{R}$, avec $h>0$, et $f : [a,a+2h] \to \mathbb{R}$ une fonction deux fois dérivable sur $[a,a+2h]$. Montrer qu'il existe $\theta \in \,]0,1[$ tel que :

$$f(a)-2f(a+h)+f(a+2h) = h^2 f''(a+2\theta h)$$

IV.2.23 Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. On suppose que, pour tous $x,\ y \in \mathbb{R}$, $f(x+y)f(x-y) \leqslant f^2(x)$. Montrer que, pour tout $x \in \mathbb{R}$, $f(x)f''(x) \leqslant f'^2(x)$.

IV.2.24 Soient $P \in \mathbb{R}[X]$ un polynôme de degré *impair* et $f : \mathbb{R} \to \mathbb{R}$ une fonction de classe $\mathcal{C}^\infty$. On suppose que, pour tout $n \in \mathbb{N}$ et tout $x \in \mathbb{R}$, $\left|f^{(n)}(x)\right| \leqslant \left|P(x)\right|$. Montrer que f est identiquement nulle. Montrer que le résultat similaire est faux pour un polynôme de degré pair.

Fonctions convexes

IV.2.25

1. Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction définie par $t \in \mathbb{R} \mapsto f(t) := \ln(1 + e^t)$.
   Montrer que f est convexe sur $\mathbb{R}$.
2. Soient λ et μ deux réels positifs tels que $\lambda + \mu = 1$.
   Montrer que pour tous réels strictement positifs x et y, on a $1 + x^\lambda y^\mu \leqslant (1 + x)^\lambda (1 + y)^\mu$.

IV.2.26 *Comparaison des moyennes arithmétiques et géométriques.*
Soient $n \in \mathbb{N}^*$ et $x_1, \ldots, x_n \in \mathbb{R}_+^*$. Montrer que $\sqrt[n]{x_1 \ldots x_n} \leqslant \frac{x_1 + \cdots + x_n}{n}$.

IV.2.27 Soit $f : I \to \mathbb{R}$ une fonction. On suppose que, pour tous $x,\ y,\ z \in I$ tels que $x < y < z$, l'on a :

$$\begin{vmatrix} 1 & 1 & 1 \\ x & y & z \\ f(x) & f(y) & f(z) \end{vmatrix} \geqslant 0.$$

Montrer que f est convexe sur I.

IV.2.28 *Inégalité de Hölder*
Soient p et q deux réels strictement positifs tels que $\frac{1}{p} + \frac{1}{q} = 1$. Soient $n \in \mathbb{N}^*$ et $a_1,\ a_2, \ldots,\ a_n$, $b_1,\ b_2, \ldots,\ b_n$ des nombres réels positifs. Montrer l'inégalité :

$$\sum_{i=1}^{n} a_i b_i \leqslant \left(\sum_{i=1}^{n} a_i^p \right)^{1/p} \left(\sum_{i=1}^{n} b_i^q \right)^{1/q}. \tag{27}$$

IV.2.29 *Inégalité de Minkowski*
Soit $p \in \mathbb{R}$, $p \geqslant 1$. Soit $f : [0, 1] \to \mathbb{R}$ définie par $x \mapsto f(x) := (1 - x^{1/p})^p$.

1. Montrer que f est convexe sur I.
2. Soient $n \in \mathbb{N}^*$ et $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}_+$. Montrer l'inégalité :

$$\left(\sum_{i=1}^{n} (a_i + b_i)^p \right)^{1/p} \leqslant \left(\sum_{i=1}^{n} a_i^p \right)^{1/p} + \left(\sum_{i=1}^{n} b_i^p \right)^{1/p}. \tag{28}$$

3. Soit $n \in \mathbb{N}^*$. Pour tout $x := (x_1, \ldots, x_n) \in \mathbb{R}^n$ on note $\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$.
   Montrer que :
   (i) $\|x\|_p \geqslant 0$ et $\|x\|_p = 0$ si, et seulement si, $x = 0$;
   (ii) pour tout $\lambda \in \mathbb{R}$, on a $\|\lambda x\|_p = |\lambda| \|x\|_p$;
   (iii) pour tous $x, y \in \mathbb{R}^n$, on a $\|x + y\|_p \leqslant \|x\|_p + \|y\|_p$.
   Que peut-on dire si $p = 2$?

IV.2.30 Retrouver l'inégalité de Hölder (27) de l'exercice IV.2.28 en utilisant l'inégalité de Young (26) (*cf.* l'exercice 19 de la page 667).

IV.2.31 Soit $f : \mathbb{R}_+^* \to \mathbb{R}$ une fonction convexe.

1. Montrer que $f(x)/x$ admet une limite $\ell \in \mathbb{R} \cup +\infty$ quand x tend vers $+\infty$.
2. On suppose que $\ell \leqslant 0$. Montrer que f est décroissante.

Études de fonctions

IV.2.32 Étudier et tracer la courbe représentative de la fonction $f(x) := \dfrac{x^4}{x^2-1}$.

IV.2.33 Étudier les branches infinies des graphes des fonctions définies par :

$$f(x) := 3x + 2 - \sqrt{4x^2 - x}, \qquad g(x) := 3x - 5 + \frac{x^2+3}{e^{2x}-1}, \qquad h(x) := 2x + 1 - \sqrt{x}.$$

Fonctions transcendantes IV.3

Les fonctions réelles les plus simples sont celles que l'on peut construire à partir de la variable en utilisant des opérations élémentaires (addition, multiplication etc.), c'est-à-dire les polynômes et les fractions rationnelles. En ajoutant des extractions de racines carrées ou autres et, plus généralement, des solutions d'équations polynomiales, on obtient des fonctions plus générales, comme par exemple $x \mapsto 1/\sqrt{1+x^2}$ ou $x \mapsto \sqrt[3]{\sqrt{x-1}+\sqrt{2x-3}}$ (pour $x \geqslant 3/2$). Toutes ces fonctions sont dites *algébriques*.
Plus précisément une fonction f, définie sur un intervalle réel I (d'intérieur non vide), est dite *algébrique* si elle est *solution d'une équation algébrique à coefficients polynomiaux* c'est-à-dire s'il existe $n \in \mathbb{N}^*$ et des *polynômes* $p_0, \ldots, p_n \in \mathbb{R}[X]$ tels que $p_n \neq 0$ et $p_n(x)f(x)^n + p_{n-1}(x)f(x)^{n-1} + \cdots + p_0 = 0$ pour tout $x \in I$.
Une fonction qui n'est pas algébrique est dite *transcendante*[1]. Le lecteur comparera avec les notions de *nombre* algébrique et de *nombre* transcendant (*cf.* page 312).
On a depuis longtemps constaté que les fonctions algébriques *ne suffisent pas* aux besoins de l'analyse. Pour la résolution d'équations différentielles par exemple (une tâche fondamentale en physique), ou même simplement pour trouver des primitives, il faut construire de nouvelles fonctions qui sont transcendantes : par exemple la fonction logarithme qui est une primitive de $x \mapsto 1/x$.
Les fonctions étudiées dans ce module sont *toutes* transcendantes. Certaines d'entre elles ont des dérivées très simples (algébriques), et sont par suite essentielles dans la recherche de primitives. L'utilisation de réciproques (comme l'exponentielle, vue comme réciproque du logarithme), est aussi une manière naturelle d'obtenir des fonctions transcendantes à partir d'autres déjà construites.
Nous décrirons dans ce module les plus « utiles » et les plus anciennement introduites des fonctions transcendantes. On les appelle *fonctions transcendantes usuelles*. On peut (heureusement...) montrer qu'elles sont effectivement transcendantes[2] : nous l'admettrons et nous donnerons certaines preuves sous forme d'une série d'exercices (de IV.3.26 à IV.3.30).
Au cours du temps (principalement aux XVIIIe et XIXe siècles), pour satisfaire aux besoins des mathématiques pures ou appliquées, de la physique, de la mécanique et des sciences de l'ingénieur, de nouvelles fonctions transcendantes ont été progressivement introduites : fonction Γ, fonction ζ *de Riemann*, fonction *d'Airy*, fonction *exponentielle intégrale*, fonction *d'erreur*, fonctions *hypergéo-*

[1] Transcendant est défini dans le *Littré* par : *Qui monte, s'élève au delà du reste*.
[2] Ceci est affirmé dans beaucoup d'ouvrages et d'articles « sur la toile » mais n'est pas souvent prouvé...

métriques, *de Bessel*, *d'Hermite*, *de Legendre*, *elliptiques*, *de Mathieu*, ... On les désigne du terme (un peu vague...) de *fonctions spéciales*[3]. De nombreux (et volumineux...) traités leur sont consacrés et les systèmes de calcul formel permettent aujourd'hui de « manipuler » la plupart d'entre-elles.

Nous aurons besoin dans ce module d'un résultat d'intégration. Nous l'énonçons ci-dessous et nous l'admettrons provisoirement. Il sera démontré plus loin (*cf.* le théorème 21 de la page 758).

Théorème 1 (Cauchy). Étant donnée une fonction réelle f continue sur un intervalle I, il existe une fonction F dérivable sur I dont la dérivée est égale à f. Une telle fonction est appelée une *primitive* de f.

Deux primitives de f diffèrent d'une constante puisque leur différence a une dérivée nulle sur l'intervalle I. En particulier, si a est un point fixé de I, il existe une unique primitive F_a de f s'annulant en a. On la note $F_a \ : \ x \mapsto \displaystyle\int_a^x f(t)\,\mathrm{d}t$.

1 Fonctions logarithme et exponentielle

Le point de départ d'une présentation des fonctions transcendantes usuelles est la construction du logarithme népérien *ou* celle de l'exponentielle. Ces deux fonctions se déduisant l'une de l'autre par passage à la fonction réciproque, il suffit de construire *l'une* d'entre elles. Nous aurions pu partir de la construction de l'exponentielle faite dans le module II.5 (théorème et définition 21 de la page 272) et en déduire le logarithme. Nous avons choisi de faire le contraire : partant de la construction du logarithme népérien par intégration, nous en déduirons l'exponentielle et la *caractériserons* en la comparant avec sa dérivée. Cette méthode est « plus élémentaire » et plus proche de la définition de l'exponentielle donnée en terminale. Nous vérifierons bien sûr que nos deux constructions différentes de l'exponentielle conduisent à la même fonction.

1.1 Logarithme népérien

Définition 1. La fonction *logarithme népérien*, notée $\ln$, est l'unique primitive sur $\mathbb{R}_+^*$ de $x \mapsto 1/x$ qui s'annule en 1. Elle est donc définie sur $\mathbb{R}_+^*$ par :

$$\ln x = \int_1^x \frac{\mathrm{d}u}{u}\cdot$$

Proposition 2. La fonction logarithme népérien vérifie :

$$\forall (x,y) \in \left(\mathbb{R}_+^*\right)^2 \ , \ \ln(x\,y) = \ln x + \ln y.$$

Démonstration. Soit $y \in \mathbb{R}_+^*$. L'application $u_y \ : \ \begin{array}{ccc} \mathbb{R}_+^* & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \ln(xy) \end{array}$ est dérivable et,

[3]Certaines de ces fonctions contiennent des *paramètres* et peuvent être algébriques ou se réduire à des transcendantes usuelles pour certaines valeurs « exceptionnelles » de ceux-ci.

par la formule donnant la dérivée d'une fonction composée :

$$\forall x \in \mathbb{R}_+^*\,,\ u_y'(x) = \frac{y}{x\,y} = \frac{1}{x}\cdot$$

Les fonctions u_y et $\ln$ ayant même dérivée sur l'intervalle $]\,0, +\infty\,[$, leur différence $u_y - \ln$ est constante sur $\mathbb{R}_+^*$. Comme, en 1, cette différence vaut $\ln y$, on en déduit :

$$\forall x \in \mathbb{R}_+^*\,,\ \ln(x\,y) = \ln x + \ln y.$$

■

Conséquences. De la proposition précédente, on déduit :

1. $\forall (x, y) \in \mathbb{R}_+^*\,,\ \ln\left(\dfrac{x}{y}\right) = \ln x - \ln y,$
2. $\forall x \in \mathbb{R}_+^*\,,\ \forall n \in \mathbb{Z}\,,\ \ln(x^n) = n \ln x.$

Proposition 3. La fonction logarithme népérien est une bijection strictement croissante de $]\,0, +\infty\,[$ sur $\mathbb{R}$ vérifiant :

$$\lim_{x \to +\infty} \ln x = +\infty \qquad \text{et} \qquad \lim_{x \to +\infty} \frac{\ln x}{x} = 0$$

$$\lim_{x \to 0} \ln x = -\infty \qquad \text{et} \qquad \lim_{x \to 0} x\,\ln x = 0.$$

Les premières limites indiquent que la courbe représentative de la fonction logarithme présente une branche parabolique de direction asymptotique Oy.

Démonstration. La fonction logarithme est strictement croissante sur $]\,0, +\infty\,[$ puisque sa dérivée est strictement positive sur cet intervalle. Par conséquent :

$$\ln 3 > \ln 1 = 0,$$

et la suite :

$$\big(\ln(3^n)\big)_{n \in \mathbb{N}} = (n\,\ln 3)_{n \in \mathbb{N}}$$

tend vers $+\infty$. Donc la fonction logarithme népérien n'est pas majorée, et comme elle est croissante sur $]\,0, +\infty\,[$, on en déduit $\lim\limits_{x \to +\infty} \ln x = +\infty$.

On a alors $\lim\limits_{x \to 0} \ln(1/x) = +\infty$, c'est-à-dire :

$$\lim_{x \to 0} \ln x = -\infty.$$

La fonction logarithme népérien est donc une bijection strictement croissante de $]\,0, +\infty\,[$ sur $\mathbb{R}$.

Soit $x \geqslant 1$. Il existe $n \in \mathbb{N}$ tel que $3^n \leqslant x \leqslant 3^{n+1}$ et l'on a donc :

$$0 \leqslant \frac{\ln x}{x} \leqslant \frac{(n+1)\,\ln 3}{3^n} \leqslant \left(\frac{2}{3}\right)^n \ln 3$$

puisque, par une récurrence immédiate, on a $\forall n \in \mathbb{N}\,,\ n + 1 \leqslant 2^n$.

On en déduit donc $\lim\limits_{x \to +\infty} \frac{\ln x}{x} = 0$ et de la même façon que précédemment, en remplaçant x par $1/x$, on obtient $\lim\limits_{x \to 0} x\,\ln x = 0$. ■

Remarque. La fonction logarithme népérien est un isomorphisme de groupes de $(\mathbb{R}_+^*, \times)$ sur $(\mathbb{R}, +)$.

Exercice 1.
Montrer que pour tout $x > -1$, $\ln(1+x) \leqslant x$.
Solution. On définit $f : x \in]-1, +\infty[\mapsto f(x) := x - \ln(1+x)$.
On a $f'(x) = 1 - \frac{1}{1+x} = \frac{x}{1+x}$. Par suite $f'(x)$ est du signe de x et l'on en déduit que f atteint son *minimum* en $x = 0$. Puisque $f(0) = 0$, $f \geqslant 0$.

Exercice 2.
Montrer que la fonction logarithme népérien est solution d'une équation différentielle linéaire homogène du second ordre.
Solution. On a $(\ln x)' = 1/x$, donc $x(\ln x)' - 1 = 0$. En dérivant à nouveau, on obtient $0 = (x(\ln x)')' = x(\ln x)'' + (\ln x)' = 0$. Par suite $y := \ln$ vérifie l'équation différentielle linéaire homogène du second ordre $xy'' + y' = 0$.

1.2 Exponentielle

Proposition et définition 4. La fonction *exponentielle*, notée $\exp$, est la fonction réciproque de la fonction logarithme népérien.
C'est une bijection strictement croissante de $\mathbb{R}$ sur $\mathbb{R}_+^*$ vérifiant :

$$\lim_{x \to -\infty} \exp x = 0 \qquad \text{et} \qquad \lim_{x \to +\infty} \exp x = +\infty.$$

Elle est dérivable sur $\mathbb{R}$ et elle est égale à sa dérivée.

Démonstration.

- C'est la réciproque de la fonction logarithme népérien qui est une bijection strictement croissante de $\mathbb{R}_+^*$ sur $\mathbb{R}$, c'est donc une bijection strictement croissante de $\mathbb{R}$ sur $\mathbb{R}_+^*$.
- Elle est croissante et, comme son image est $\mathbb{R}_+^*$, on a :
$$\lim_{x \to -\infty} \exp x = 0 \qquad \text{et} \qquad \lim_{x \to +\infty} \exp x = +\infty.$$
- La fonction logarithme népérien est dérivable sur $\mathbb{R}_+^*$ et sa dérivée est strictement positive. Sa fonction réciproque, l'exponentielle, est donc dérivable en tout point x de $\mathbb{R}$ avec :
$$\exp'(x) = \frac{1}{\ln'(\exp x)} = \exp x.$$

■

On dit aussi que l'exponentielle est une solution de l'équation différentielle linéaire $y' = y$.

Proposition 5.

1. $\exp 0 = 1$ et $\forall x \in \mathbb{R}\,,\ \exp x > 0$.
2. $\forall (x, y) \in \mathbb{R}^2\,,\ \exp(x + y) = \exp x \exp y$ et $\exp(x - y) = \dfrac{\exp x}{\exp y}$.
3. $\forall x \in \mathbb{R}\,,\ \forall n \in \mathbb{Z}\,,\ \exp(n\,x) = (\exp x)^n$.

Démonstration. Comme la fonction logarithme est bijective, il suffit de montrer, pour chaque égalité, que les deux membres (strictement positifs) ont même logarithme et ceci est une conséquence des propriétés correspondantes de la fonction $\ln$. ■

Remarque. L'exponentielle étant la réciproque d'un isomorphisme de groupes, on en déduit que c'est un isomorphisme de groupes de $\mathbb{R}$ dans $\mathbb{R}_+^*$, ce qui permet de retrouver les résultats précédents.

Corollaire 6. La fonction exponentielle est *l'unique* fonction $f : \mathbb{R} \to \mathbb{R}$, dérivable sur $\mathbb{R}$, égale à sa dérivée et telle que $f(0) = 1$.

Démonstration. Soit $f : \mathbb{R} \to \mathbb{R}$, une fonction dérivable sur $\mathbb{R}$, égale à sa dérivée. La fonction exponentielle ne s'annule pas sur $\mathbb{R}$ donc la fonction $g : x \mapsto g(x) := f(x)/\exp x$ est dérivable sur $\mathbb{R}$ et :

$$g'(x) = \frac{f'(x)\exp x - f(x)(\exp' x)}{(\exp x)^2} = 0.$$

Donc $f(x) = C \exp x$, avec $C \in \mathbb{R}$. Puisque $f(0) = \exp 0 = 1$, on a $C = 1$ et $f = \exp$. ■

Le corollaire ci-dessus permet de montrer que la fonction exponentielle $\exp$ que nous venons de définir coïncide avec la restriction à $\mathbb{R}$ de l'exponentielle complexe définie dans le module II.5. Notons $g : \mathbb{R} \to \mathbb{R}$ cette restriction. On a $g' = g$ et $g(0) = 1$ (*cf.* page 272), donc $g = \exp$.

Théorème 7. La restriction à $\mathbb{R}$ de la fonction exponentielle complexe définie page 272 est égale à la fonction exponentielle définie ci-dessus dans la proposition-définition 4.

Exercice 3.

Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction dérivable telle que $f' = f$ et $f(0) = 1$. Montrer, *sans utiliser la connaissance des fonctions exponentielle et logarithme népérien*, qu'elle ne s'annule jamais et en déduire qu'elle est déterminée de façon unique.

Solution. On considère la fonction $\varphi : x \in \mathbb{R} \mapsto \varphi(x) := f(x)f(-x)$. On a, pour tout $x \in \mathbb{R}$, $\varphi'(x) = f'(x)f(-x) - f(x)f'(-x) = \varphi(x) - \varphi(x) = 0$, donc φ est constante. On a $\varphi(0) = f(0)f(0) = 1$, donc $\varphi = 1$. S'il existait un réel a tel que $f(a) = 0$, on aurait $\varphi(a) = 0$ et une contradiction.

Si $h' = h$ et $h(0) = 1$, on montre, en raisonnant comme plus haut que $g := h/f$ est de dérivée nulle donc constante. Cette constante est 1 puisque $g(0) = 1$. Ainsi $h = f$.

1.3 Représentation graphique des fonctions logarithme népérien et exponentielle

x	0	1	$+\infty$
$\ln$	$-\infty$ ↗	0 ↗	$+\infty$

x	$-\infty$	0	$+\infty$
$\exp$	0 ↗	1 ↗	$+\infty$

On note $\mathrm{e} = \exp 1$ l'unique réel dont le logarithme vaut 1.

FIGURE IV.3.1 – Fonctions logarithme et exponentielle

1.4 Logarithmes et exponentielles de base quelconque

Définition 2. Si a est un réel strictement positif et différent de 1, on appelle *logarithme de base* a, l'application, notée $\log_a$, définie sur $\mathbb{R}_+^*$ par :

$$\log_a x = \frac{\ln x}{\ln a}.$$

Exemple. Le *logarithme décimal* ($a = 10$) et le *logarithme binaire* ($a = 2$) sont couramment employés en sciences appliquées.

Propriétés.

1. $\log_a 1 = 0$ et $\log_a a = 1$.
2. $\forall (x, y) \in \mathbb{R}_+^*,\ \log_a(x\,y) = \log_a x + \log_a y$.

3. $\forall (x,y) \in \mathbb{R}_+^*\,,\ \log_a\left(\dfrac{x}{y}\right) = \log_a x - \log_a y\,.$
4. $\forall x > 0\,,\ \forall n \in \mathbb{Z}\,,\ \log_a(x^n) = n \log_a x\,.$

Définition 3. Si a est un réel strictement positif et différent de 1, la fonction logarithme de base a est une bijection de $\mathbb{R}_+^*$ sur $\mathbb{R}$. Sa réciproque est donc une bijection de $\mathbb{R}$ sur $\mathbb{R}_+^*$. On l'appelle *exponentielle de base* a, et on la note $\exp_a$.

Propriétés.

1. $\exp_a 0 = 1$ et $\exp_a 1 = a\,.$
2. $\forall (x,y) \in \mathbb{R}^2\,,\ \exp_a(x+y) = \exp_a x \, \exp_a y\,.$
3. $\forall (x,y) \in \mathbb{R}^2\,,\ \exp_a(x-y) = \dfrac{\exp_a x}{\exp_a y}\cdot$
4. $\forall x \in \mathbb{R}\,,\ \forall n \in \mathbb{Z}\,,\ \exp_a(n\,x) = (\exp_a x)^n\,.$

Remarques.

- Si $a = \mathrm{e}$, la fonction $\exp_a$ est la fonction exponentielle déjà vue précédemment.
- Pour $x \in \mathbb{R}$, si l'on pose $y = \exp_a x$, on a :
$$x = \log_a y = \frac{\ln y}{\ln a}$$
et donc $\ln y = x \ln a$, ce qui entraîne :
$$\boxed{\exp_a x = \exp(x \ln a)\,.}$$

Proposition 8. Soit $f : \mathbb{R} \to \mathbb{R}$ une application *continue* telle que, pour tous $x, y \in \mathbb{R}$, l'on ait $f(x+y) = f(x) + f(y)$. Alors f est une application linéaire.

Autrement dit les seuls endomorphismes continus du groupe additif $(\mathbb{R}, +)$ sont les applications linéaires.

Démonstration. L'application f est un endomorphisme du groupe $(\mathbb{R}, +)$.
Pour tout $q \in \mathbb{N}$, on a $f(qx) = qf(x)$.
Donc, pour tout $q \in \mathbb{N}^*$, $qf\left(\frac{y}{q}\right) = f\left(q\frac{y}{q}\right) = f(y)$; c'est-à-dire $f\left(\frac{y}{q}\right) = \frac{1}{q}f(y)$.
Pour tout $p \in \mathbb{Z}$, on a $f\left(\frac{p}{q}y\right) = p\left(\frac{y}{q}\right) = \frac{p}{q}f(y)$.
Notons $k := f(1)$. On a, pour tous $p \in \mathbb{Z}$ et $q \in \mathbb{N}^*$, $f\left(\frac{p}{q}y\right) = k\frac{p}{q}$.
Ainsi, pour tout $x \in \mathbb{Q}$, on a $f(x) = kx$.
Soit $x \in \mathbb{R}$. Puisque $\mathbb{Q}$ est dense dans $\mathbb{R}$ (*cf.* proposition 12 de la page 519), il existe une suite (x_n) de $\mathbb{Q}$ telle que $\lim\limits_{n\to+\infty} x_n = x$. Par continuité de f, on a :
$$f(x) = \lim_{n\to+\infty} f(x_n) = \lim_{n\to+\infty} kx_n = kx.$$
■

Corollaire 9. Soit $g : \mathbb{R}_+^* \to \mathbb{R}$ une application continue telle que, pour tous $x, y \in \mathbb{R}_+^*$, l'on ait $g(xy) = g(x) + g(y)$. Alors g est l'application nulle ou bien il existe $a > 0$, $a \neq 1$, tel que $g = \log_a$.

Démonstration. Notons $f := g \circ \exp$. L'application $f : \mathbb{R} \to \mathbb{R}$ est continue et, pour tous $x, y \in \mathbb{R}$, l'on a $f(x+y) = f(x) + f(y)$. Donc, d'après la proposition 8 ci-dessus, il existe $k \in \mathbb{R}$ tel que $f(x) = kx$. Donc, pour tout $t \in \mathbb{R}$, $g(\exp t) = f(t) = kt$ et, pour tout $x \in \mathbb{R}_+^*$, $g(x) = k \ln x$. Si $k = 0$, g est l'application nulle. Sinon, il existe $a \in \mathbb{R}_+^*$ tel que $\log_a = 1/k$ ($a := \exp(1/k)$) et $g(x) = \frac{\ln x}{\ln a} = \log_a(x)$. ∎

2 Fonctions racines et puissances

2.1 Fonctions racines

Soit $n \in \mathbb{N}^*$. La fonction $x \mapsto x^n$ est continue sur $\mathbb{R}$.

- Elle est strictement croissante sur $\mathbb{R}$ lorsque n est impair et définit ainsi une bijection de $\mathbb{R}$ sur $\mathbb{R}$. Par restriction elle induit une bijection de $\mathbb{R}_+$ sur $\mathbb{R}_+$.
- Elle est strictement croissante sur $\mathbb{R}_+$ lorsque n est pair et définit ainsi une bijection de $\mathbb{R}_+$ sur $\mathbb{R}_+$.

Définition 4. Soit $n \in \mathbb{N}^*$. La fonction *racine* $n^{\text{ème}}$ est la réciproque de $x \mapsto x^n$:

- de $\mathbb{R}$ sur $\mathbb{R}$ lorsque n est impair ; elle est donc définie sur $\mathbb{R}$;
- de $\mathbb{R}_+$ sur $\mathbb{R}_+$ lorsque n est pair ; elle est donc définie sur $\mathbb{R}_+$.

On la note $x \mapsto \sqrt[n]{x}$.

Proposition 10. Soit $n \in \mathbb{N}^*$. La fonction $\sqrt[n]{}$ est dérivable sur $\mathbb{R}^*$ si n est impair, sur $\mathbb{R}_+^*$ si n est pair et sa dérivée est $x \mapsto \dfrac{1}{n\,(\sqrt[n]{x})^{n-1}} = \dfrac{\sqrt[n]{x}}{n\,x}$.

Démonstration. La dérivée de $x \mapsto x^n$ est non nulle en tout point sauf en 0, donc sa réciproque est dérivable en tout point de son intervalle de définition sauf en $0 = 0^n$ (théorème 56 de la page 657). De plus, cette dérivée est donnée en tout point par :

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\sqrt[n]{x}\right) = \frac{1}{n\left(\sqrt[n]{x}\right)^{n-1}} = \frac{\sqrt[n]{x}}{n\,x}. \qquad ∎$$

2.2 Fonctions puissances

Soit a un réel strictement positif différent de 1.
On a $\exp_a 1 = a$, donc, d'après les propriétés de la page 683 :

$$\forall n \in \mathbb{Z},\ \exp_a n = \exp(n \ln a) = a^n.$$

Il est donc naturel de *définir* a^x, pour tout $x \in \mathbb{R}$, par $a^x := \exp_a x = \exp(a \ln x)$.

Si $a = 1$, la fonction $\exp_a$ n'est pas définie, mais on peut encore *définir* a^x, pour tout $x \in \mathbb{R}$, par $a^x := 1$.

Définition 5. Soit a un réel strictement positif. On *définit*, pour tout $x \in \mathbb{R}$:

$$a^x := \exp(x \ln a).$$

En particulier, on a : $\exp x = \mathrm{e}^x$. Si, de plus, $a \neq 1$, alors $a^x := \exp_a x$.

Pour tout $a > 0$, la fonction $\varphi_a : x \in \mathbb{R} \mapsto a^x \in \mathbb{R}_+^*$ est appelée *fonction puissance*. Si $a \neq 1$, on vérifie que $\varphi_a : \mathbb{R} \mapsto \mathbb{R}_+^*$ est bijective.
Les propriétés de l'exponentielle et du logarithme nous donnent immédiatement les règles de calcul habituelles sur les fonctions puissances :

Proposition 11. Pour tout $a \in \mathbb{R}_+^*$ et tout $(x, y) \in \mathbb{R}^2$, on a :

$$a^x\, a^y = a^{x+y} \quad \text{et} \quad (a^x)^y = a^{xy}.$$

Pour tout $(a, b) \in (\mathbb{R}_+^*)^2$ et tout $x \in \mathbb{R}$, on a $a^x\, b^x = (a\, b)^x$.

Pour tout $a \in \mathbb{R}_+^*$, la fonction φ_a est un *isomorphisme de groupes* entre le groupe additif $(\mathbb{R}, +)$ et le groupe multiplicatif $(\mathbb{R}_+^*, \times)$.

Soient $a, b, c > 0$. On a, en général $(a^b)^c \neq a^{(b^c)}$.
Par exemple, si $a = b := 2$ et $c := 3$, on a $(a^b)^c = a^{bc} = 2^6$ et $a^{(b^c)} = 2^8$.

Exercice 4.
Soit $\psi : \mathbb{R} \to \mathbb{R}$ une application continue telle que, pour tous $x, y \in \mathbb{R}$, on ait :

$$\psi(x + y) = \psi(x)\psi(y).$$

Alors ψ est l'application nulle ou il existe $a \in \mathbb{R}_+$ tel que $\psi = \varphi_a$.
Solution.

- Supposons qu'il existe $x_0 \in \mathbb{R}$ tel que $\psi(x_0) = 0$.
  Alors, pour tout $x \in \mathbb{R}$, on a $\psi(x) = \psi(x - x_0 + x_0) = \psi(x - x_0)\psi(x_0) = 0$.
- Supposons que ψ ne s'annule pas sur $\mathbb{R}$.
  On a, pour tout $x \in \mathbb{R}$, $\psi(x) = \psi(x/2 + x/2) = \psi(x/2)^2 > 0$. Ainsi ψ est strictement positive. Posons $f := \ln \circ \psi$.
  L'application f est continue et, pour tous $x, y \in \mathbb{R}$, on a $f(x + y) = f(x) + f(y)$. Donc, d'après le corollaire 9 de la page précédente, il existe $k \in \mathbb{R}$ tel que, pour tout $x \in \mathbb{R}$, on ait $f(x) = kx$. Par suite $\psi(x) = \exp(kx) = (\mathrm{e}^k)^x = a^x$, en posant $a := \mathrm{e}^k$.

Soit $\alpha \in \mathbb{R}$. La relation $x^\alpha = \exp(\alpha \ln x)$ permet de prouver le résultat suivant.

Proposition 12. La fonction $\varphi_\alpha : x \mapsto x^\alpha$ est dérivable sur $\mathbb{R}_+^*$ et :

$$\frac{\mathrm{d}}{\mathrm{d}x}(x^\alpha) = \frac{\alpha}{x} x^\alpha = \alpha\, x^{\alpha - 1}.$$

On a ainsi $\varphi'_\alpha = \alpha\,\varphi_{\alpha-1}$ ce qui donne les variations de φ_α en fonction du signe de α. On a immédiatement les limites aux bornes. D'où les variations :

x	0		1		$+\infty$
φ_α $(\alpha>0)$	0	$\nearrow$	1	$\nearrow$	$+\infty$
φ_α $(\alpha<0)$	$+\infty$	$\searrow$	1	$\searrow$	0

Si $\alpha > 0$, on peut prolonger la fonction φ_α par continuité en 0 en posant $\varphi_\alpha(0) = 0$. Comme $\varphi_\alpha/x = \varphi_{\alpha-1}(x)$, on en déduit :

- si $\alpha > 1$, la fonction φ_α est dérivable en 0 et $\varphi'_\alpha(0) = 0$;
- si $0 < \alpha < 1$, la fonction φ_α n'est pas dérivable en 0 mais son graphe possède une tangente verticale à l'origine ;
- si $\alpha = 1$, la fonction φ_α est l'identité qui est donc dérivable en 0 de dérivée 1.

Représentation graphique des fonctions puissances

FIGURE IV.3.2 – Fonctions puissances

Point Méthode

Pour dériver une expression de la forme $f(x) = u(x)^{v(x)}$, avec u et v dérivables, $u > 0$, il faut au préalable l'écrire $f(x) = \exp\big(v(x)\,\ln u(x)\big)$, puis utiliser la formule de dérivation d'une fonction composée. On obtient ainsi :

$$f'(x) = f(x)\left(v'(x)\,\ln u(x) + v(x)\,\frac{u'(x)}{u(x)}\right).$$

Remarques.

- Soit $n \in \mathbb{N}^*$. On a pour tout $x > 0$:
$$\left(x^{1/n}\right)^n = x = \left(\sqrt[n]{x}\right)^n$$
et donc les fonctions $x \mapsto x^{1/n}$ et $x \mapsto \sqrt[n]{x}$ coïncident sur $\mathbb{R}_+^*$. Elles ne sont pas pour autant égales, puisque la première est définie sur $\mathbb{R}_+^*$ et la deuxième sur $\mathbb{R}$ ou $\mathbb{R}_+$ suivant la parité de n.
- La formule de dérivation de la fonction racine $n^{ème}$ devient facile à retenir en pensant à la fonction $x \mapsto x^{1/n}$. En tout $x > 0$, la dérivée vaut :
$$\frac{1}{n}\, x^{\frac{1}{n}-1} = \frac{1}{n}\, x^{(1-n)/n} = \frac{1}{n\,(\sqrt[n]{x})^{n-1}}.$$
On remarque que le résultat, ainsi obtenu pour $x > 0$, reste valable pour tout $x \in \mathbb{R}^*$ lorsque n est impair.
- Faire attention toutefois de ne pas confondre les fonctions puissances et les fonctions racines $n^{èmes}$. En particulier, ne pas écrire $\sqrt[3]{x} = x^{1/3}$ pour $x \leqslant 0$ sous peine d'arriver à des contradictions en utilisant les règles de calcul habituelles sur les puissances. Par exemple :
$$-1 = \sqrt[3]{-1} = (-1)^{1/3} = (-1)^{2/6} = \left((-1)^2\right)^{1/6} = 1^{1/6} = 1$$
et même $(-1)^{1/3} = \left((-1)^{1/6}\right)^2$ qui n'a pas de sens.

Exercice 5.

Montrer que $\left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = 2$.

En déduire qu'il existe deux nombres *irrationnels* positifs α et β tels que α^β soit *rationnel*.

Solution. On a $\left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = \sqrt{2}^{\sqrt{2}\times\sqrt{2}} = \sqrt{2}^2 = 2$.

On a l'alternative suivante : $\sqrt{2}^{\sqrt{2}} \in \mathbb{Q}$ *ou* $\sqrt{2}^{\sqrt{2}} \notin \mathbb{Q}$.

- Si $\sqrt{2}^{\sqrt{2}} \in \mathbb{Q}$, on pose $\alpha = \beta := \sqrt{2}$; alors α et β sont *irrationnels* et α^β est *rationnel*.
- Si $\sqrt{2}^{\sqrt{2}} \notin \mathbb{Q}$, on pose $\alpha := \sqrt{2}^{\sqrt{2}}$ et $\beta := \sqrt{2}$; alors α et β sont *irrationnels* et $\alpha^\beta = 2$ est *rationnel*.

L'argument ci-dessus est simple mais *non constructif* : on *montre qu'il existe* α et β répondant à la question mais on ne les connaît pas. On peut en fait montrer que $\sqrt{2}^{\sqrt{2}}$ est irrationnel et même *transcendant*[4]. Pour d'autres exemples *explicites*, *cf.* l'exercice IV.3.18.

[4]Cela résulte d'un théorème *difficile* de Gelfond-Schneider ; *cf.* par exemple : Serge Lang *Algebra*, Appendix.

2.3 Croissances comparées des fonctions puissances, logarithme et exponentielle

Soit $\alpha > 0$. Les trois fonctions $x \mapsto \ln x$, $x \mapsto x^\alpha$ et $x \mapsto \exp x$ tendent vers $+\infty$ en $+\infty$. La proposition suivante compare « la vitesse » avec laquelle chacune tend vers $+\infty$.

Proposition 13. Pour tout $\alpha > 0$, on a :

$$\lim_{x\to+\infty} \frac{\ln x}{x^\alpha} = 0 \qquad \text{et} \qquad \lim_{x\to+\infty} \frac{\exp x}{x^\alpha} = +\infty.$$

Démonstration.

D'après la proposition 3 de la page 679, on sait déjà que $\lim\limits_{x\to+\infty} \dfrac{\ln x}{x} = 0$. Comme $\alpha > 0$, on a $\lim\limits_{x\to+\infty} x^\alpha = +\infty$ et donc $\dfrac{\ln x}{x^\alpha} = \dfrac{\ln(x^\alpha)}{\alpha\, x^\alpha} \underset{x\to+\infty}{\longrightarrow} 0$.

De même, en partant de $\lim\limits_{x\to+\infty} \dfrac{\mathrm{e}^x}{x} = +\infty$, on obtient :

$$\frac{\mathrm{e}^x}{x^\alpha} = \left(\frac{\exp(\frac{x}{\alpha})}{x}\right)^\alpha = \frac{1}{\alpha^\alpha}\left(\frac{\exp(\frac{x}{\alpha})}{\frac{x}{\alpha}}\right)^\alpha \underset{x\to+\infty}{\longrightarrow} +\infty. \qquad \blacksquare$$

On en déduit immédiatement les résultats suivants :

Corollaire 14. Pour tout $\alpha > 0$, on a :

$$\lim_{x\to 0} x^\alpha \ln x = 0 \qquad \text{et} \qquad \lim_{x\to+\infty} x^\alpha \exp(-x) = 0.$$

Exercice 6.

Soient α et β deux réels strictement positifs. Étudier les limites suivantes :

$$\lim_{x\to+\infty} \frac{(\ln x)^\alpha}{x^\beta} \qquad \text{et} \qquad \lim_{x\to+\infty} \frac{\exp(x^\alpha)}{x^\beta}.$$

Solution. En écrivant $\dfrac{(\ln x)^\alpha}{x^\beta} = \left(\dfrac{\ln x}{x^{\beta/\alpha}}\right)^\alpha$ et $\dfrac{\exp(x^\alpha)}{x^\beta} = \dfrac{\exp(x^\alpha)}{(x^\alpha)^{\beta/\alpha}}$ on obtient, à l'aide des résultats précédents :

$$\lim_{x\to+\infty} \frac{(\ln x)^\alpha}{x^\beta} = 0 \qquad \text{et} \qquad \lim_{x\to+\infty} \frac{\exp(x^\alpha)}{x^\beta} = +\infty.$$

---

3 Fonctions trigonométriques

3.1 Fonctions sinus et cosinus

Dans cette partie nous utiliserons la *définition* de l'exponentielle *complexe* et les *propriétés élémentaires* de cette fonction (*cf.* le théorème et définition 21 de la page 272). Il n'est pas nécessaire pour lire ce qui suit de connaître *la preuve* de ce théorème ; *l'énoncé* suffit.

On définit *rigoureusement* les fonctions sinus et cosinus à partir de l'exponentielle complexe :

$$\forall x \in \mathbb{R}, \quad \sin x := \frac{\mathrm{e}^{ix} - \mathrm{e}^{-ix}}{2i} \quad \text{et} \quad \cos x := \frac{\mathrm{e}^{ix} + \mathrm{e}^{-ix}}{2}.$$

Autrement dit $\cos x$ et $\sin x$ sont respectivement la partie réelle et imaginaire de e^{ix}. Pour les propriétés élémentaires de ces fonctions, le lecteur se reportera au module II.5 page 273. Rappelons le résultat suivant.

Proposition 15. La fonction sinus est 2π-périodique impaire, la fonction cosinus est 2π-périodique paire. Elles sont dérivables et l'on a pour tout $x \in \mathbb{R}$:

$$\sin'(x) = \cos(x) \qquad \text{et} \qquad \cos'(x) = -\sin(x).$$

De plus : $\sin^2 x + \cos^2 x = 1$.

Les variations sur $[0, \pi]$ des fonctions $\sin$ et $\cos$, dont on déduit les variations sur $\mathbb{R}$ par périodicité et (im)parité, sont données par le tableau ci-contre :

x	0		$\frac{\pi}{2}$		π
$\sin x$	0	$\nearrow$	1	$\searrow$	0
$\cos x$	1	$\searrow$	0	$\searrow$	-1

Les fonctions $\sin$ et $\cos$ vérifient les formules d'addition :

Proposition 16. Pour tout $(a, b) \in \mathbb{R}^2$:

$$\sin(a+b) = \sin a \cos b + \sin b \cos a \qquad \text{et} \qquad \cos(a+b) = \cos a \cos b - \sin a \sin b.$$

Démonstration. On *montre rigoureusement* les formules d'addition en utilisant la relation $\mathrm{e}^{i(a+b)} = \mathrm{e}^{ia}\mathrm{e}^{ib}$, puis en prenant les parties réelle et imaginaire. ∎

Remarque. En changeant b en $-b$, on obtient les formules :

$$\sin(a-b) = \sin a \cos b - \sin b \cos a \qquad \text{et} \qquad \cos(a-b) = \cos a \cos b + \sin a \sin b.$$

En particulier, on a :

$$\boxed{\sin 2a = 2 \sin a \cos a} \quad \text{et} \quad \boxed{\cos 2a = \cos^2 a - \sin^2 a = 1 - 2\sin^2 a = 2\cos^2 a - 1}$$

d'où l'on déduit :

$$\boxed{\cos^2 a = \frac{1 + \cos 2a}{2}} \qquad \text{et} \qquad \boxed{\sin^2 a = \frac{1 - \cos 2a}{2}}$$

(pour ne pas confondre ces deux formules, penser à prendre $a = 0$).
Pour tout $n \in \mathbb{N}$, $\mathrm{e}^{inx} = (\mathrm{e}^{ix})^n$. On en déduit une preuve très simple de la *formule de Moivre* :

$$\cos nx + i \sin nx = (\cos x + i \sin x)^n.$$

Soient $p, q \in \mathbb{R}^2$. En utilisant $p := \frac{p+q}{2} + \frac{p-q}{2}$ et $q := \frac{p+q}{2} - \frac{p-q}{2}$, on déduit de la proposition 16 de la page précédente et de la remarque qui la suit (avec $a := \frac{p+q}{2}$ et $b := \frac{p-q}{2}$) les résultats suivants.

Corollaire 17.

$$\cos p + \cos q = \quad 2\cos\frac{p+q}{2}\cos\frac{p-q}{2}, \quad \sin p + \sin q = 2\sin\frac{p+q}{2}\cos\frac{p-q}{2},$$
$$\cos p - \cos q = -2\cos\frac{p+q}{2}\cos\frac{p-q}{2}, \quad \sin p - \sin q = 2\cos\frac{p+q}{2}\sin\frac{p-q}{2}.$$

Remarque. Pour retrouver rapidement ces formules et ne pas les confondre, penser à regarder les cas particuliers $q = \pm p$, ainsi que $p = 0$ pour le signe.

FIGURE IV.3.3 – Graphes des fonctions sinus et cosinus

Exercice 7.
Calculer $\cos 3x$ et $\frac{\sin 3x}{\sin x}$ en fonction de $\cos x$.
Solution. On trouve $\cos 3x = 4\cos^3 x - 3\cos x$ et $\frac{\sin 3x}{\sin x} = 4\cos^2 x - 1$.

Le résultat de l'exercice ci-dessus s'étend au calcul, pour tout $n \in \mathbb{N}^*$, de $\cos nx$ et $\frac{\sin nx}{\sin x}$ en fonction de $\cos x$. On obtient ainsi les polynômes de Tchebychef de première et deuxième espèce T_n et U_n (*cf.* l'exemple de la page 267 : $\cos nx = T_n(\cos x)$ et $\frac{\sin nx}{\sin x} = U_n(\cos x)$.

Proposition 18. Soit $x,\ x_0 \in \mathbb{R}$.

(i) On a $\cos x = \cos x_0$ si, et seulement s'il existe $k \in \mathbb{Z}$ tel que $x = \pm x_0 + 2\pi k$;

(ii) On a $\sin x = \sin x_0$ si, et seulement s'il existe $k \in \mathbb{Z}$ tel que $x = x_0 + 2\pi k$ ou $x = \pi - x_0 + 2k\pi$;

Démonstration. On considère d'abord l'équation $\sin x = 0$. Par 2π-périodicité et imparité, il suffit de résoudre cette équation pour $x \in [0, \pi]$. En étudiant les variations de la fonction $\sin$, on montre qu'elle et strictement positive sur $]0, \pi[$. Par suite, pour $x \in [0, \pi]$,

$\sin x = 0$ si, et seulement si, $x = 0$ ou $x = \pi$. On en déduit que, pour $x \in \mathbb{R}$, $\sin x = 0$ si, et seulement s'il existe $k \in \mathbb{Z}$ tel que $x = k\pi$.

- D'après le corollaire 17, $\cos x - \cos x_0 = 0$ si, et seulement si, $\sin \frac{x-x_0}{2} = 0$ ou $\sin \frac{x+x_0}{2} = 0$. On a $\sin \frac{x-x_0}{2} = 0$ si, et seulement s'il existe $k \in \mathbb{Z}$ tel que $x = x_0 + 2k\pi$ et $\sin \frac{x+x_0}{2} = 0$ si, et seulement s'il existe $k \in \mathbb{Z}$ tel que $x = -x_0 + 2k\pi$.
  Par suite $\cos x = \cos x_0$ si, et seulement s'il existe $k \in \mathbb{Z}$ tel que $x = \pm x_0 + 2k\pi$.
- On résout $\sin x = \sin x_0$ en se ramenant au cas précédent à l'aide des formules :
  $$\sin x = \cos(\pi/2 - x) \quad \text{et} \quad \sin x_0 = \cos(\pi/2 - x_0).$$
  On utilise ensuite $\pi/2 - x_0 = x - \pi/2 + 2k\pi \Leftrightarrow x = \pi - x_0 - 2k\pi$. ■

3.2 Fonctions tangente et arc-tangente

3.2.1 Fonction tangente

Définition 6. On appelle *tangente* la fonction notée $\tan$ définie par :
$$\forall x \in \mathbb{R},\ x \notin \pi/2 + \mathbb{Z}\pi, \quad \tan x := \sin x / \cos x,$$
et *cotangente* la fonction notée $\cot$ définie par :
$$\forall x \in \mathbb{R},\ x \notin \mathbb{Z}\pi, \quad \cot x := \cos x / \sin x.$$

En France, le nom de la tangente était écrit tg jusqu'à une date assez récente ; cet usage est maintenant « contraire aux normes établies ». La co-tangente est peu utilisée, ce n'est rien d'autre que l'inverse de la tangente (les inverses des cosinus et sinus étaient aussi nommés autrefois *sécante* et *cosécante*[5], mais cet usage est maintenant très rare).

FIGURE IV.3.4 – Graphes des fonctions tangente et cotangente.

[5]Suivant une terminologie introduite en France au XVIIe siècle.

Le lecteur vérifiera facilement les propriétés suivantes : la tangente est de période π, elle est définie sur tout intervalle de la forme $]k\pi - \pi/2, k\pi + \pi/2[$ avec $k \in \mathbb{Z}$, et elle est strictement croissante sur de tels intervalles. Sa dérivée est :

$$\tan' x = \frac{1}{\cos^2 x} = 1 + \tan^2 x. \tag{1}$$

D'après la proposition 16 de la page 689, pour tous les réels a et b tels que $\tan a$, $\tan b$ et $\tan(a+b)$ existent, on a :

$$\tan(a+b) = \frac{\tan a + \tan b}{1 - \tan a \tan b}. \tag{2}$$

En changeant b en $-b$, on en déduit :

$$\tan(a-b) = \frac{\tan a - \tan b}{1 + \tan a \tan b}.$$

En particulier : $\tan 2a = \dfrac{2\tan a}{1 - \tan^2 a}$.

L'étude de la cotangente se fait en notant que $\cot x = \tan\left(\dfrac{\pi}{2} - x\right)$.

On a, pour tous $p, q \in \mathbb{R} \setminus \{\pi/2 + \mathbb{Z}\pi\}$:

$$\tan p - \tan q = \frac{\sin(p-q)}{\cos p \cos q}.$$

On en déduit que $\tan x = \tan x_0$ si, et seulement s'il existe $k \in \mathbb{Z}$ tel que $x = x_0 + k\pi$.

3.3 Fonction arc-tangente

La fonction tangente, restreinte à l'intervalle $]-\pi/2, \pi/2[$, est strictement croissante, et a pour limites respectives $-\infty$ et $+\infty$ quand x tend vers $\mp\pi/2$.
Elle admet donc une réciproque définie sur $\mathbb{R}$, et à valeurs dans $]-\pi/2, \pi/2[$.

Définition 7. On appelle *arc-tangente*, et l'on note $\arctan$, la fonction réciproque de la restriction de la tangente à $]-\pi/2, \pi/2[$. Cette fonction est définie sur $\mathbb{R}$ tout entier.

Le nom est une abréviation pour « l'arc dont la tangente est... », le terme « arc » étant évidemment une référence à la géométrie. L'ancienne notation française arctg est maintenant abandonnée. La notation anglo-saxonne est atan ou $\tan^{-1}$.
On a donc par définition :

$$\forall x \in \mathbb{R}, \qquad \tan(\arctan x) = x \tag{3}$$

$$\forall x \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[, \qquad \arctan(\tan x) = x. \tag{4}$$

On prendra garde à la restriction sur la valeur de x dans la dernière relation : comme la tangente est de période π, il n'est pas vrai que $\arctan(\tan x) = x$ pour tout x (même ceux pour lesquels $\tan x$ est définie). Par exemple $\arctan(\tan \pi) = \arctan 0 = 0$.
Ces relations deviennent quasi-évidentes grâce au résultat suivant qui n'est qu'une réécriture de la définition :

FIGURE IV.3.5 – Graphe de la fonction arc-tangente.

Proposition 19. Soit $x \in \mathbb{R}$.
Le réel $\arctan x$ est l'unique élément de $\left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$ dont la tangente vaut x.

Ainsi, pour $x \in \mathbb{R} \setminus \{\frac{\pi}{2} + k\,\pi\,;\ k \in \mathbb{Z}\}$, pour déterminer $\arctan(\tan x)$, il suffit de trouver un (en fait l'unique) élément de $\left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$ qui a même tangente que x, ce qui donne, en notant $\lfloor . \rfloor$ la *partie entière* :

$$\arctan(\tan x) = x - k\pi \quad \text{avec } k = \left\lfloor \frac{x}{\pi} + \frac{1}{2} \right\rfloor \in \mathbb{Z}.$$

En effet, on a :

$$\forall k \in \mathbb{Z}\,,\ x - k\pi \in \left]-\pi/2, \pi/2\right[\iff x + \pi/2 \in \left]k\pi, (k+1)\pi\right[\Longrightarrow k = \left\lfloor \frac{x}{\pi} + \frac{1}{2} \right\rfloor.$$

Exemple. Soit $x \in \mathbb{R}^*$; posons $\alpha := \frac{\pi}{2} - \arctan x$. Comme $\tan(\arctan x) = x$, on a $\tan(\alpha) = \frac{1}{x}$ et donc $\arctan(\frac{1}{x}) = \alpha$ lorsque α est strictement compris entre $-\frac{\pi}{2}$ et $\frac{\pi}{2}$, c'est-à-dire lorsque $x \in \mathbb{R}_+^*$.
Lorsque $x < 0$, on a $\alpha \in \left]\frac{\pi}{2}, \pi\right[$ et donc :

$$\alpha - \pi \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[\qquad \text{avec} \qquad \tan(\alpha - \pi) = \tan\alpha = \frac{1}{x}\cdot$$

On en déduit $\arctan(\frac{1}{x}) = \alpha - \pi = -\frac{\pi}{2} - \arctan x$.
Finalement :

$$\forall x \in \mathbb{R}^*, \qquad \arctan x + \arctan\frac{1}{x} = \begin{cases} \pi/2 & \text{si } x > 0 \\ -\pi/2 & \text{si } x < 0. \end{cases} \tag{5}$$

Proposition 20.

La dérivée de la fonction arc-tangente est la fonction fraction rationnelle $x \mapsto \dfrac{1}{1+x^2}$.

Démonstration. On utilise la formule de dérivation d'une fonction réciproque :

$$\arctan'(x) = \frac{1}{\tan'(\arctan(x))} = \frac{1}{1+\tan^2(\arctan(x))} = \frac{1}{1+x^2}. \qquad \blacksquare$$

En utilisant la proposition 20 et les propriétés des progressions géométriques, on peut obtenir par intégration un *développement en série* de la fonction $arctan$, qui peut être utilisé pour le calcul des décimales de π. Ce résultat a été obtenu à la fin du XIVe siècle par le mathématicien et astronome indien Madhava de Sangamagramma[6] et retrouvé vers 1670 par le mathématicien écossais James Gregory, puis par G. W. Leibniz. Le développement a été longtemps appelé *formule de Gregory* (ou *de Gregory-Leibniz*). On dit aujourd'hui *formule de Madhava-Gregory* (*cf.* l'exercice IV.3.13 de la page 709).

Remarque. Le lecteur pourra redémontrer les relations (5) en dérivant la fonction $x \mapsto \arctan x + \arctan \frac{1}{x}$ pour montrer qu'elle est constante sur chacun des deux intervalles $\mathbb{R}_+^*$ et $\mathbb{R}_-^*$.

Exercice 8.

Montrer que $\arctan 1/2 + \arctan 1/3 = \pi/4$.

Solution. On a $\tan(\arctan 1/2 + \arctan 1/3) = \dfrac{\frac{1}{2}+\frac{1}{3}}{1-\frac{1}{6}} = 1$.

Par ailleurs, de $0 < 1/2 < 1$ et $0 < 1/3 < 1$ et de la croissance de la fonction $\arctan$, on déduit :

$$0 < \arctan\frac{1}{2} < \pi/4 \quad \text{et} \quad 0 < \arctan\frac{1}{3} < \pi/4,$$

puis $0 < \arctan 1/2 + \arctan 1/3 < \pi/2$. Notons $\alpha := \arctan\frac{1}{2} + \arctan\frac{1}{3}$.
On a $\tan \pi/4 = 1 = \tan\alpha$ et $\alpha, \pi/4 \in]-\pi/2, \pi/2[$, donc $\alpha = \pi/4 = \arctan 1$.

La formule de l'exercice ci-dessus date de 1738 et elle est due à Leonhard Euler. Il y a de nombreuses variantes (avec deux termes ou plus) qui ont été utilisées[7], avec la formule de Madhava-Gregory, pour calculer un certain nombre de décimales de π (comme par exemple la formule de Machin, *cf.* l'exercice IV.3.14 de la page 709).

[6] Madhava, fondateur de la remarquable école mathématique du Kerala, a calculé π avec 11 décimales exactes.
[7] C'est une histoire assez mouvementée qui s'est poursuivie jusqu'à une date récente : en 2002 Yasumara Kamada et sont équipe de l'université de Tokyo ont obtenu plus d'un trillion (10^{18}) de décimales de π avec une formule de ce type ; *cf.* Jonathan Borwein, David Bailey, *Mathematics by Experiment,* 2008. La compétition continue, mais avec d'autres formules...

3.4 Expressions de $\sin x$ et $\cos x$ en fonction de $\tan \frac{x}{2}$ Paramétrage du cercle

De $\sin 2a = 2\sin a\,\cos a$ et $\cos 2a = \cos^2 a - \sin^2 a$ et $\cos^2 a + \sin^2 a = 1$, l'on déduit, pour tout $x \in \mathbb{R}$ non congru à π modulo 2π :

$$\sin x = 2\sin\frac{x}{2}\,\cos\frac{x}{2} \quad = \frac{2\sin\frac{x}{2}\,\cos\frac{x}{2}}{\cos^2\frac{x}{2}+\sin^2\frac{x}{2}} = \frac{2\tan\frac{x}{2}}{1+\tan^2\frac{x}{2}}. \tag{6}$$

$$\cos x = \cos^2\frac{x}{2} - \sin^2\frac{x}{2} = \frac{\cos^2\frac{x}{2}-\sin^2\frac{x}{2}}{\cos^2\frac{x}{2}+\sin^2\frac{x}{2}} = \frac{1-\tan^2\frac{x}{2}}{1+\tan^2\frac{x}{2}}. \tag{7}$$

Ainsi les fonctions $\sin$ et $\cos$ peuvent être exprimées comme *fonctions fractions rationnelles* de la tangente de « l'arc moitié ». Nous allons voir que l'on peut en déduire un *paramétrage rationnel* du cercle unité C (privé d'un point) dans le plan vectoriel euclidien.

En coordonnées euclidiennes (u, v), une équation de C est $u^2 + v^2 = 1$. Les fonctions $\cos$ et $\sin$ permettent un *paramétrage transcendant* de C : à $x \in \mathbb{R}$, on associe $u = \cos x$, $v = \sin x$ (on peut se limiter à $x \in [0, 2\pi[$). En posant $t := \tan x/2$, on en déduit le paramétrage $t \in \mathbb{R}^* \mapsto \left(u = \frac{1-t^2}{1+t^2}, v = \frac{2t}{1+t^2}\right)$ du cercle C privé du point $(-1, 0)$. Nous allons l'interpréter de façon géométrique.

On se place dans le plan vectoriel euclidien (P) muni d'un repère orthonormé $(O, \vec{\imath}, \vec{\jmath})$: $(u, v) \in \mathbb{R}^2$ est identifié au point $M \in (P)$ tel que $\overrightarrow{OM} = u\vec{\imath} + v\vec{\jmath}$.

Nous allons utiliser les notions d'angles de vecteurs et de droites (le lecteur se reportera au module II.5, page 268).

Nous noterons $(\widehat{\vec{V}_1, \vec{V}_2})$ l'angle orienté de deux vecteurs $\vec{V}_1$ et $\vec{V}_2$ (il est défini modulo 2π) et $(\widehat{D_1, D_2})$ l'angle orienté de deux droites D_1 et D_2 (il est défini modulo π).

On a $C = \{M \in (P) \mid \overrightarrow{OM} = u\vec{\imath} + v\vec{\jmath},\ u^2 + v^2 = 1\}$. On note $\theta := (\widehat{\vec{\imath}, \overrightarrow{OM}})$.

Alors $u = \cos\theta$ et $v = \sin\theta$. On dit que θ est *l'angle au centre* correspondant à l'arc de cercle BM (B étant le point de coordonnées $(1, 0)$).

Soit A le point de coordonnées $(-1, 0)$. On considère les droites D passant par A et non tangentes à C. On note t la pente d'une telle droite. C'est la tangente de l'angle de droites φ défini par l'axe des abscisses et D. Ainsi $t = \tan\varphi$ et $v = t(u + 1)$ est une équation de D. L'application qui à $t \in \mathbb{R}$ associe l'unique droite $D(t)$ de pente t passant par A est injective et son image est l'ensemble des droites passant par A qui sont différentes de la tangente à C en A.

Une droite $D(t)$, passant par A et non tangente à C, de pente t, coupe C en deux points *distincts* A et $M(t)$. Par suite, on peut définir une application $t \in \mathbb{R} \mapsto M(t) \in C \setminus \{A\}$; elle est évidemment bijective. On a ainsi un paramétrage du cercle. Nous allons le calculer explicitement. Notons $(u(t), v(t))$ les coordonnées de $M(t)$. On a :

$$v(t) = t\,(u(t) + 1) \quad \text{et} \quad u(t)^2 - 1 + v(t)^2 = 0,$$

donc $(u(t) + 1)\left(u(t) - 1 + t^2\,(u(t) + 1)\right) = 0$ et, en utilisant $M(t) \neq A$, $u(t) = \dfrac{1-t^2}{1+t^2}$.

On en déduit $v(t) = \dfrac{2t}{1+t^2}\cdot$ Le paramétrage est *rationnel*.

Si $\theta(t) := (\widehat{\overrightarrow{OB}, \overrightarrow{OM}(t)})$, on a $\cos\theta(t) = u(t) = \frac{1-t^2}{1+t^2}$ et $\sin\theta(t) = v(t) = \frac{2t}{1+t^2}\cdot$

En comparant avec (6) et (7), on en déduit facilement $t = \tan\frac{\theta(t)}{2}$ c'est-à-dire, en notant $\varphi(t) := \arctan t$, l'angle de droites défini par l'axe des abscisses et $D(t)$, $\varphi(t) = \frac{\theta(t)}{2}$. En géométrie, on dit que φ est *l'angle inscrit* correspondant à l'arc de cercle BM (*cf.* la figure IV.3.6). Nous avons donc montré un résultat classique en géométrie élémentaire : *l'angle inscrit est la moitié de l'angle au centre*.

FIGURE IV.3.6 – Angle inscrit et angle au centre

Le paramétrage du cercle décrit ci-dessus se généralise au cas d'une conique C et d'un point A de celle-ci. De tels paramétrages ont de nombreuses applications, en particulier en calcul intégral ou en géométrie projective (*cf.* aussi l'exercice II.1.14 de la page 107 pour une application arithmétique).

3.5 Arc-sinus et arc-cosinus

La fonction sinus n'est évidemment pas strictement monotone, et n'a donc pas de réciproque. Mais on peut la restreindre à un intervalle où elle est strictement monotone. La fonction étant impaire, il est naturel de choisir un tel intervalle centré en 0. La fonction sinus est strictement monotone sur $[-\pi/2, \pi/2]$ (avec pour image l'intervalle $[-1, 1]$) et elle n'est pas monotone sur un intervalle plus grand. On choisit donc $[-\pi/2, \pi/2]$.

Définition 8. On appelle *arc-sinus*, et l'on note $\arcsin$ la fonction réciproque de la restriction du sinus à $[-\pi/2, \pi/2]$. Cette fonction est définie sur $[-1, 1]$, impaire et strictement croissante.

On a donc par définition :

Proposition 21. Soit $x \in [-1, 1]$.
Le réel $\arcsin x$ est l'unique élément de $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ dont le sinus vaut x.

Ainsi :

$$\forall x \in [-1, 1], \qquad \sin(\arcsin x) = x \tag{8}$$

$$\forall x \in \left[-\frac{\pi}{2}, +\frac{\pi}{2}\right], \qquad \arcsin(\sin x) = x. \tag{9}$$

Comme pour l'arc-tangente, la dernière relation n'est pas vraie pour tout $x \in \mathbb{R}$: on peut seulement dire en général que x et $\arcsin(\sin x)$ ont le même sinus d'après (8), ce qui implique :

$$\forall x \in \mathbb{R}, \quad \arcsin(\sin x) = \begin{cases} x - 2k\pi & \text{si } x \in [2k\pi - \pi/2, 2k\pi + \pi/2], \\ \pi - x + 2k\pi & \text{si } x \in [2k\pi + \pi/2, 2k\pi + 3\pi/2] \end{cases}$$

(avec $k \in \mathbb{Z}$ dans tous les cas).

FIGURE IV.3.7 – Fonctions $\sin$ et $\arcsin$ (à gauche), $\cos$ et $\arccos$ (à droite) ; les fonctions $\sin$ et $\cos$ sont restreintes aux intervalles adéquats.

Proposition 22. La fonction arc-sinus est continue sur $[-1, 1]$, dérivable sur $]-1, 1[$, sa dérivée sur $]-1, 1[$ étant $x \mapsto 1/\sqrt{1 - x^2}$.

Noter que la dérivée n'existe pas en ± 1 (la courbe représentative admet une tangente verticale), parce que la dérivée du sinus s'annule en $\pm\pi/2$.

Démonstration. La continuité provient du théorème 29 de la page 637 concernant la continuité d'une fonction réciproque.

La dérivabilité sur $]-1, 1[$ ainsi que l'expression de la dérivée est une conséquence du théorème 56 de la page 657 :

$$\arcsin' x = \frac{1}{\sin'(\arcsin x)} = \frac{1}{\cos(\arcsin x)}.$$

On a $\arcsin x \in]-\pi/2, \pi/2[$, donc $\cos(\arcsin x) \geqslant 0$ et :

$$\cos(\arcsin x) = \sqrt{1 - \sin^2(\arcsin x)} = \sqrt{1 - x^2}.$$

Par suite, pour tout $x \in]-1, 1[$, on a $\arcsin' x = 1/\sqrt{1 - x^2}$. ∎

On peut définir de même la réciproque de la fonction cosinus, à condition de la restreindre à $[0, \pi]$ par exemple (on ne peut pas choisir un intervalle centré en 0, à cause de la parité du cosinus). La réciproque ainsi définie sur $[-1, 1]$ est décroissante mais n'est pas paire On l'appelle *arc-cosinus*, et on la note $\arccos$. En fait, cette fonction n'est pas vraiment

nouvelle car l'on déduit de $\cos(\pi/2 - x) = \sin x$ que :

$$\forall t \in [-1, 1], \qquad \arccos t = \frac{\pi}{2} - \arcsin t. \tag{10}$$

Par suite, la dérivée de $\arccos$ est $x \mapsto -1/\sqrt{1-x^2}$.
La fonction $\arccos$ permet de *définir* l'angle θ de deux vecteurs $\vec{u}$ et $\vec{v}$ (non nuls) dans l'espace à partir de leur produit scalaire et de leurs normes : $\theta := \dfrac{(\vec{u} \mid \vec{v})}{\|\vec{u}\| \, \|\vec{u}\|} \in [0, \pi]$.

4 Trigonométrie hyperbolique

4.1 Sinus, cosinus et tangente hyperboliques

Les notions de sinus et cosinus hyperboliques proviennent de la géométrie. Au lieu de considérer l'ordonnée v et l'abscisse u d'un point du cercle unité d'équation $u^2+v^2 = 1$ dans le plan vectoriel euclidien muni d'un repère orthonormé, on considère l'ordonnée et l'abscisse d'un point de la branche d'hyperbole équilatère d'équation $u^2 - v^2 = 1$ définie par $u > 0$.

Les fonctions trigonométriques hyperboliques ont été introduites vers 1760 par le jésuite italien Vincenzo Ricatti. elles ont été étudiées systématiquement un peu plus tard par le mathématicien français Jean-Henri Lambert.
Les fonctions trigonométriques hyperboliques sont utilisées en *relativité* restreinte (pour les calculs dans le *groupe de Lorentz*, voir le livre L2) ; *cf.* aussi les exercices IV.3.7 et IV.3.8.
Rappelons, pour tout $t \in \mathbb{R}$, que $\cos t = (\mathrm{e}^{it}+\mathrm{e}^{-it})/2$ et $\sin t = (\mathrm{e}^{it}-\mathrm{e}^{-it})/2i$. Autrement dit $\cos$ et $\sin$ sont les *parties réelles et imaginaires* de la fonction $t \mapsto \mathrm{e}^{it}$. On peut, de façon analogue, considérer les *partie paire et impaire* (*cf.*proposition et définition 1 de la page 620) de la fonction exponentielle $t \mapsto \mathrm{e}^t$.
On obtient ainsi de nouvelles fonctions $\sinh$ et $\cosh$ vérifiant certaines formules qui rappellent les formules de trigonométrie.

Définition 9. On appelle *cosinus hyperbolique* et *sinus hyperbolique*, et l'on note $\cosh$ et $\sinh$, les parties paire et impaire de la fonction exponentielle, c'est-à-dire les fonctions définies sur $\mathbb{R}$ par :

$$\cosh t := \frac{\mathrm{e}^t + \mathrm{e}^{-t}}{2} \qquad \text{et} \qquad \sinh t := \frac{\mathrm{e}^t - \mathrm{e}^{-t}}{2}.$$

On appelle *tangente hyperbolique* et l'on note $\tanh$ la fonction définie sur $\mathbb{R}$ par :

$$\tanh t := \frac{\sinh t}{\cosh t} = \frac{1 - \mathrm{e}^{-2t}}{1 + \mathrm{e}^{-2t}}.$$

Noter la différence avec la tangente pour ce dernier cas : le cosinus hyperbolique ne s'annule jamais, puisque c'est une somme d'exponentielles, donc la tangente hyperbolique est définie partout. L'inverse de la tangente hyperbolique s'appelle aussi *cotangente hyperbolique* (cotanh) ; elle est définie sur $\mathbb{R}^*$.
Auparavant, on utilisait en France les notations sh (au lieu de $\sinh$), ch (au lieu de $\cosh$) et th (au lieu de $\tanh$).

On vérifie facilement la propriété suivante :

$$\forall t \in \mathbb{R}, \qquad \cosh^2 t - \sinh^2 t = 1. \tag{11}$$

En particulier $\cosh t = \sqrt{1 + \sinh^2 t} \geqslant 1$ pour tout t.
Donnons quelques autres propriétés dont la démonstration est laissée au lecteur :

Proposition 23.

- $\cosh 0 = 1$, $\sinh 0 = 0 = \tanh 0$;
- $\cosh$ est paire, $\sinh$ et $\tanh$ sont impaires ;
- $\cosh t > |\sinh t|$ pour tout t ;
- la dérivée de $\cosh$ est $\sinh$;
- la dérivée de $\sinh$ est $\cosh$;
- la dérivée de $\tanh$ est $t \mapsto 1 - \tanh^2 t = 1/\cosh^2 t$;
- les trois fonctions $\cosh$, $\sinh$ et $\tanh$ sont strictement croissantes sur $\mathbb{R}_+$.

FIGURE IV.3.8 – Graphes des fonctions $\cosh$, $\sinh$, $\tanh$ et $x \mapsto \mathrm{e}^x/2$ (en pointillés).

La dernière propriété implique notamment que toutes ces fonctions *ne sont pas périodiques* : c'est une différence majeure avec les fonctions trigonométriques.
Lorsque t tend vers $+\infty$, e^{-t} tend vers zéro très rapidement et les deux fonctions $\cosh$ et $\sinh$ sont « très voisines » de $t \mapsto \mathrm{e}^t/2$ (voir figure IV.3.8) et tendent vers l'infini très rapidement. La ressemblance du graphe de $\cosh$ avec une parabole est donc « trompeuse » : en réalité la courbe se redresse bien plus vite qu'une parabole. Cette courbe s'appelle une *chaînette*, parce que l'on peut montrer que c'est la courbe d'équilibre d'une chaîne fine pesante suspendue entre deux points fixes à la même hauteur (dans le champ de pesanteur terrestre).
Notons que d'après leur définition, on a aussi $\cosh t + \sinh t = \mathrm{e}^t$ et $\cosh t - \sinh t = \mathrm{e}^{-t}$.

En écrivant ces relations pour $t = a$ et $t = b$, en utilisant $\mathrm{e}^a\,\mathrm{e}^b = \mathrm{e}^{a+b}$ et $\mathrm{e}^{-a}\,\mathrm{e}^{-b} = \mathrm{e}^{-a-b}$, on obtient par addition et soustraction la proposition suivante.

Proposition 24. Pour tous a, b réels, on a :

$$\begin{aligned}\cosh(a+b) &= \cosh a \cosh b + \sinh a \sinh b,\\ \sinh(a+b) &= \cosh a \sinh b + \sinh a \cosh b.\\ \tanh(a+b) &= \frac{\tanh a + \tanh b}{1 + \tanh a \tanh b}.\end{aligned} \tag{12}$$

Remarque. Par parité de la fonction cosinus hyperbolique et imparité des fonctions sinus et tangente hyperboliques, on en déduit facilement, en remplaçant b par $-b$, des formules similaires donnant $\cosh(a-b)$, $\sinh(a-b)$ et $\tanh(a-b)$.

Noter la similarité avec les formules de la proposition 16 de la page 689 et la relation (2) de la page 692 : seul un signe a changé. Le lecteur est invité à retrouver ses formules de trigonométrie favorites sous une forme équivalente en trigonométrie hyperbolique (noter les nombreux changements de signes qui interviennent, dans des formules par ailleurs similaires).

Une autre façon de retrouver les formules est de noter que l'exponentielle étant définie pour des nombres complexes, le cosinus, sinus, cosinus hyperbolique et sinus hyperbolique le sont aussi : par somme ou différence d'exponentielles dans tous les cas (*cf.* 4.4). On a de plus $\cos z = \cosh(iz)$, $i \sin z = \sinh(iz)$ et inversement $\cosh z = \cos(iz)$, $i \sinh z = \sin(iz)$; on peut donc « déduire » les formules de trigonométrie hyperbolique des autres (ou inversement), en utilisant des variables complexes.

4.2 Expressions de $\sinh x$ et $\cosh x$ en fonction de $\tanh \frac{x}{2}$ Paramétrage de l'hyperbole équilatère

De $\sinh 2a = 2 \sinh a \cosh a$ et $\cosh 2a = \cosh^2 a + \sinh^2 a$ et $\cosh^2 a - \sinh^2 a = 1$, l'on déduit :

$$\forall x \in \mathbb{R},\ \ \sinh x = 2 \sinh \frac{x}{2} \cosh \frac{x}{2} = \frac{2 \sinh \frac{x}{2} \cosh \frac{x}{2}}{\cosh^2 \frac{x}{2} - \sinh^2 \frac{x}{2}} = \frac{2 \tanh \frac{x}{2}}{1 - \tanh^2 \frac{x}{2}}; \tag{13}$$

$$\forall x \in \mathbb{R},\ \ \cosh x = \cosh^2 \frac{x}{2} + \sinh^2 \frac{x}{2} = \frac{\cosh^2 \frac{x}{2} + \sinh^2 \frac{x}{2}}{\cosh^2 \frac{x}{2} - \sinh^2 \frac{x}{2}} = \frac{1 + \tan^2 \frac{x}{2}}{1 - \tan^2 \frac{x}{2}}. \tag{14}$$

Les fonctions $\cosh$ et $\sinh$ permettent un *paramétrage transcendant* de la branche d'hyperbole équilatère d'équation $u^2 - v^2 = 1$ définie par $u > 0$: à $x \in \mathbb{R}$, on associe $u = \cosh x$, $v = \sinh x$. En posant $t := \tanh x/2$, on en déduit le *paramétrage rationnel* $t \in \mathbb{R}^* \mapsto \left(u = \frac{1+t^2}{1-t^2}, v = \frac{2t}{1-t^2}\right)$ de l'hyperbole équilatère. On peut donner une interprétation géométrique de ce paramétrage en utilisant les droites passant par le point $A := (-1, 0)$: on reprend *mutatis mutandis* la méthode de 3.4 à la page 695. Nous laissons les détails au lecteur à titre d'exercice.

FIGURE IV.3.9 – Graphes de la tangente hyperbolique et de sa réciproque $\operatorname{argtanh}$.

4.3 Réciproques des fonctions hyperboliques

Les fonctions $\sinh$ et $\tanh$ sont strictement croissantes sur $\mathbb{R}$, et ont pour images respectivement $\mathbb{R}$ et $]-1,1[$. La fonction $\cosh$, comme toute fonction paire non constante, n'est pas monotone, mais sa restriction à $\mathbb{R}_+$ l'est, et a pour image $[1,+\infty[$.

Définition 10. On appelle *argument du sinus, de la tangente ou du cosinus hyperbolique*, et l'on note $\operatorname{argsinh}$, $\operatorname{argtanh}$, $\operatorname{argcosh}$, les réciproques des fonctions $\sinh$, $\tanh$, et de la restriction de $\cosh$ à $\mathbb{R}_+$.
Ces fonctions sont respectivement définies sur $\mathbb{R}$, sur $]-1,1[$ et sur $[1,+\infty[$.

On utilise ici le terme général « argument », et non « arc », car ces fonctions sont sans lien avec la mesure des angles. On trouve parfois la notation $arcsinh$ (resp. $arccosh$) au lieu de $argsinh$ (resp. $argcosh$), en particulier dans certains langages de programmation.
On a donc :

$$\begin{aligned}
\forall x \in \mathbb{R},\ \sinh(\operatorname{argsinh} x) &= x,\\
\forall x \in \mathbb{R},\ \operatorname{argsinh}(\sinh x) &= x,\\
\forall x \in]-1,1[,\ \tanh(\operatorname{argtanh} x) &= x,\\
\forall x \in \mathbb{R},\ \operatorname{argtanh}(\tanh x) &= x,\\
\forall x \in [1,+\infty[,\ \cosh(\operatorname{argcosh} x) &= x,\\
\forall x \in \mathbb{R},\ \operatorname{argcosh}(\cosh x) &= |x|,
\end{aligned}$$

(le dernier cas s'obtient en étudiant séparément le cas $x < 0$ et en notant que $\cosh$ est paire).

FIGURE IV.3.10 – Graphes de $\operatorname{argsinh}$ et $\operatorname{argcosh}$ avec ceux de $\sinh$ et $\cosh$ restreinte à $\mathbb{R}_+$ (pointillés).

Les trois fonctions sont croissantes, et $\operatorname{argsinh}$ et $\operatorname{argtanh}$ sont impaires. On a de plus, toujours en utilisant la formule de la dérivée d'une fonction réciproque :

$$\begin{aligned} \forall t \in \mathbb{R}\,,\ \operatorname{argsinh}' t &= \frac{1}{\sqrt{1+t^2}}, \\ \forall t \in \,]1,+\infty[\,,\ \operatorname{argcosh}' t &= \frac{1}{\sqrt{t^2-1}}, \\ \forall t \in \,]-1,1[\,,\ \operatorname{argtanh}' t &= \frac{1}{1-t^2}. \end{aligned} \tag{15}$$

Comme les fonctions trigonométriques inverses ces fonctions se distinguent donc par une dérivée assez simple (algébrique). Mais, contrairement au cas trigonométrique, on peut les exprimer à l'aide du logarithme népérien.

Proposition 25. Pour tout t dans le domaine de définition des fonctions $\operatorname{argsinh}$, $\operatorname{argtanh}$ et $\operatorname{argcosh}$ respectivement, on a :

$$\operatorname{argsinh} t = \ln\left(t + \sqrt{1+t^2}\right), \qquad \operatorname{argcosh} t = \ln\left(t + \sqrt{t^2-1}\right),$$
$$\operatorname{argtanh} t = \frac{1}{2}\ln\left(\frac{1+t}{1-t}\right).$$

Démonstration. La preuve suivante est basée sur la notion de primitive (*cf.* l'exercice IV.3.16 pour une preuve différente). On va utiliser les formules (15) et l'unicité d'une primitive à l'addition d'une constante près.

- Cherchons une primitive de $t \mapsto \frac{1}{\sqrt{1+t^2}}$. On a $(\sqrt{1+t^2})' = \frac{t}{\sqrt{1+t^2}}$.

On en déduit $1+(\sqrt{1+t^2})' = \frac{t+\sqrt{1+t^2}}{\sqrt{1+t^2}}$, puis $(t+\sqrt{1+t^2})' = \frac{t+\sqrt{1+t^2}}{\sqrt{1+t^2}}$.
Ainsi $\frac{(t+\sqrt{1+t^2})'}{t+\sqrt{1+t^2}} = \frac{1}{\sqrt{1+t^2}}$ et $t \mapsto \ln\,(t+\sqrt{1+t^2})$ est une primitive de $\frac{1}{\sqrt{1+t^2}}$ sur $\mathbb{R}$. Elle s'annule en $t=0$ et elle est donc égale à $\operatorname{argsinh}$.

- On vérifie de façon similaire que $\frac{(t+\sqrt{t^2-1})'}{t+\sqrt{t^2-1}} = \frac{1}{\sqrt{t^2-1}}$. Ainsi $\frac{(t+\sqrt{t^2-1})'}{t+\sqrt{t^2-1}} = \frac{1}{\sqrt{t^2-1}}$ et $t \mapsto \ln\,(t+\sqrt{t^2-1})$ est une primitive de $\frac{1}{\sqrt{t^2-1}}$ sur $]1,+\infty[$. Il existe donc une constante C telle que, pour tout $t>1$, $\ln\,(t+\sqrt{t^2-1}) = \operatorname{argcosh} t + C$.
  On a $\ln\,(1+\sqrt{1-1}) = \operatorname{argcosh} 1 = 0$. En faisant tendre t vers 1, on obtient $0 = 0 + C$, donc $C=0$. Ainsi, pour tout $t \geqslant 1$, $\ln\,(t+\sqrt{t^2-1}) = \operatorname{argcosh} t$.
- On a $\frac{1}{1-t^2} = \frac{1}{2}\left(\frac{1}{1+t}+\frac{1}{1-t}\right)$, donc $t \mapsto \frac{1}{2}\ln\left(\frac{1+t}{1-t}\right)$ est une primitive de $\frac{1}{1-t^2}$ sur $]-1,1[$. Elle s'annule en $t=0$ et elle est donc égale à $\operatorname{argtanh} t$. ■

Avant que ces fonctions soient standardisées aux normes anglo-saxonnes, les fonctions trigonométriques hyperboliques étaient notées par les symboles ch, sh, th, Argch, Argsh et Argth, que l'on rencontre donc encore dans certains ouvrages.

Exercice 9.

On appelle *cotangente hyperbolique* la fonction inverse multiplicative de la fonction tangente hyperbolique et *argument de la cotangente hyperbolique* la fonction réciproque de la cotangente hyperbolique. Montrer que la fonction argument de la cotangente hyperbolique est définie sur $]-\infty,-1] \cup [1,+\infty[$, et l'exprimer par un logarithme, puis calculer sa dérivée.

Solution. Posons, pour $x \in \mathbb{R}$ et $y \in \mathbb{R}^*$:

$$x = \operatorname{cotanh} y = \frac{1}{\tanh y} = \frac{\cosh y}{\sinh y} = \frac{\mathrm{e}^y + \mathrm{e}^{-y}}{\mathrm{e}^y - \mathrm{e}^{-y}} = \frac{\mathrm{e}^{2y}+1}{\mathrm{e}^{2y}-1}.$$

Posant $u := \mathrm{e}^{2y} \neq 1$, on en déduit les relations :

$$x\,(u-1) = x\,(\mathrm{e}^{2y}-1) = \mathrm{e}^{2y}+1 = u+1,\ x = \frac{u+1}{u-1} \neq 1,\ u = \frac{x+1}{x-1},\ y = \frac{1}{2}\,\ln\left(\frac{x+1}{x-1}\right).$$

La dernière égalité définit bien la fonction inverse recherchée, impaire et décroissante sur chacun des deux intervalles $]-\infty,-1[$ et $]1,+\infty[$: puisque u est une exponentielle, on doit avoir $u>0$, ce qui équivaut à $x<-1$ *ou* $x>1$.
La dérivée de la fonction argument de la cotangente hyperbolique peut être obtenue à partir de celle de sa réciproque $1/\tanh$ mais l'on peut aussi la calculer en dérivant $\frac{1}{2}\,\ln\left(\frac{x+1}{x-1}\right)$.
On trouve $1/(1-x^2)$.
On peut noter que la fonction argument de la cotangente hyperbolique *et* la fonction $\operatorname{argtanh}$ permettent de définir une fonction sur la réunion $\mathbb{R} \setminus \{-1,1\}$ de leurs ensembles de définition. Cette dernière s'écrit :

$$x \mapsto \frac{1}{2}\,\ln\left|\frac{1+x}{1-x}\right|,$$

de dérivée égale à $1/(1-x^2)$ en tout point tel que $x^2 \neq 1$.

4.4 Extension au domaine complexe

Les fonctions transcendantes élémentaires se déduisent toutes de l'exponentielle complexe $z \mapsto \mathrm{e}^z$ (et plus précisément de l'exponentielle réelle et de la fonction $y \in \mathbb{R} \mapsto \mathrm{e}^{iy}$) par des opérations algébriques simples et le passage à une fonction réciproque (après éventuellement une restriction).

On peut étendre les fonctions trigonométriques et trigonométriques hyperboliques aux valeurs *complexes* de la variable en utilisant les *définitions* suivantes :

$$\forall z \in \mathbb{C},\ \sin z := \frac{\mathrm{e}^{iz} - \mathrm{e}^{-iz}}{2i}, \qquad \cos z := \frac{\mathrm{e}^{iz} + \mathrm{e}^{-iz}}{2},$$

$$\forall z \in \mathbb{C},\ \sinh z := \frac{\mathrm{e}^{z} - \mathrm{e}^{-z}}{2}, \qquad \cosh z := \frac{\mathrm{e}^{z} + \mathrm{e}^{-z}}{2},$$

$$\forall z \in \mathbb{C} \setminus \{\pi/2 + \pi\mathbb{Z}\},\ \tan z := \frac{\sin z}{\cos z}; \qquad \forall z \in \mathbb{C} \setminus \{i\pi/2 + i\pi\mathbb{Z}\},\ \tanh z := \frac{\sinh z}{\cosh z}\cdot$$

On vérifie $\cos z = \cosh(iz)$, $i \sin z = \sinh(iz)$, $i \tan z = \tan(iz)$ et inversement $\cosh z = \cos(iz)$, $i \sinh z = \sin(iz)$, $i \tan z = \tan z$.

On peut aussi se proposer d'étendre la fonction logarithme népérien aux valeurs complexes de la variable. Nous allons voir qu'il y a une difficulté. Rappelons (*cf.* page 263) que si z est un nombre complexe non nul, on peut l'écrire sous la forme $z = \rho \mathrm{e}^{i\theta}$, avec $\rho > 0$ et $\theta \in \mathbb{R}$. On a $\rho = |z|$, mais θ *n'est pas unique*, il n'est déterminé que modulo 2π. On dit que c'est *un* argument de z et l'on note $\theta \in \arg z$. On peut écrire (improprement) $z = |z|\mathrm{e}^{i \arg z} = \mathrm{e}^{\ln|z| + i \arg z}$ et l'on est tenté de définir le logarithme complexe de z par $\ln|z| + i \arg z$. Cette définition est incorrecte car $z \mapsto \arg z$ *n'est pas une fonction*. La difficulté est sérieuse : on peut montrer qu'il n'existe pas de fonction *continue* $f : \mathbb{C}^* \to \mathbb{C}$ telle que $\mathrm{e}^{f(z)} = z$, pour tout $z \in \mathbb{C}^*$ (*cf.* l'exercice IV.3.22). On évite cette difficulté en restreignant le domaine de définition : on suppose que $z \in \mathbb{C} \setminus \mathbb{R}_-$. On peut alors définir l'*argument principal* de z, noté $\operatorname{Arg} z$; il est caractérisé par $\mathrm{e}^{i \operatorname{Arg} z} = z/|z$ et $\operatorname{Arg} z \in]-\pi, \pi[$. On a $\mathrm{e}^{\ln|z| + i \operatorname{Arg} z} = |z|\mathrm{e}^{i \operatorname{Arg} z} = |z|\frac{z}{|z|} = z$. On peut donc *définir* une fonction Log sur le domaine $\mathbb{C} \setminus \mathbb{R}_-$ par $\operatorname{Log} z := \ln|z| + i \operatorname{Arg} z$. On l'appelle *la détermination principale* du logarithme complexe. Pour $x \in \mathbb{R}_+^*$, on a $\operatorname{Arg} x = 0$, donc $\operatorname{Log} x = \ln x$: la restriction à $\mathbb{R}_+^*$ du logarithme complexe est le logarithme réel.

Notons que la fonction Log n'est pas la seule solution de $\mathrm{e}^{f(z)} = z$. On a, pour tout $n \in \mathbb{Z}$, $\mathrm{e}^{\operatorname{Log} z + 2in\pi} = z$ et $f(z) := \operatorname{Log} z + 2in\pi$ est aussi une solution. Inversement on peut montrer que l'on obtient ainsi *toutes les solutions* f *continues* (*cf.* l'exercice IV.3.22).

Soient $z_1, z_2 \in \mathbb{C} \setminus \mathbb{R}_-$ tels que $z_1 z_2 \in \mathbb{C} \setminus \mathbb{R}_-$.
Il est possible que $\operatorname{Log} z_1 z_2 \neq \operatorname{Log} z_1 + \operatorname{Log} z_2$. Par exemple

$$\operatorname{Log}(\mathrm{e}^{2i\pi/3})^2 = \operatorname{Log} \mathrm{e}^{-2i\pi/3} = -2i\pi/3 \neq 2 \operatorname{Log} \mathrm{e}^{2i\pi/3} = 2 \times 2i\pi/3 = 4i\pi/3.$$

On vérifie que $\operatorname{Log} z_1 z_2 = \operatorname{Log} z_1 + \operatorname{Log} z_2$ si, et seulement si, $\operatorname{Arg} z_1 + \operatorname{Arg} z_2 \in]-\pi, \pi[$.

Nous reviendrons sur le logarithme complexe et sur l'extension au domaine complexe des fonctions trigonométriques inverses et des fonctions trigonométriques hyperboliques inverses (voir les « fonctions analytiques » dans le L2).

5 Formulaires

5.1 Dérivées des fonctions élémentaires

Fonction	*Dérivée*
$x^\alpha \qquad \alpha \in \mathbb{R}$	$\alpha\, x^{\alpha-1}$
e^x	e^x
$\ln x$	$\dfrac{1}{x}$
$\cos x$	$-\sin x$
$\sin x$	$\cos x$
$\tan x$	$1+\tan^2 x = \dfrac{1}{\cos^2 x}$
$\cosh x$	$\sinh x$
$\sinh x$	$\cosh x$
$\tanh x$	$1-\tanh^2 x = \dfrac{1}{\cosh^2 x}$

Fonction	*Dérivée*
$\operatorname{cotan} x$	$-1-\operatorname{cotan}^2 x = -\dfrac{1}{\sin^2 x}$
$\coth x$	$1-\coth^2 x = -\dfrac{1}{\sinh^2 x}$
$\arctan x$	$\dfrac{1}{1+x^2}$
$\arcsin x$	$\dfrac{1}{\sqrt{1-x^2}}$
$\arccos x$	$-\dfrac{1}{\sqrt{1-x^2}}$
$\operatorname{argsinh} x$	$\dfrac{1}{\sqrt{x^2+1}}$
$\operatorname{argcosh} x$	$\dfrac{1}{\sqrt{x^2-1}}$
$\operatorname{argtanh} x$	$\dfrac{1}{1-x^2}$

5.2 Fonctions trigonométriques inverses

$$x = \arcsin y \iff \big(y = \sin x \quad \text{et} \quad -\pi/2 \leqslant x \leqslant \pi/2\big)$$
$$x = \arccos y \iff \big(y = \cos x \quad \text{et} \quad 0 \leqslant x \leqslant \pi\big)$$
$$x = \arctan y \iff \big(y = \tan x \quad \text{et} \quad -\pi/2 < x < \pi/2\big)$$

$$\forall y \in [-1,1]\,,\ \sin(\arcsin y) = \cos(\arccos y) = y \qquad \forall y \in \mathbb{R}\,,\ \tan(\arctan y) = y$$
$$\forall y \in [-1,1]\,,\ \sin(\arccos y) = \cos(\arcsin y) = \sqrt{1-y^2}$$

$$\forall y \in \,]-1,1[\,,\ \tan(\arcsin y) = \frac{y}{\sqrt{1-y^2}} \quad \forall y \in \mathbb{R}\,,\ \cos(\arctan y) = \frac{1}{\sqrt{1-y^2}}$$
$$\forall y \in [-1,1] \setminus \{0\}\,,\ \tan(\arccos y) = \frac{\sqrt{1-y^2}}{y} \quad \forall y \in \mathbb{R}\,,\ \sin(\arctan y) = \frac{y}{\sqrt{1-y^2}}$$

$$y = \sin x \iff \exists k \in \mathbb{Z} \,:\, x = \arcsin y + 2k\pi \quad \text{ou} \quad x = \pi - \arcsin y + 2k\pi$$
$$y = \cos x \iff \exists k \in \mathbb{Z} \,:\, x = \arccos y + 2k\pi \quad \text{ou} \quad x = -\arccos y + 2k\pi$$
$$y = \tan x \iff \exists k \in \mathbb{Z} \,:\, x = \arctan y + k\pi$$

5.3 Fonctions trigonométriques hyperboliques inverses

$$\forall y \in \mathbb{R}\,,\ \operatorname{argsh} y = \ln\left(y + \sqrt{1+y^2}\right)$$
$$\forall y \in [1,+\infty[\,,\ \operatorname{argch} y = \ln\left(y + \sqrt{y^2-1}\right)$$
$$\forall y \in]-1,1[\,,\ \operatorname{argtanh} y = \frac{1}{2}\ln\frac{1+y}{1-y}$$

$$x = \operatorname{argsinh} y \iff y = \sinh x$$
$$x = \operatorname{argcosh} y \iff \big(y = \cosh x \quad \text{et} \quad x \geqslant 0\big)$$
$$x = \operatorname{argtanh} y \iff y = \tanh x$$

$$\forall y \in \mathbb{R},\quad \sinh(\operatorname{argsinh} y) = y \quad \cosh(\operatorname{argsinh} y) = \sqrt{1+y^2} \quad \tanh(\operatorname{argsinh} y) = \frac{y}{\sqrt{1+y^2}}$$

$$\forall y \in [1,+\infty[\,,\ \cosh(\operatorname{argcosh} y) = y \qquad \forall y \in]-1,1[\ \ \tanh(\operatorname{argtanh} y) = y$$
$$\forall y \in [1,+\infty[\,,\ \sinh(\operatorname{argcosh} y) = \sqrt{y^2-1} \qquad \forall y \in]-1,1[\,,\ \sinh(\operatorname{argtanh} y) = \frac{y}{\sqrt{1-y^2}}$$
$$\forall y \in [1,+\infty[\,,\ \tanh(\operatorname{argcosh} y) = \frac{\sqrt{y^2-1}}{y} \qquad \forall y \in]-1,1[\,,\ \cosh(\operatorname{argtanh} y) = \frac{1}{\sqrt{1-y^2}}$$

Exercice type corrigé

IV.3.0

Nous avons donné dans la partie 4.4 un lien entre les fonctions trigonométriques hyperboliques et les fonctions trigonométriques circulaires. L'objet de cet exercice est d'en étudier un autre.

1. On considère tous les couples $(\varphi,\theta) \in \mathbb{R}^2$ pour lesquels il existe $t > 0$ tel que $\varphi := \ln t$ et $\theta := 2\arctan t$. Montrer qu'il existe deux fonctions réciproques l'une de l'autre Φ et Θ telles que $\theta = \Theta(\varphi)$ et $\varphi = \Phi(\theta)$. Préciser leurs intervalles de définition et leurs écritures explicites.
2. Dans cette question et les suivantes, $\varphi \in \mathbb{R}$, $\theta \in]0,\pi[$, $t > 0$ et l'on suppose que $t = \mathrm{e}^{\varphi} = \tan\theta/2$. Les fonctions Φ et Θ sont définies comme dans la première question.

   Que peut-on dire si l'on change t en $1/t$?

   Que peut-on en déduire pour les graphes des fonctions Φ et Θ ?
3. Calculer les fonctions circulaires de θ en fonction de t puis de φ.
4. Calculer les fonctions hyperboliques de φ en fonction de t puis de θ.
5. Calculer les dérivées des deux fonctions Φ et Θ. En déduire des primitives de $1/\sin x$ et $1/\cosh x$.

Solution.

1. L'application $f : \varphi \mapsto t = \mathrm{e}^{\varphi}$ est une bijection de $\mathbb{R}$ sur $\mathbb{R}_+^*$ et $g : t \mapsto \theta = 2\arctan t$ est une bijection de $\mathbb{R}_+^*$ sur $]0, \pi[$. Notant $\Theta := g \circ f$ l'application composée, on a $\Theta : \varphi \mapsto \theta = 2\arctan \mathrm{e}^{\varphi}$ et l'on voit que Θ envoie bijectivement $\mathbb{R}$ sur $]0, \pi[$. Sa réciproque est notée Φ.
   On a $\Phi : \theta \mapsto \varphi = \ln(\tan\theta/2)$ et Φ envoie bijectivement $]0, \pi[$ sur $\mathbb{R}$.
2. Le changement involutif $t \in \mathbb{R}_+^* \mapsto 1/t \in \mathbb{R}_+^*$ transforme φ en $-\varphi$ et θ en $\pi - \theta$.
   Les graphes des fonctions Θ et Φ admettent donc un centre de symétrie.
3. On rappelle les formules suivantes :
$$\cos\theta = \frac{1-t^2}{1+t^2}, \qquad \sin\theta = \frac{2t}{1+t^2}, \qquad \tan\theta = \frac{2t}{1-t^2}$$
   (on notera que $\tan\theta$ n'est pas défini pour $t := 1$, c'est-à-dire pour $\theta = \pi/2$ ou $\varphi = 0$).
   En résultent aussitôt les égalités suivantes, là où elles ont un sens :
$$\cos\theta = -\tanh\varphi, \qquad \sin\theta = \frac{1}{\cosh\varphi}, \qquad \tan\theta = -\frac{1}{\sinh\varphi}\cdot$$
4. Les expressions « inverses » s'en déduisent immédiatement :
$$\cosh\varphi = \frac{1}{\sin\theta}, \qquad \sinh\varphi = -\frac{1}{\tan\theta}, \qquad \tanh\varphi = -\cos\theta.$$
5. Les calculs sont immédiats et donnent :
$$\frac{d\Theta}{d\varphi} = \frac{2t}{1+t^2} = \sin\theta = \frac{1}{\cosh\varphi}, \qquad \frac{d\Phi}{d\theta} = \frac{1+t^2}{2t} = \frac{1}{\sin\theta} = \cosh\varphi.$$
   On en déduit que $x \mapsto \ln(\tan x/2)$ est une primitive de $x \mapsto 1/\sin x$ sur $\mathbb{R}_+^*$ et sur $\mathbb{R}_-^*$, et que $x \mapsto 2\arctan \mathrm{e}^x$ est une primitive de $x \mapsto 1/\cosh x$ sur $\mathbb{R}$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

IV.3.1 Calculer explicitement les nombres :

$$A := \cos(2\arcsin 0{,}6), \qquad B := \arctan\frac{1}{2} + \arctan\frac{1}{3} \qquad \text{et} \qquad C := \tan\left(2\arctan\frac{\sqrt{3}}{2}\right)\cdot$$

IV.3.2 Simplifier les expressions $A(x) := \operatorname{argcosh}(\cosh x)$, $B(x) := \cosh(\operatorname{argth} x)$.

IV.3.3 Résoudre l'équation : $\arctan 2x + \arctan x = \dfrac{\pi}{4}\cdot$

IV.3.4 Simplifier les expressions : $f(x) := \cos^2(\arctan x)$; $g(x) := \arctan\sqrt{\dfrac{1-\cos x}{1+\cos x}}\cdot$

IV.3.5 Résoudre dans $\mathbb{R}$ les équations suivantes :

$$2\sinh x - \cosh x = 1; \qquad \operatorname{argcosh} x = \operatorname{argsinh}(2-x).$$

IV.3.6

1. Calculer $\cosh 3x$ en fonction de $\cosh x$.
2. Quel est le domaine de définition $\mathcal{D}_f$ de la fonction $f(x) := \operatorname{argch}(4x^3 - 3x)$?
3. Montrer que pour tout $x \geqslant 1$, on a $\operatorname{argcosh}(4x^3 - 3x) = 3\operatorname{argcosh} x$.

IV.3.7

1. Soient $a, b \in [0,1[$. Montrer que $\operatorname{argtanh}\frac{a+b}{1+ab}$ est bien défini et que l'on a la formule d'addition : $\operatorname{argtanh} a + \operatorname{argtanh} b = \operatorname{argtanh}\frac{a+b}{1+ab}$.
2. En relativité restreinte la formule de composition des vitesses (colinéaires) s'écrit : $v_1 \oplus v_2 = \dfrac{v_1+v_2}{1+\frac{v_1 v_2}{c^2}}$ (où c est la vitesse de la lumière dans le vide et $0 \leqslant v_1 < c$, $0 \leqslant v_2 < c$). Écrire cette formule de façon « intuitive » en utilisant la fonction $\operatorname{argtanh}$.

IV.3.8 $*$ *Rotations hyperboliques et transformations de Lorentz.*
On rappelle que l'on note $GL_2(\mathbb{R})$ le groupe multiplicatif des matrices 2×2 à coefficients réels dont le déterminant n'est pas nul. À $\theta \in \mathbb{R}$ on associe $\Phi(\theta) := \begin{pmatrix} \cosh\theta & -\sinh\theta \\ -\sinh\theta & \cosh\theta \end{pmatrix}$.
On dit que $\Phi(\theta)$ est une (matrice de) *rotation hyperbolique.*

1. Montrer que Φ définit une application de $\mathbb{R}$ dans $GL_2(\mathbb{R})$ et que cette application est un homomorphisme de groupes de $(\mathbb{R},+)$ dans $GL_2(\mathbb{R})$. Montrer que Φ est injective. Montrer que l'image de Φ est un groupe et la caractériser.
2. En relativité restreinte, dans un espace-temps de dimension 2, les formules de passage d'un repère galiléen à un autre, en translation uniforme de vitesse relative γ par rapport au premier, s'expriment par une *transformation de Lorentz*[8] $L_\gamma := \frac{1}{\sqrt{1-\gamma^2}}\begin{pmatrix} 1 & -\gamma \\ -\gamma & 1 \end{pmatrix}$ (en supposant la vitesse de la lumière c égale à 1 et $|\gamma| < 1$). Écrire $\Phi(\theta)$ sous forme de transformation de Lorentz L_γ et inversement. Étudier le produit de deux transformations de Lorentz et comparer à l'exercice précédent IV.3.7.
3. Soit $(u,v) \in \mathbb{R}^2$. Pour $\theta \in \mathbb{R}$, on note $\begin{pmatrix} u_\theta \\ v_\theta \end{pmatrix} := \Phi(\theta)\begin{pmatrix} u \\ v \end{pmatrix}$. Montrer que $u_\theta^2 - v_\theta^2 = u^2 - v^2$. Interpréter géométriquement le résultat en utilisant des hyperboles équilatères.
4. Pour cette question on utilise l'exponentielle des matrices (*cf.* la définition 25 de la page 803 au module IV.6). Soit $A := \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$.
   Montrer que, pour tout $\theta \in \mathbb{R}$, on a $\Phi(\theta) = \mathrm{e}^{\theta A}$. Expliquer le choix de A.
5. Reprendre les questions précédentes en remplaçant les fonctions trigonométriques hyperboliques par les fonctions trigonométriques (en adaptant convenablement les signes).

[8] On déduit facilement cette formule des deux postulats suivants : cette formule est linéaire et la vitesse de la lumière est la même dans tous les repères galiléens.

IV.3.9 Étudier, pour $x > 0$, la fonction $x \mapsto x^{1/x}$ et tracer sa courbe représentative.

IV.3.10

Étudier, pour $x \in \mathbb{R}$, la fonction $x \mapsto f(x) := \dfrac{\tan x}{1 + 2\cos x}$ et tracer sa courbe représentative.

Pour déterminer le signe de la dérivée de f on pourra introduire la fonction polynomiale définie par $g(X) := -2X^3 + 4X + 1$ et étudier le signe de g.

IV.3.11 Montrer que pour tout $x < 1$, on a $x + 1 \leqslant \mathrm{e}^x \leqslant \dfrac{1}{1-x}$.

IV.3.12 *Une formule de Lewis Carroll.*

1. Soient $x, y \in \mathbb{R}_+$ tels que $xy < 1$. Montrer que $\arctan x + \arctan y = \arctan\left(\frac{x+y}{1-xy}\right)$.
2. Soient $p, q, r > 0$ tels que $1 + p^2 = qr$. Montrer[9] :
$$\arctan\left(\frac{1}{p+r}\right) + \arctan\left(\frac{1}{p+q}\right) = \arctan\left(\frac{1}{p}\right).$$
Calculer $\arctan \frac{1}{\sqrt{3}+2}$. Retrouver la formule de l'exercice 8 de la page 694.
3. Soient p et q strictement positifs. Montrer : $\arctan \frac{1}{p+q} + \arctan \frac{q}{p^2+pq+1} = \arctan \frac{1}{p}$.

IV.3.13 *Formule de Madhava-Gregory : calcul de $\arctan x$ sous forme de somme de série.*
Pour tout $n \in \mathbb{N}$, on considère le polynôme u_n défini par :
$$u_n(x) := \sum_{k=0}^{n} \frac{(-1)^k}{2k+1} x^{2k+1}.$$

1. Déterminer la dérivée de u_n, et montrer que $|u_n'(x) - 1/(1+x^2)| \leqslant x^{2n}$, pour tout $x \in \mathbb{R}$.
2. En déduire que, si $a > 0$ est fixé, alors $\forall x \in [-a, a]$, $|\arctan x - u_n(x)| \leqslant a^{2n+1}$.
3. Quelle est la limite de la suite $\big(u_n(x)\big)_{n\in\mathbb{N}}$, pour $x \in \,]-1, 1[$ fixé ?
Écrire le résultat en utilisant la somme d'une série.

IV.3.14 *Formule de Machin et calcul de π.*

1. Donner une expression de $\tan 4x$ en fonction de $\tan x$.
En déduire la valeur de $\tan(4 \arctan 1/5)$.
2. Calculer $\tan(4 \arctan 1/5 - \arctan 1/239)$.
3. Déduire de ce qui précède la formule $\pi/4 = 4 \arctan \dfrac{1}{5} - \arctan \dfrac{1}{239}$ (formule de John Machin 1706).
4. Pourquoi $1/239$?

[9] Cette formule est attribuée au révérend Charles Lutwidge Dodgson, plus connu sous le nom de Lewis Carroll (auteur de *Alice's Adventures in Wonderland*, 1865).

5. Pour tout $n \in \mathbb{N}$, on pose :

$$s_n := 4 \sum_{k=0}^{n} \frac{(-1)^k}{2k+1} \left(\frac{4}{5^{2k+1}} - \frac{1}{239^{2k+1}} \right).$$

Montrer que s_n est un nombre rationnel, et que la suite (s_n) converge vers une limite ℓ dont on donnera la valeur (utiliser l'exercice précédent).

6. Plus précisément, montrer que $|s_n - \ell| \leqslant 5^{1-2n}$ (utiliser encore l'exercice précédent) ; quelle valeur faut-il donner à n pour avoir une précision sur le calcul de ℓ inférieure à 10^{-2} ? à 10^{-16} ? à 10^{-100} ? Faire le calcul dans le premier cas.

IV.3.15 *Formules du type de Machin.*[10]
Montrer que $2 \arctan 1/2 - \arctan 1/7 = \pi/4$ et $2 \arctan 1/3 + \arctan 1/7 = \pi/4$.

IV.3.16 Donner une autre preuve de la proposition 25 de la page 702 en résolvant les équations $y = \sinh x$, $y = \cosh x$, $y = \tanh x$.
On utilisera des équations du second degré en $X := \mathrm{e}^x$ ou $X := \mathrm{e}^{2x}$.

IV.3.17 Soient $n \geqslant 1$ un entier, et a, b des réels avec $b \neq 0$. On pose $E := \sum_{k=0}^{n-1} \mathrm{e}^{(a+kb)}$.

1. Montrer que $E = \mathrm{e}^{a+(n-1)b/2} \dfrac{\sinh nb/2}{\sinh b/2}$.
2. En déduire la valeur de $C := \sum_{k=0}^{n-1} \cosh(a+kb)$ et $S := \sum_{k=0}^{n-1} \sinh(a+kb)$.
   On pourra d'abord chercher la valeur de $C+S$ et $C-S$.
3. Déterminer de même la valeur de $\sum_{k=0}^{n-1} \cos(a+kb)$ et $\sum_{k=0}^{n-1} \sin(a+kb)$.

IV.3.18 On revient sur une question de l'exercice 5 de la page 687.

1. Montrer que, pour tout $m \in \mathbb{N}^*$, $m \geqslant 2$, $\sqrt[m]{10}$ est irrationnel.
2. Montrer que $\log_{10} 2$ est irrationnel. Montrer que $\sqrt{10}^{\log_{10} 4} = 2$. Conclure.
3. Montrer que $\log_2 3$ est irrationnel. Montrer que $\sqrt{2}^{\log_2 6} = 3$. Conclure.

IV.3.19 *Une autre définition des fonctions trigonométriques.*

1. Soit $f : \mathbb{R} \mapsto \mathbb{C}$ une fonction dérivable telle que $f' = if$ et $f(0) = 1$.
   Montrer que, pour tout $x \in \mathbb{R}$, $f(x) = \mathrm{e}^{ix}$.
2. On « oublie provisoirement » la définition des fonctions $\cos$ et $\sin$. On rappelle que les solutions sur $\mathbb{R}$ de l'équation différentielle linéaire homogène du second ordre à coefficients constants :

$$y'' + y = 0 \tag{16}$$

[10] La première formule a été découverte par le mathématicien suisse Jakob Hermann (1706), la seconde par le mathématicien britannique Charles Hutton (1776). Elle a été retrouvée plus tard par Euler (1779).

forment un $\mathbb{R}$-espace vectoriel de dimension 2 (*cf.* le théorème 37 de la page 812 au module IV.6). On le note E.

Montrer qu'il existe une base de E formée d'une solution *paire* prenant la valeur 1 en 0 et de l'opposée de sa dérivée. On les note respectivement $\cos$ et $\sin$ et l'on pose $g := \cos + i \sin$. Calculer g' et conclure.

IV.3.20 Soient $a, z_0 \in \mathbb{C}$.

1. Résoudre, pour $z \in \mathbb{C}$, l'équation $\mathrm{e}^z = a$.
2. Résoudre, pour $z \in \mathbb{C}$, les équations $\cos z = \cos z_0$ et $\sin z = \sin z_0$.
3. On suppose que $z_0 \notin \{\pi/2 + \pi\mathbb{Z}\}$. Résoudre, pour $z \in \mathbb{C} \setminus \{\pi/2 + \pi\mathbb{Z}\}$, l'équation $\tan z = \tan z_0$.
4. Résoudre, pour $z \in \mathbb{C}$, l'équation $\cos z = a$.

IV.3.21 On désigne par Arg (resp. Log) la détermination principale de l'argument (resp. du logarithme complexe) sur $\mathbb{C} \setminus \mathbb{R}_-$ (*cf.* la section 4.4 de la page 704).

1. Montrer que, pour tout $x > 0$, on a : $\dfrac{x+i}{x-i} = \mathrm{e}^{2i \arctan \frac{1}{x}}$.
2. Montrer que, pour tout $x > 0$, on a : $\arctan \frac{1}{x} = \frac{1}{2i} \mathrm{Log}\, \frac{x+i}{x-i}$.
3. Montrer la relation $\frac{2+i}{2-i}\frac{3+i}{3-i} = i$ et la comparer à la formule de l'exercice 8 de la page 694 en utilisant la question 1.

   Mêmes questions pour les relations :

$$\left(\frac{5+i}{5-i}\right)^4 \left(\frac{239+i}{239-i}\right)^{-1} = i, \quad \left(\frac{2+i}{2-i}\right)^2 \left(\frac{7+i}{7-i}\right)^{-1} = i \quad \text{et} \quad \left(\frac{3+i}{3-i}\right)^2 \left(\frac{7+i}{7-i}\right) = i,$$

   en utilisant les formules des exercices IV.3.14 et IV.3.15.
4. Soient u et v des réels strictement positifs, $m \in \mathbb{N}^*$ et $n \in \mathbb{Z}$.

   Trouver une condition *nécessaire*[11], reliant « algébriquement » u, v, m et n, pour que l'on ait $m \arctan 1/u + n \arctan 1/v = \pi/4$.

IV.3.22 *Déterminations continues de l'argument d'un nombre complexe.*

Soit $z \in \mathbb{C}^*$.

Rappelons que si $\mathrm{e}^{i\theta} = z/|z|$, on dit que θ est un argument de z et l'on note $\theta \in \arg z$.

Si $z \notin \mathbb{R}_-$ et $\theta \in]-\pi, \pi[$, alors $\theta = \mathrm{Arg}\, z$.

1. Soit $z \in \mathbb{C} \setminus \mathbb{R}$. Calculer $\mathrm{Arg}\, z$ en utilisant les fonctions $\arccos$ et $\arcsin$. En déduire que la fonction Arg est continue sur $\mathbb{C} \setminus \mathbb{R}_-$.
2. Soit $g : \mathbb{C} \setminus \mathbb{R}_- \mapsto \mathbb{R}$ une application continue dont l'image est contenue dans $\mathbb{Z}$. Montrer qu'elle est constante. *On pourra considérer sa restriction aux segments* $[1, u]$, $u \notin \mathbb{R}_-$.
3. Soit $F : \mathbb{C} \setminus \mathbb{R}_- \mapsto \mathbb{R}$ une application continue telle que, pour tout $z \in \mathbb{C}$, $F(z) \in \arg z$. Montrer qu'il existe $n \in \mathbb{Z}$ tel que $F = \mathrm{Arg} + 2n\pi$.

[11] En utilisant cette condition le mathématicien danois Carl Størmer a montré en 1897 que, si l'on suppose u et v *entiers*, les *seules possibilités* correspondent aux 4 formules des exercices 8, IV.3.14 et IV.3.15. Il y a d'autres solutions avec u et v *rationnels* : par exemple $m := 20$, $n := 8$, $u := 7$, $v := 3/79$ (Euler 1755).

4. Montrer qu'il n'existe pas de fonction continue $G : \mathbb{C}^* \to \mathbb{R}$, dont la restriction à $\mathbb{C} \setminus \mathbb{R}_-$ est égale à la fonction Arg.
5. Montrer qu'il n'existe pas de fonction continue $f : \mathbb{C}^*$ telle que, pour tout $z \in \mathbb{C}^*$, l'on ait $\mathrm{e}^{f(z)} = z$.

IV.3.23 Soient $P, Q \in \mathbb{R}[X]$, avec $Q \neq 0$ et I un intervalle réel ne contenant pas de zéro de Q. Montrer que la fonction rationnelle $f : \begin{array}{ccc} I & \longrightarrow & \mathbb{R} \\ x & \longmapsto & P(x)/Q(x) \end{array}$ est algébrique (voir page 677).

IV.3.24 Soit $f : [2, +\infty[$, définie par $f : x \mapsto \sqrt[3]{\sqrt{x-1} + \sqrt{2x-3}}$.

1. Montrer que f est une fonction algébrique.
2. Montrer que f est strictement croissante sur $[2, +\infty[$. On note $g : [\sqrt[3]{2}, +\infty[\to \mathbb{R}$ la fonction réciproque de f. Montrer que g est une fonction algébrique et la calculer en utilisant un radical.
3. Soit $f_1 : [26, +\infty[\to \mathbb{R}$, définie par $f_1 : x \mapsto \sqrt[3]{\sqrt{2x-3} - \sqrt{x-1}}$.
   Que peut-on dire de cette fonction ?

IV.3.25 On définit trois fonctions sur $\mathbb{R}$ par $f : x \mapsto |x|$, $g : x \mapsto |x-1|$ et $h := f + g$. Montrer qu'elles sont algébriques.

« Les fonctions transcendantes usuelles sont transcendantes »

Les exercices suivants portent sur la transcendance des fonctions $\ln$, $\exp$, $\sinh$, $\cosh$, argsh et argch. Nous les compléterons en revenant sur les autres cas dans une prochaine édition du module II.6 du livre L2.

IV.3.26 Montrer que la fonction logarithme népérien n'est pas une fraction rationnelle, c'est-à-dire qu'il n'existe pas de polynômes $P \in \mathbb{R}[X]$ et $Q \in \mathbb{R}[X]^*$ tels que $\ln x = P(x)/Q(x)$, pour tout $x > 0$ tel que $Q(x) \neq 0$.

IV.3.27 $*$ Montrer, en utilisant l'exercice précédent, que $\ln$ est une fonction transcendante.

IV.3.28

1. Soit r une fraction rationnelle. Montrer que $\lim\limits_{x \to +\infty} r(x)\mathrm{e}^{-x} = 0$.
2. Montrer, en utilisant les relations de comparaison que les fonctions exponentielle et logarithme népérien sont transcendantes.

IV.3.29 $**$ Soient I un intervalle réel d'intérieur non vide et $f : I \to \mathbb{R}$ une fonction algébrique.

1. On suppose que f ne s'annule pas sur I. Montrer que $1/f$ est une fonction algébrique.
2. On suppose que f admet une fonction réciproque $g : J \to \mathbb{R}$. Montrer que g est une fonction algébrique.

IV.3.30 $*$

1. Montrer, en utilisant l'exercice IV.3.28, que les fonctions $\sinh$ et $\cosh$ sont transcendantes.
2. Montrer, en utilisant l'exercice IV.3.29, que les fonctions $\mathrm{argsinh}$ et $\mathrm{argcosh}$ sont transcendantes.

Séries numériques IV.4

Ce module est un premier contact avec la notion de *série* ou *somme infinie*. Une étude plus poussée des séries sera faite en L2. Par ailleurs, les séries considérées ici sont des séries *numériques*, *i.e.* formées de nombres. En L2 interviendront d'autres types de séries, à savoir les *séries de fonctions*, notamment les séries entières et les séries de Fourier, qui sont fondamentales.
La notion de série est très liée à la notion de suite, nous ferons donc souvent référence au module IV.1.

1 Convergence d'une série

1.1 Définitions

Les séries sont essentiellement des « sommes infinies ». Il faut bien sûr préciser ce que nous entendons par là. Partons par exemple de la somme suivante, n étant un entier strictement positif :

$$\sum_{k=1}^{n} \frac{1}{2^k} \;=\; \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^n} \;=\; \frac{1}{2}\left(\frac{1-\frac{1}{2^n}}{1-\frac{1}{2}}\right) \;=\; 1 - \frac{1}{2^n}\cdot \tag{1}$$

La deuxième égalité vient de la formule donnant la somme des termes d'une progression géométrique (proposition 23 de la page 533), la troisième égalité est évidente.
Nous avons d'écrire les égalités suivantes :

$$\sum_{k=1}^{+\infty} \frac{1}{2^k} \;=\; \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^n} + \cdots \;=\; 1. \tag{2}$$

Comme il n'est évidemment pas question d'effectuer une infinité d'additions, la seule possibilité raisonnable, pour donner un sens à ces égalités, est de « passer à la limite », en faisant tendre n vers $+\infty$ dans les égalités (1). C'est ce qui motive la définition 2 de la page suivante ci-après.

Soit $(u_n)_{n\geqslant 0}$ une suite à termes réels ou complexes. Nous appellerons *série de terme général* u_n l'expression $\sum\limits_{n\geqslant 0} u_n$, qui n'a pour l'instant qu'un sens « formel » : nous ne parlons pas de la « valeur » de cette expression.
Pour tout entier $n \geqslant 0$, posons :

$$U_n \;:=\; u_0 + u_1 + \cdots + u_n \;=\; \sum_{k=0}^{n} u_k. \tag{3}$$

Définition 1. Si $n \in \mathbb{N}$, la somme U_n définie dans la formule (3) est appelée *somme partielle d'indice* n de la série $\sum\limits_{n\geqslant 0} u_n$. La suite $(U_n)_{n\geqslant 0}$ est la *suite des sommes partielles* de cette série.

Définition 2. On dit qu'une série $\sum\limits_{n\geqslant 0} u_n$ *converge* ou est *convergente* si la suite $(U_n)_{n\geqslant 0}$ de ses sommes partielles *converge*, *i.e.* a une limite finie S. Si c'est le cas, S est appelée *somme* de cette série, et l'on écrit :

$$\sum_{n=0}^{+\infty} u_n \;=\; S. \tag{4}$$

Dans le cas contraire, on dit que la série $\sum\limits_{n\geqslant 0} u_n$ *diverge* ou est *divergente*.

Ainsi la nature (convergence ou divergence) d'une série est la même, par définition, que la nature de la suite de ses sommes partielles. Quant à la somme d'une série (convergente), ce n'est pas une notion algébrique, comme pour des sommes finies. En effet, elle repose sur la notion de limite d'une suite, notion qui fait partie de l'analyse.
Par ailleurs, il faut se rappeler que l'égalité (4) n'est pas une égalité ordinaire : elle signifie *d'abord* que la série $\sum\limits_{n\geqslant 0} u_n$ converge, et *ensuite* que sa somme est le nombre réel S.
La définition précédente se résume ainsi :

$$\sum_{n=0}^{+\infty} u_n \;:=\; \lim_{n\to+\infty} \left(\sum_{k=0}^{n} u_k \right), \tag{5}$$

en ce sens que la série converge si, et seulement si, la limite existe (dans $\mathbb{R}$ ou dans $\mathbb{C}$), auquel cas la somme de la série est *égale* à cette limite.

Remarques.

1. La somme d'une série n'est définie que si cette série converge.
2. Il faut bien distinguer la suite (u_n) de la suite (U_n), formée des sommes partielles de la série de terme général u_n.
3. Si (u_n) est définie pour $n \geqslant n_0$, où $n_0 \in \mathbb{Z}$ est fixé, on apporte les modifications suivantes : d'une part on pose $U_n := \sum\limits_{k=n_0}^{n} u_k$ pour tout $n \geqslant n_0$, d'autre

part la série $\sum_{n \geqslant n_0} u_n$ converge si, et seulement si, la suite $(U_n)_{n \geqslant n_0}$ converge ; dans ce cas, la somme de la série $\sum_{n \geqslant n_0} u_n$ est, par définition, la limite de la suite $(U_n)_{n \geqslant n_0}$, notée $\sum_{n=n_0}^{+\infty} u_n$.

4. On observera que l'expression $\sum_{n \geqslant 0} u_n$ désigne la série elle-même, alors que, *lorsque cette série converge*, sa somme est notée $\sum_{n=0}^{+\infty} u_n$.
5. Adoptons une convention : si (a_n) est une suite, notons (A_n) la suite des sommes partielles de la série $\sum_{n \geqslant 0} a_n$, idem pour (u_n) et (U_n), etc. Remarquer la correspondance entre lettres minuscules et lettres majuscules.
6. Soient $\sum_{n \geqslant 0} u_n$ une série et $p \in \mathbb{N}^*$. Pour tout $n \geqslant p$, on a évidemment :

$$U_n = U_{p-1} + \sum_{k=p}^{n} u_k.$$

Il en résulte que $\sum_{n \geqslant 0} u_n$ converge si, et seulement si, $\sum_{n \geqslant p} u_n$ converge, et, dans ce cas, les sommes de ces deux séries sont reliées ainsi :

$$\sum_{n=0}^{+\infty} u_n = U_{p-1} + \sum_{n=p}^{+\infty} u_n.$$

Ainsi l'on ne modifie pas la *nature* d'une série en lui enlevant un nombre fini de termes (consécutifs ou non) mais, en cas de convergence, la *somme* de la série est modifiée, comme le montre la formule précédente.

Exemples.

1. La série $\sum_{n \geqslant 0} 1$ diverge : ici $u_n := 1$, d'où $U_n = n+1$ pour tout n, et la suite $(n+1)$ est divergente.
2. Soit (u_n) une suite dont les termes sont nuls à partir d'un certain rang N. Pour tout $n \geqslant N$, on a $U_n = U_N$. Ainsi la suite (U_n) est stationnaire, donc convergente, et sa limite est U_N. Par définition, la série $\sum_{n \geqslant 0} u_n$ converge, et sa somme est U_N (c'est ici une somme d'un nombre fini de termes).
3. Prenons $u_n := \dfrac{1}{n(n+1)} = \dfrac{1}{n} - \dfrac{1}{n+1}$ pour tout $n \geqslant 1$.

   Les sommes partielles de la série $\sum_{n \geqslant 1} u_n$ se calculent simplement ; pour tout $n \geqslant 1$:

$$U_n = u_1 + u_2 + \cdots + u_n = \left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{n+1}\right).$$

Dans chaque parenthèse, sauf la première, le terme de gauche se simplifie avec le terme de droite de la parenthèse précédente. On obtient ainsi $U_n = 1 - 1/(n+1)$. Lorsque n tend vers $+\infty$, U_n tend donc vers 1. Il en résulte que la série $\sum_{n\geqslant 1} u_n$ converge, sa somme valant 1 :

$$\frac{1}{1.2} + \frac{1}{2.3} + \frac{1}{3.4} + \cdots + \frac{1}{n(n+1)} + \cdots = 1.$$

Voici maintenant un exemple fondamental de série :

Définition 3. On appelle *série géométrique* toute série de la forme $\sum_{n\geqslant 0} ar^n$, où a et r sont deux nombres complexes fixés ; a est le premier terme de cette série, et r en est la *raison*.

Théorème 1. Considérons une série géométrique $\sum_{n\geqslant 0} ar^n$, où a et r sont deux nombres complexes fixés et $a \neq 0$.

1. Si $|r| \geqslant 1$, cette série diverge.
2. Si $|r| < 1$, cette série converge, sa somme est donnée par la formule :

$$\sum_{n=0}^{+\infty} ar^n = \frac{a}{1-r}\cdot \tag{6}$$

Démonstration. Pour tout $n \in \mathbb{N}$, posons $u_n := ar^n$. La suite (u_n) est une *suite* géométrique, de premier terme a et de raison r. La proposition 23 de la page 533 permet de calculer les sommes partielles U_n de la série donnée ; pour tout $n \in \mathbb{N}$, on a :

$$\begin{cases} U_n = \sum_{k=0}^{n} u_k = a\left(\dfrac{1-r^{n+1}}{1-r}\right) & \text{si} \quad r \neq 1 \\ U_n = (n+1)a \quad \text{si} \quad r = 1. \end{cases} \tag{7}$$

Si $r = 1$, la suite $(U_n) = \big((n+1)a\big)$ est divergente. Supposons $r \neq 1$. Alors la suite (U_n) converge si, et seulement si, la suite (r^{n+1}) converge, autrement dit si $|r| < 1$ (théorème 38 de la page 549). Si c'est le cas, (r^{n+1}) converge vers 0, donc (U_n) converge vers $a/(1-r)$, ce qui signifie que la série $\sum_{n\geqslant 0} u_n$ converge et que sa somme vaut $a/(1-r)$. ∎

Dans le théorème précédent, prenons $a = r = 1/2$. La série $\sum_{n\geqslant 1} \frac{1}{2^n}$ est donc convergente, et sa somme vaut 1, ce qui justifie l'égalité (2) de la page 713.

Définition 4. Soit $\sum_{n\geqslant 0} u_n$ une série convergente, de somme S. Si $n \in \mathbb{N}$, posons :

$$R_n := \sum_{k=n+1}^{+\infty} u_k. \tag{8}$$

Le nombre R_n est appelé *reste d'indice* n de la série donnée. Ainsi :

$$S = \sum_{k=0}^{+\infty} u_k = U_n + R_n \tag{9}$$

pour tout $n \in \mathbb{N}$, où U_n est la somme partielle d'indice n de la série.

Le fait que cette définition soit licite, c'est-à-dire que la série définissant R_n converge, ainsi que l'égalité (9), résultent de la remarque 6 de la page 715.
En général, la somme S n'est pas connue de façon exacte, il s'agit plutôt d'en donner une valeur approchée, avec une précision imposée. Si l'on remplace S par la somme partielle U_n, pour un certain n, l'égalité (9) montre que l'erreur commise est R_n. Cette erreur n'est donc pas connue exactement, on cherche le plus souvent à la *majorer* (en module).
Dans l'exemple 3 de la page 715, on a $R_n = \dfrac{1}{n+1}$ pour tout $n \geqslant 1$. La convergence de la série considérée n'est pas très rapide : en prenant $1\,000$ termes, on obtient une erreur voisine de $1/1\,000$. Pour une série géométrique $\sum_{n\geqslant 0} ar^n$, où $a \neq 0$ et $|r| < 1$, les formules (6) et (7) impliquent la suivante, valable pour tout $n \geqslant 0$:

$$R_n = \frac{ar^{n+1}}{1-r}\cdot \tag{10}$$

Si par exemple $a = 1/2$ et $r = 1/2$, il suffit de 9 termes pour avoir une erreur de $2^{-10} = \dfrac{1}{1\,024} < \dfrac{1}{1\,000}$: la convergence est donc nettement plus rapide.

1.2 Premiers résultats

Comme pour les suites, on peut effectuer des opérations (addition, soustraction) sur les séries :

Théorème 2. Soient $\sum_{n\geqslant 0} u_n$ et $\sum_{n\geqslant 0} v_n$ deux séries convergentes et $\lambda \in \mathbb{C}$. Les séries de termes généraux u_n+v_n, u_n-v_n et λu_n respectivement convergent, et leurs sommes sont données par les formules suivantes :

$$\sum_{n=0}^{+\infty}(u_n+v_n) = \sum_{n=0}^{+\infty} u_n + \sum_{n=0}^{+\infty} v_n\,,\quad \sum_{n=0}^{+\infty}(u_n-v_n) = \sum_{n=0}^{+\infty} u_n - \sum_{n=0}^{+\infty} v_n\,,$$

$$\sum_{n=0}^{+\infty}\lambda u_n = \lambda\left(\sum_{n=0}^{+\infty} u_n\right).$$

Démonstration. Pour tout $n \in \mathbb{N}$, on a, avec la convention de la page 715 :

$$\sum_{k=0}^{n}(u_k+v_k) = \sum_{k=0}^{n} u_k + \sum_{k=0}^{n} v_k = U_n + V_n.$$

Par définition, la suite (U_n) converge, et sa limite U est la somme de la série $\sum_{n\geqslant 0} u_n$; de même la suite (V_n) converge, et sa limite V est la somme de la série $\sum_{n\geqslant 0} v_n$. D'après le théorème 33 de la page 546, la suite $(U_n + V_n)$ converge vers $U + V$. Ainsi la suite des sommes partielles de la série $\sum_{n\geqslant 0}(u_n + v_n)$ converge vers $U + V$, ce qui montre que cette série converge et que sa somme vaut $U + V$. D'où la conclusion voulue, dans le cas de l'addition. Pour la soustraction et la multiplication par λ, les preuves sont analogues. ■

Voici une conséquence utile de ce théorème. La preuve en est évidente, nous la laissons au lecteur (raisonner par l'absurde).

Corollaire 3. Soient $\sum_{n\geqslant 0} u_n$ et $\sum_{n\geqslant 0} v_n$ deux séries.
Si l'une de ces deux séries converge et si l'autre diverge, la série $\sum_{n\geqslant 0}(u_n + v_n)$ diverge.

L'étude des séries à termes complexes se ramène essentiellement à celle des séries à termes réels, grâce au résultat suivant :

Théorème 4. Soit $\sum_{n\geqslant 0} u_n$ une série à termes complexes. Pour tout n, notons a_n (resp. b_n) la partie réelle (resp. imaginaire) de u_n. Pour que la série $\sum_{n\geqslant 0} u_n$ converge il faut et il suffit que chacune des deux séries (à termes réels) $\sum_{n\geqslant 0} a_n$ et $\sum_{n\geqslant 0} b_n$ converge. Si tel est le cas, on a :

$$\sum_{n=0}^{+\infty} u_n = \sum_{n=0}^{+\infty} a_n + i\sum_{n=0}^{+\infty} b_n = \sum_{n=0}^{+\infty} \operatorname{Re} u_n + i\sum_{n=0}^{+\infty} \operatorname{Im} u_n. \tag{11}$$

Démonstration. Avec nos conventions habituelles on a, pour tout $n \in \mathbb{N}$:

$$U_n = \sum_{k=0}^{n} u_k = \sum_{k=0}^{n}(a_k + ib_k) = \sum_{k=0}^{n} a_k + i\sum_{k=0}^{n} b_k = A_n + iB_n.$$

La conclusion vient du résultat analogue pour les suites (*cf.* le module IV.1). ■

Exercice 1.

Soit x un nombre réel fixé. Étudier la série $\sum_{n\geqslant 0} \dfrac{\cos nx}{2^n}$.

Solution. Notons a_n son terme général.

On a $a_n = \operatorname{Re} r^n$, où $r = \dfrac{\mathrm{e}^{ix}}{2}$, puisque $(\mathrm{e}^{ix})^n = \mathrm{e}^{inx} = \cos nx + i\sin nx$.

Comme $|r| = \dfrac{1}{2}$, la série géométrique $\sum_{n\geqslant 0} r^n$ converge, et sa somme vaut :

$$\sum_{n=0}^{+\infty} \frac{\mathrm{e}^{inx}}{2^n} = \frac{1}{1-r} = \frac{2}{2-\mathrm{e}^{ix}} = \frac{2(2-\mathrm{e}^{-ix})}{(2-\mathrm{e}^{ix})(2-\mathrm{e}^{-ix})}.$$

Or $(2-\mathrm{e}^{ix})(2-\mathrm{e}^{-ix}) = 4 - 2(\mathrm{e}^{ix}+\mathrm{e}^{-ix}) + 1 = 5 - 4\cos x$.
D'après le théorème précédent, la série $\sum\limits_{n\geqslant 0} \dfrac{\cos nx}{2^n}$ converge, et sa somme est la partie réelle de la somme de $\sum\limits_{n\geqslant 0} r^n$, soit :

$$\sum_{n=0}^{+\infty} \frac{\cos nx}{2^n} \;=\; \mathrm{Re}\left(\frac{2(2-\mathrm{e}^{-ix})}{5-4\cos x}\right) \;=\; \frac{2(2-\cos x)}{5-4\cos x}\,.$$

Soient $\sum\limits_{n\geqslant 0} u_n$ une série à termes réels ou complexes et (U_n) la suite de ses sommes partielles. Pour tout entier $n \geqslant 1$, on a :

$$u_n \;=\; \sum_{k=0}^{n} u_k - \sum_{k=0}^{n-1} u_k \;=\; U_n - U_{n-1}. \tag{12}$$

On en déduit deux résultats importants, le second généralisant le premier :

Théorème 5. Si une série $\sum\limits_{n\geqslant 0} u_n$ converge, alors u_n tend vers 0 lorsque n tend vers $+\infty$. La réciproque est fausse en général.

Démonstration. Supposons que la série $\sum\limits_{n\geqslant 0} u_n$ converge, soit S sa somme. Si n tend vers $+\infty$, U_n tend vers S par définition ; alors U_{n-1} tend aussi vers S, et la formule (12) montre que u_n tend vers 0. Le fait que la réciproque soit fausse sera vérifié plus loin. ■

Théorème 6. Soient $\sum\limits_{n\geqslant 0} u_n$ une série convergente et $\varepsilon \in \mathbb{R}_+^*$.
Il existe un entier $N \geqslant 0$ tel que, pour tous entiers p, q vérifiant $N \leqslant p \leqslant q$, on ait :

$$\left|\, u_p + u_{p+1} + \cdots + u_q \,\right| = \left|\sum_{k=p}^{q} u_k\right| \;\leqslant\; \varepsilon. \tag{13}$$

Démonstration. Reprenons les notations de la preuve précédente. Puisque la suite (U_n) converge vers S, il existe un entier $M \geqslant 0$ tel que $|U_n - S| \leqslant \dfrac{\varepsilon}{2}$ pour tout $n \geqslant M$. Posons $N := M+1$. Si p, q sont deux entiers tels que $N \leqslant p \leqslant q$, on a $q \geqslant M$ et $p-1 \geqslant M$, d'où :

$$\left|u_p + u_{p+1} + \cdots + u_q\right| = |U_q - U_{p-1}| \leqslant |U_q - S| + |S - U_{p-1}| \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \quad ■$$

En se limitant au cas $p = q$, on retrouve le théorème 5.

Exemple. Considérons la série $\sum\limits_{n\geqslant 1} \dfrac{1}{n}$, appelée *série harmonique*. Cette série diverge, bien que son terme général tende vers 0. Supposons en effet qu'elle converge, et soit $\varepsilon = 1/3$.

D'après le théorème précédent, il existe un entier $N \geqslant 1$ tel que, pour tous entiers p, q vérifiant $N \leqslant p \leqslant q$, on ait :

$$\frac{1}{p} + \frac{1}{p+1} + \cdots + \frac{1}{q} \leqslant \frac{1}{3}.$$

Prenons en particulier $p := N+1$ et $q := 2N$. Alors :

$$\frac{1}{N+1} + \frac{1}{N+2} + \cdots + \frac{1}{2N} \leqslant \frac{1}{3}.$$

Le membre de gauche est somme de N réels, rangés par ordre décroissant, il est donc minoré par N fois le plus petit de ces termes, à savoir $\frac{1}{2N}$, c'est-à-dire par $1/2$.

D'où $\frac{1}{2} \leqslant \frac{1}{3}$, ce qui est absurde.

L'étude d'une suite peut se ramener à l'étude d'une série :

Théorème 7. Pour qu'une suite (a_n) converge, il faut, et il suffit, que la série $\sum_{n \geqslant 0} (a_{n+1} - a_n)$ converge. Lorsque c'est le cas, on a :

$$\lim_{n \to +\infty} a_n = a_0 + \sum_{n=0}^{+\infty} (a_{n+1} - a_n).$$

Démonstration. Posons $u_n := a_{n+1} - a_n$ pour tout n.
Les sommes partielles U_n de $\sum_{n \geqslant 0} u_n$ sont données par :

$$U_n = (a_1 - a_0) + (a_2 - a_1) + (a_3 - a_2) + \cdots + (a_{n+1} - a_n) = a_{n+1} - a_0.$$

Ainsi la série $\sum_{n \geqslant 0} u_n$ converge si, et seulement si, la suite $(a_{n+1} - a_0)$ converge, *i.e.* si (a_n) converge, auquel cas l'égalité de l'énoncé est claire. ■

Exemple. Prenons $a_n := 1/(n+1)$. La suite (a_n) converge vers 0.
Ici $u_n = \frac{1}{n+2} - \frac{1}{n+1} = \frac{-1}{(n+1)(n+2)}$; on retrouve l'exemple 3 de la page 715.

2 Séries à termes réels positifs

2.1 Convergence par comparaison

Les séries à termes réels positifs sont plus simples à étudier que les autres, à cause des deux résultats évidents suivants :

Proposition 8. Soit $\sum_{n \geqslant 0} u_n$ une série à termes réels positifs. La suite (U_n) des sommes partielles de cette série est croissante.

Démonstration. Pour tout entier $n \geqslant 1$, la formule (12) montre que $U_n - U_{n-1} = u_n \geqslant 0$, d'où la conclusion. ■

Théorème 9. Pour qu'une série $\sum_{n\geqslant 0} u_n$ à termes réels positifs converge , il faut et il suffit que la suite (U_n) de ses sommes partielles soit *majorée*, *i.e.* qu'il existe un réel M tel que :

$$U_n \;=\; u_0 + u_1 + \cdots + u_n \;\leqslant\; M \quad \text{pour tout} \quad n \geqslant 0.$$

Si cette condition est vérifiée, la somme de la série $\sum_{n\geqslant 0} u_n$ est la *borne supérieure* de $\{U_n \mid n \geqslant 0\}$.

Démonstration. Puisque la suite (U_n) est croissante, elle converge si, et seulement si, elle est majorée : la série $\sum_{n\geqslant 0} u_n$ converge si, et seulement si, (U_n) est majorée. La dernière assertion vient de la preuve du théorème 44 de la page 556. ■

Si une série $\sum_{n\geqslant 0} u_n$ à termes réels positifs diverge, la suite (U_n) de ses sommes partielles diverge. Cette suite étant croissante, elle a comme limite $+\infty$, et l'on écrit $\sum_{n=0}^{+\infty} u_n = +\infty$. Ainsi $\sum_{n=1}^{+\infty} \frac{1}{n} = +\infty$. Voici une application importante du théorème précédent :

Théorème 10 (Comparaison). Soient $\sum_{n\geqslant 0} a_n$ et $\sum_{n\geqslant 0} b_n$ deux séries à termes réels positifs. On suppose que l'on a $a_n \leqslant b_n$ pour tout $n \geqslant 0$.

1. Si la série $\sum_{n\geqslant 0} b_n$ converge, alors la série $\sum_{n\geqslant 0} a_n$ converge aussi, et l'on a :
$$\sum_{n=0}^{+\infty} a_n \;\leqslant\; \sum_{n=0}^{+\infty} b_n.$$
2. Si la série $\sum_{n\geqslant 0} a_n$ diverge, alors la série $\sum_{n\geqslant 0} b_n$ diverge aussi.

Supposons maintenant qu'il existe deux nombres réels m, M, avec $0 < m \leqslant M$, tels que l'on ait, à partir d'un certain rang n_0 :

$$m a_n \;\leqslant\; b_n \;\leqslant\; M a_n.$$

Alors les séries $\sum_{n\geqslant 0} a_n$ et $\sum_{n\geqslant 0} b_n$ sont de même nature.

Démonstration.

Notons A_n, B_n les sommes partielles des séries $\sum_{n\geqslant 0} a_n$ et $\sum_{n\geqslant 0} b_n$ respectivement.

Supposons que la série $\sum_{n\geqslant 0} b_n$ converge. D'après le théorème 9, il existe un réel M majorant tous les B_n. Or, pour tout n :

$$A_n = a_0 + a_1 + \cdots + a_n \leqslant b_0 + b_1 + \cdots + b_n = B_n.$$

Ainsi M majore tous les A_n, donc $\sum_{n\geqslant 0} a_n$ converge, vu le théorème précédent.

En faisant tendre n vers $+\infty$ dans l'inégalité $A_n \leqslant B_n$, nous obtenons $\sum_{n=0}^{+\infty} a_n \leqslant \sum_{n=0}^{+\infty} b_n$. D'où l'assertion 1, puis l'assertion 2, par contraposition.
Pour la dernière assertion, supposons que la série $\sum_{n\geqslant 0} a_n$ converge. Alors $\sum_{n\geqslant 0} Ma_n$ converge ; comme $b_n \leqslant Ma_n$ dès que $n \geqslant n_0$, $\sum_{n\geqslant n_0} b_n$ converge, d'après 1, donc la série $\sum_{n\geqslant 0} b_n$ converge.
Supposons que la série $\sum_{n\geqslant 0} a_n$ diverge. Dans ce cas, $\sum_{n\geqslant 0} ma_n$ diverge. Or $b_n \geqslant ma_n$ dès que $n \geqslant n_0$, donc la série $\sum_{n\geqslant 0} b_n$ diverge, d'après 2. ■

Théorème 11 (Équivalence). Soient $\sum_{n\geqslant 0} a_n$ et $\sum_{n\geqslant 0} a'_n$ deux séries à termes réels strictement positifs, on suppose que a'_n est équivalent à a_n lorsque n tend vers $+\infty$, c'est-à-dire que a'_n/a_n tend vers 1 lorsque n tend vers $+\infty$.

1. Les deux séries sont de même nature.
2. On suppose que les deux séries convergent. En notant R_n, R'_n les restes d'indice n respectifs des deux séries, R'_n est équivalent à R_n lorsque n tend vers $+\infty$.
3. On suppose que les deux séries divergent. En notant A_n, A'_n les sommes partielles respectives des deux séries, A'_n est équivalent à A_n lorsque n tend vers $+\infty$.

Démonstration. 1. Par hypothèse, la suite (a'_n/a_n) converge vers 1. Soit $\varepsilon \in]0,1[$. À partir d'un certain rang $N \geqslant 1$, $a'_n/a_n \in [1-\varepsilon, 1+\varepsilon]$, c'est-à-dire $(1-\varepsilon)a_n \leqslant a'_n \leqslant (1+\varepsilon)a_n$. D'après la dernière assertion du théorème 10 de la page précédente, les deux séries sont de même nature.
2. Supposons que les deux séries convergent. Si ε et N sont comme ci-dessus, et si $n \geqslant N$, on a $(1-\varepsilon)a_k \leqslant a'_k \leqslant (1+\varepsilon)a_k$ pour tout entier $k > N$, d'où par sommation $(1-\varepsilon)R_n \leqslant R'_n \leqslant (1+\varepsilon)R_n$. Autrement dit $R'_n/R_n \in [1-\varepsilon, 1+\varepsilon]$ dès que $n \geqslant N$. Cela montre que la suite (R'_n/R_n) converge vers 1.
3. Supposons que les deux séries divergent : les suites (A_n) et (A'_n) ont pour limite $+\infty$. Soient ε, N comme ci-dessus et $n > N$.
En sommant les inégalités $(1-\varepsilon)a_k \leqslant a'_k \leqslant (1+\varepsilon)a_k$ pour $k = N+1, \ldots, n$, il vient :

$$(1-\varepsilon)\big[A_n - A_N\big] \leqslant A'_n - A'_N \leqslant (1+\varepsilon)\big[A_n - A_N\big] \leqslant (1+\varepsilon)A_n\,, \text{ d'où}$$

$$1-\varepsilon+\frac{A'_N-(1-\varepsilon)A_N}{A_n} \leqslant \frac{A'_n}{A_n} \leqslant 1+\varepsilon+\frac{A'_N}{A_n}\,.$$

Mais (A_n) a pour limite $+\infty$, et ε, N sont fixés, donc il existe un entier $N' > N$ tel que, pour tout $n \geqslant N'$, on ait $1-2\varepsilon \leqslant A'_n/A_n \leqslant 1+2\varepsilon$. On en déduit que la suite (A'_n/A_n) converge vers 1. ■

Exemple. La série de terme général $u_n := \sin(\pi/2^n)$ converge. En effet, $\sin x/x$ tend vers 1 quand le réel x tend vers 0 (pourquoi ?), donc u_n est équivalent à $\pi/2^n$, terme général d'une série géométrique convergente, car sa raison est $1/2$ (on a $u_n > 0$ dès que $n \geqslant 1$).

Voici un autre résultat important sur les séries à termes réels positifs :

Théorème 12 (Règle de d'Alembert). Soit $\sum\limits_{n\geqslant 0} u_n$ une série à termes réels strictement positifs. Supposons que, lorsque n tend vers $+\infty$, le quotient u_{n+1}/u_n ait une limite ℓ, finie ou égale à $+\infty$. Alors :

1. Si $\ell < 1$, la série converge.
2. Si $\ell > 1$, la série diverge.
3. Si $\ell = 1$, on ne peut rien dire *a priori*.

Démonstration. Supposons $\ell < 1$, et soit $r := (\ell+1)/2 \in \,]\ell, 1[$. À partir d'un certain rang m, on a $u_{n+1}/u_n \leqslant r$, soit $u_{n+1} \leqslant r u_n$. On en déduit, par récurrence sur n, que $u_n \leqslant a r^n$ pour tout $n \geqslant m$, où $a := u_m r^{-m}$. Or la série géométrique $\sum\limits_{n\geqslant m} a r^n$ converge, puisque $0 < r < 1$ (théorème 1 de la page 716). D'après le théorème de comparaison, $\sum\limits_{n\geqslant m} u_n$ converge, donc $\sum\limits_{n\geqslant 0} u_n$ converge.

Supposons $\ell > 1$, avec éventuellement $\ell = +\infty$. À partir d'un certain rang m, on a $u_{n+1}/u_n \geqslant 1$, soit $u_{n+1} \geqslant u_n$. On en déduit que $u_n \geqslant u_m > 0$ pour tout $n \geqslant m$. Il en résulte que la suite (u_n) ne converge pas vers 0, donc la série $\sum\limits_{n\geqslant m} u_n$ diverge, en vertu du théorème 5 de la page 719. ■

Exercice 2.

Étudier la série suivante :

$$\sum_{n\geqslant 0} \frac{1}{n!} = 1 + \frac{1}{1} + \frac{1}{2} + \frac{1}{6} + \cdots$$

Solution. Notons u_n son terme général. Pour tout n, on a $u_n = \dfrac{1}{n!}$ et $u_{n+1} = \dfrac{1}{(n+1)!}$, d'où $\dfrac{u_{n+1}}{u_n} = \dfrac{1}{n+1}\cdot$

Ainsi $\dfrac{u_{n+1}}{u_n}$ tend vers 0 lorsque n tend vers $+\infty$, donc la série donnée converge, en vertu de la règle de d'Alembert. Nous verrons plus tard que la somme de cette série est e, base des logarithmes népériens.

Exercice 3.

Étudier la série suivante :

$$\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} + \cdots = \sum_{n\geqslant 1} \frac{1}{n^2}\cdot \tag{14}$$

Solution. Posons $u_n := 1/n^2$. Si $n \geqslant 1$, $u_n = \dfrac{1}{n^2}$ et $u_{n+1} = \dfrac{1}{(n+1)^2}$, d'où :

$$\frac{u_{n+1}}{u_n} = \frac{n^2}{(n+1)^2} = \left(1+\frac{1}{n}\right)^{-2}.$$

Il en résulte que u_{n+1}/u_n tend vers 1 lorsque n tend vers $+\infty$. La règle de d'Alembert ne permet pas de conclure, quant à la nature de la série donnée. Par contre $u_n \sim 1/\big(n(n+1)\big)$, donc $\sum\limits_{n\geqslant 1} u_n$ converge, à cause de l'exemple 3 de la page 715 et du théorème d'équivalence.

2.2 Utilisation d'une intégrale

Il peut être fructueux, dans l'étude d'une série, de faire intervenir des intégrales appropriées. Voici un type d'encadrement utile dans ce contexte :

Proposition 13. Soient f une fonction réelle continue et décroissante sur un intervalle I et $p, q \in \mathbb{Z}$, $p \leqslant q$. Si p et $q+1$ appartiennent à I, on a :

$$f(p+1)+f(p+2)+\cdots+f(q+1) \leqslant \int_p^{q+1} f(x)\mathrm{d}x \leqslant f(p)+f(p+1)+\cdots+f(q). \quad (15)$$

Démonstration. Soit $k \in [\![p, q]\!]$. Puisque f est décroissante, on a pour tout $x \in [k, k+1]$, $f(k+1) \leqslant f(x) \leqslant f(k)$. En intégrant ces inégalités sur $[k, k+1]$, il vient :

$$f(k+1) = \int_k^{k+1} f(k+1)\mathrm{d}x \leqslant \int_k^{k+1} f(x)\mathrm{d}x \leqslant \int_k^{k+1} f(k)\mathrm{d}x = f(k).$$

Sommons ces inégalités pour $k = p, p+1, \ldots, q$. D'après la relation de Chasles, la somme des intégrales $\displaystyle\int_k^{k+1} f(x)\mathrm{d}x$ vaut $\displaystyle\int_p^{q+1} f(x)\mathrm{d}x$. On en déduit aussitôt les inégalités annoncées. ∎

Voici un exemple d'application de la proposition précédente :

Théorème 14 (Séries de Riemann).
Soit a un nombre réel fixé. Considérons la série suivante, appelée *série de Riemann* :

$$\sum_{n\geqslant 1} \frac{1}{n^a} = \sum_{n\geqslant 1} n^{-a}. \quad (16)$$

Cette série converge si $a > 1$ et diverge si $a \leqslant 1$.

Démonstration. Si $a = 1$, c'est la série harmonique, qui diverge. Si $a < 1$, $n^a \leqslant n$, donc $\dfrac{1}{n^a} \geqslant \dfrac{1}{n}$ pour tout n. Par comparaison, la série (16) diverge.

Supposons $a > 1$. Considérons la fonction $f : x \mapsto \dfrac{1}{x^a}$, qui est continue et strictement décroissante sur $]0, +\infty[$. Soit n un entier supérieur ou égal à 2. Appliquons la proposition 13 de la page précédente, en prenant $p := 1$ et $q := n - 1$. La première inégalité de la proposition donne :

$$\begin{aligned}\frac{1}{2^a} + \frac{1}{3^a} + \cdots + \frac{1}{n^a} &\leqslant \int_1^n \frac{\mathrm{d}x}{x^a} = \int_1^n x^{-a}\mathrm{d}x \\ &= \left[\frac{x^{-a+1}}{-a+1}\right]_{x=1}^{x=n} = \frac{1}{a-1}\left[1 - n^{-a+1}\right] \leqslant \frac{1}{a-1}\cdot\end{aligned}$$

Il en résulte que la suite des sommes partielles de la série (16) est majorée par $1 + \dfrac{1}{a-1}$, donc cette série converge, d'après le théorème 9 de la page 721. ■

Par cet exemple, nous retrouvons le fait que la série (14) converge. On peut démontrer que la somme de cette série est égale à $\pi^2/6$, ce qui n'est pas évident (voir, par exemple, l'exercice II.6.19 de la page 328).
Dans certains cas, les séries de Riemann servent de tests pour déterminer la nature d'une série à termes réels positifs, à cause du théorème suivant :

Théorème 15. Considérons une série $\sum\limits_{n\geqslant 0} u_n$, à termes réels positifs.

1. Supposons qu'il existe un réel $a > 1$ tel que la suite $(n^a u_n)$ soit *majorée*. Alors la série donnée converge.
2. Supposons qu'il existe un réel $a \leqslant 1$ tel que la suite $(n^a u_n)_{n\geqslant 1}$ soit *minorée* par un nombre $m > 0$. Alors la série donnée diverge.

Démonstration. Pour 1, soit M un réel tel que $n^a u_n \leqslant M$, donc $u_n \leqslant \frac{M}{n^a}$, pour tout $n \geqslant 1$. Puisque $a > 1$, la série $\sum\limits_{n\geqslant 1} \frac{M}{n^a}$ converge, donc $\sum\limits_{n\geqslant 0} u_n$ converge, par comparaison.
Pour 2, on a de même $u_n \geqslant \frac{m}{n^a}$ pour tout $n \geqslant 1$. Puisque $a \leqslant 1$ et $m > 0$, la série $\sum\limits_{n\geqslant 1} \frac{m}{n^a}$ diverge, donc $\sum\limits_{n\geqslant 0} u_n$ diverge, par comparaison. ■

Exercice 4.

Soit a un réel fixé. Étudier la série suivante :

$$\sum_{n\geqslant 1} u_n \quad \text{où} \quad u_n := \frac{\ln n}{n^a}\cdot$$

Solution. Montrons que cette série diverge si $a \leqslant 1$ et converge si $a > 1$.
Supposons d'abord $a \leqslant 1$. Pour tout $n \geqslant 3$, on a $\ln n \geqslant 1$, d'où $u_n \geqslant n^{-a}$. D'après le théorème 14 de la page ci-contre, la série de terme général n^{-a} diverge (car $a \leqslant 1$), donc la série donnée diverge, par comparaison.

Supposons $a > 1$. La conclusion est ici moins directe. On sait que, lorsque x tend vers $+\infty$, $\ln x/x$ tend vers 0. Si $b > 0$ est un réel fixé et n est un entier tendant vers $+\infty$, $\ln(n^b)/n^b$ tend donc vers 0, *i.e.* $\ln n/n^b$ tend vers 0. Cela étant, on a :

$$n^{a-b}u_n \;=\; n^{a-b}\frac{\ln n}{n^a} \;=\; \frac{\ln n}{n^b}\cdot$$

Choisissons $b := (a-1)/2$. On a $b > 0$ et $a-b = (a+1)/2 > 1$, car $a > 1$. Vu les égalités précédentes, $n^{a-b}u_n$ tend vers 0 quand n tend vers $+\infty$, en particulier la suite $(n^{a-b}u_n)$ est majorée. D'après l'assertion 1 du théorème 15 de la page précédente, la série $\sum\limits_{n\geqslant 1} u_n$ est donc convergente.

Exercice 5.

Pour tout entier $n \geqslant 1$, posons :

$$a_n \;:=\; \frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n}-\ln(n+1).$$

Montrer que la suite (a_n) converge, à l'aide du théorème 7 de la page 720.

Solution. Posons : $u_n \;:=\; a_{n+1}-a_n$. Il s'agit de montrer que la série de terme général u_n converge. Commençons par établir les inégalités suivantes :

$$x-\frac{x^2}{2} \;\leqslant\; \ln(1+x) \;\leqslant\; x \quad \text{pour tout} \quad x\in[0,+\infty[. \tag{17}$$

Considérons les deux fonctions f, g définies ainsi sur $]-1,+\infty[$:

$$f(x) \;:=\; x-\ln(1+x)\;,\;\; g(x) \;:=\; \frac{x^2}{2}+\ln(1+x)-x.$$

Ces fonctions sont dérivables, un calcul immédiat donne leurs dérivées :

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

Il en résulte que f et g sont croissantes sur $[0,+\infty[$. Comme f et g sont nulles en 0, on a $f(x)\geqslant 0$ et $g(x)\geqslant 0$ pour tout $x\in[0,+\infty[$, d'où les inégalités (17).

Revenons à la série de terme général u_n. Pour tout $n\geqslant 1$, il vient :

$$\begin{aligned} u_n \;&=\; \left(\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n+1}-\ln(n+2)\right)-\left(\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n}-\ln(n+1)\right)\\ &=\; \frac{1}{n+1}-\ln(n+2)+\ln(n+1) \;=\; \frac{1}{n+1}-\ln\left(\frac{n+2}{n+1}\right)\\ &=\; \frac{1}{n+1}-\ln\left(1+\frac{1}{n+1}\right) \;=\; x-\ln(1+x), \quad \text{où} \quad x := \frac{1}{n+1}\cdot \end{aligned}$$

Les inégalités (17) donnent $0\leqslant x-\ln(1+x)\leqslant \dfrac{x^2}{2}$, d'où :

$$0 \;\leqslant\; u_n \;\leqslant\; \frac{1}{2(n+1)^2} \;\leqslant\; \frac{1}{n^2}\cdot$$

Ainsi $\sum_{n\geqslant 1} u_n$ est une série à termes réels positifs, et $n^2 u_n \leqslant 1$ pour tout $n \geqslant 1$. D'après le théorème 15 de la page 725, cette série est donc convergente, *i.e.* la *suite* (a_n) converge. La limite C de cette suite est appelée *constante d'Euler*, en voici une valeur approchée par défaut : $0{,}577215$.

Lorsque n tend vers $+\infty$, $\ln(n+1) - \ln n = \ln\left(1+\dfrac{1}{n}\right)$ tend vers 0. Posons :

$$\varepsilon_n := \left(\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n}-\ln n\right) - C\,.$$

La suite (ε_n) converge donc vers 0, ce qui prouve la formule suivante :

$$1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n} = \ln n + C + \varepsilon_n\,, \quad \text{où} \quad \lim_{n\to+\infty} \varepsilon_n = 0. \tag{18}$$

2.3 Application : développement d'un réel positif

L'écriture décimale d'un réel positif a été définie lors de la construction des réels (voir la définition 31 de la page 588) ; par exemple :

$$\pi = 3{,}14159\ldots = 3 + \frac{1}{10} + \frac{4}{10^2} + \frac{1}{10^3} + \frac{5}{10^4} + \frac{9}{10^5} + \cdots$$

conduit naturellement à l'étude de séries de terme général $\beta_n\, 10^{-n}$, où β_0 appartient à $\mathbb{N}$ et les β_n, pour $n \geqslant 1$, sont des *chiffres*, c'est-à-dire des entiers compris entre 0 et 9.
Une telle série converge puisque son terme général positif est majoré par 9.10^{-n}, terme général d'une série géométrique convergente. Malheureusement, une telle écriture n'est pas unique, puisque l'on a, par exemple $1{,}000\ldots = 0{,}999\ldots$
Pour éviter ce problème, on se limite aux suites (β_n) qui ne stationnent pas à 9.

Pour les réels négatifs, la tradition veut que l'on adopte l'écriture décimale de leur valeur absolue précédée du signe $-$ (par exemple $-\sqrt{2} = -1{,}1414\ldots$). Traitons plus généralement le développement des réels dans une base b quelconque : b est un entier naturel supérieur ou égal à 2.
Nous notons donc $\mathcal{D}$ l'ensemble des *développements en base* b, c'est-à-dire :

$$\mathcal{D} = \left\{(\beta_n) \in \mathbb{N}^{\mathbb{N}} \;\middle|\; \forall n \geqslant 1\,,\ 0 \leqslant \beta_n \leqslant b-1\right\}$$

et $\mathcal{D}_p$ l'ensemble des *développements en base* b *propres*, c'est-à-dire des éléments de $\mathcal{D}$ qui prennent une infinité de fois des valeurs différentes de $b-1$.
On remarquera qu'il n'y a aucune condition sur l'entier naturel β_0.
De la même façon que dans le cas de la base 10, pour tout $(\beta_n) \in \mathcal{D}$, la série $\sum_{p\geqslant 0} \beta_p\, b^{-p}$ converge puisque son terme général (à partir du rang 1) est positif et majoré par $(b-1)\, b^{-p}$ qui est le terme général d'une série géométrique convergente.

Lemme 16. Soit $(\beta_n) \in \mathcal{D}_p$; on note $(a_n\, b^{-n})$ la suite des sommes partielles de la série de terme général $\beta_n\, b^{-n}$:

$$\forall n \in \mathbb{N}\,,\ a_n = b^n \sum_{p=0}^{n} \frac{\beta_p}{b^p}.$$

$$\text{Alors :} \qquad \forall n \in \mathbb{N}\,,\ \frac{a_n}{b^n} \leqslant \sum_{p=0}^{+\infty} \frac{\beta_p}{b^p} < \frac{1+a_n}{b^n}.$$

Démonstration. L'inégalité de gauche est immédiate puisque la série est à terme général positif. Pour l'autre, on a :

$$\sum_{p=n+1}^{+\infty} \frac{\beta_p}{b^p} \leqslant \sum_{p=n+1}^{+\infty} \frac{b-1}{b^p} = \frac{1}{b^n},$$

l'inégalité étant stricte, puisque les β_p pour $p > n$ ne sont pas constamment égaux à $b-1$. ■

Proposition 17. L'application $\delta \ :\ (\beta_n) \mapsto \displaystyle\sum_{n=0}^{+\infty} \frac{\beta_n}{b^n}$ est une application de $\mathcal{D}$ dans $\mathbb{R}_+$ dont la restriction à $\mathcal{D}_p$ est injective.

Démonstration. Nous savons déjà que δ est bien définie sur $\mathcal{D}$ et à valeurs dans $\mathbb{R}_+$. Pour prouver que sa restriction à $\mathcal{D}_p$ est injective, nous allons montrer qu'elle est strictement croissante de $(\mathcal{D}_p, \preceq)$ dans $(\mathbb{R}_+, \leqslant)$, où $\preceq$ est *l'ordre lexicographique* défini par :

$$(\beta_n) \preceq (\beta'_n) \Longleftrightarrow \begin{cases} \forall n \in \mathbb{N}\,,\ \beta_n = \beta'_n \\ \text{ou} \\ \exists n_0 \in \mathbb{N}\ :\ \forall n < n_0\,,\ \beta_n = \beta'_n \qquad \text{et} \qquad \beta_{n_0} < \beta'_{n_0}. \end{cases}$$

Soient donc $\beta = (\beta_n)$ et $\beta' = (\beta'_n)$ deux éléments distincts de $\mathcal{D}_p$ tels que $\beta \preceq \beta'$. Prenons l'entier n_0 (éventuellement nul) défini par :

$$\forall n < n_0\,,\ \beta_n = \beta'_n \qquad \text{et} \qquad \beta_{n_0} < \beta'_{n_0}.$$

On a donc plus précisément $\beta_{n_0} + 1 \leqslant \beta'_{n_0}$ (ce sont des entiers) et, en notant $(a_n\, b^{-n})$ et $(a'_n\, b^{-n})$ les suites des sommes partielles respectivement des séries $\sum\limits_{n \geqslant 0} \beta_n\, b^{-n}$ et $\sum\limits_{n \geqslant 0} \beta'_n\, b^{-n}$:

$$\delta(\beta) < \frac{1+a_{n_0}}{b^{n_0}} \leqslant \frac{a'_{n_0}}{b^{n_0}} \leqslant \delta(\beta').$$ ■

Remarque. Soit $\beta = (\beta_n) \in \mathcal{D}_p$.

En notant $x = \delta(\beta)$, l'encadrement $a_n \leqslant b^n\, x < 1 + a_n$, avec $a_n \in \mathbb{N}$, montre que a_n est la partie entière de $b^n\, x$. Cela justifie ce qui suit pour prouver la surjectivité de $\delta_{|\mathcal{D}_p}$.

Soit $x \in \mathbb{R}_+$. Pour tout $n \in \mathbb{N}$, on note a_n la partie entière de $b^n\, x$ de telle sorte que l'on ait $a_n \leqslant b^n\, x < 1 + a_n$. De plus, on pose :

$$\beta_0 = a_0 \qquad \text{et} \qquad \forall n \in \mathbb{N}\,,\ \beta_{n+1} = a_{n+1} - b a_n.$$

Proposition 18.

- Les suites $\left(\dfrac{a_n}{b^n}\right)$ et $\left(\dfrac{1+a_n}{b^n}\right)$ sont adjacentes et convergent vers x.
- La suite (β_n) est un développement en base b propre.
- $x = \displaystyle\sum_{n=0}^{+\infty} \frac{\beta_n}{b^n}$.

Démonstration. Par définition :

- a_{n+1} est le plus grand entier inférieur ou égal à $b^{n+1}\,x$,
- $1 + a_{n+1}$ est le plus petit entier strictement supérieur à $b^{n+1}\,x$.

De la double inégalité :

$$ba_n \leqslant b^{n+1}\,x < ba_n + b$$

on déduit $ba_n \leqslant a_{n+1}$ et $ba_n + b \geqslant 1 + a_{n+1}$. On a donc $0 \leqslant \beta_{n+1} \leqslant b-1$ ainsi que la croissance de la suite $\left(\dfrac{a_n}{b^n}\right)$ et la décroissance de la suite $\left(\dfrac{1+a_n}{b^n}\right)$.

Ces deux suites sont alors immédiatement adjacentes, et la double inégalité :

$$a_n\,b^{-n} \leqslant x < (1+a_n)\,b^{-n} \qquad (*)$$

montre que leur limite commune est x. Pour tout $n \in \mathbb{N}$, la somme partielle d'indice n de la série $\sum\limits_{n\geqslant 0} \beta_n\,b^{-n}$ est $a_n b^{-n}$ (le vérifier), et l'on en déduit le dernier point de la proposition.

Enfin, si le développement (β_n) n'était pas propre, on aurait à partir d'un certain rang $a_{n+1} - ba_n = \beta_{n+1} = b-1$ et donc la suite $(1+a_n)\,b^{-n}$ serait stationnaire, ce qui contredirait l'inégalité stricte dans l'encadrement $(*)$. ∎

Exemple. Dans le cas des développements décimaux ($b = 10$), les nombres $a_n\,10^{-n}$ et $(1+a_n)\,10^{-n}$ qui encadrent x s'appellent respectivement les *approximations décimales de x par défaut et par excès à 10^{-n} près*.

On en déduit la bijectivité de la restriction à $\mathcal{D}_p$ de la fonction δ :

Théorème 19. Tout réel positif x s'écrit de manière unique sous la forme :

$$x = \beta_0, \beta_1\beta_2\ldots = \sum_{n=0}^{+\infty} \frac{\beta_n}{b^n}$$

où (β_n) est une suite d'entiers naturels tels que $\forall n \in \mathbb{N}^*$, $0 \leqslant \beta_n \leqslant b-1$ et qui prend une infinité de fois des valeurs différentes de $b-1$.

Remarque. La démonstration de la proposition 17 de la page précédente fournit une méthode pratique pour comparer deux réels positifs : leur position relative est donnée par la comparaison des premiers chiffres qui diffèrent dans leurs développements en base b.

3 Séries à termes réels ou complexes

3.1 Convergence absolue

Les séries à termes réels positifs interviennent à propos de séries à termes quelconques. Le théorème ci-après illustre ce fait.

Définition 5. Soit $\sum_{n\geqslant 0} u_n$ une série à termes réels ou complexes. On dit que cette série est *absolument convergente* si la série des modules $\sum_{n\geqslant 0} |u_n|$ (qui est à termes réels positifs) converge.

Théorème 20. Toute série absolument convergente est convergente.

Démonstration. Soit donc $\sum_{n\geqslant 0} u_n$ une série absolument convergente. Pour tout n, posons $a_n := \operatorname{Re} u_n$ et $b_n := \operatorname{Im} u_n$. D'après le théorème 4 de la page 718, il suffit de prouver que les deux séries $\sum_{n\geqslant 0} a_n$ et $\sum_{n\geqslant 0} b_n$ convergent. Pour tout n, on a $|u_n|^2 = a_n^2 + b_n^2$, d'où $a_n^2 \leqslant |u_n|^2$, c'est-à-dire $|a_n| \leqslant |u_n|$, de même $|b_n| \leqslant |u_n|$. Puisque $\sum_{n\geqslant 0} |u_n|$ converge, le théorème de comparaison montre que $\sum_{n\geqslant 0} |a_n|$ converge. Autrement dit, la série $\sum_{n\geqslant 0} a_n$ converge absolument. Nous sommes donc ramenés au cas de séries à termes réels.

Partons d'une remarque simple. Si $x \in \mathbb{R}$, on a $\max(0,x) + \max(0,-x) = |x|$ et $\max(0,x) - \max(0,-x) = x$. Pour tout n, posons :

$$x_n := \max(0, a_n) \qquad \text{et} \qquad y_n := \max(0, -a_n),$$

de sorte que $x_n + y_n = |a_n|$, donc $0 \leqslant x_n \leqslant |a_n|$ et $0 \leqslant y_n \leqslant |a_n|$. Les séries $\sum_{n\geqslant 0} x_n$ et $\sum_{n\geqslant 0} y_n$ sont à termes réels positifs. Comme $\sum_{n\geqslant 0} |a_n|$ converge, le théorème de comparaison montre que $\sum_{n\geqslant 0} x_n$ et $\sum_{n\geqslant 0} y_n$ convergent. Alors $\sum_{n\geqslant 0} (x_n - y_n)$ converge, *i.e.* $\sum_{n\geqslant 0} a_n$ converge. De même $\sum_{n\geqslant 0} b_n$ converge. ∎

La réciproque de ce théorème est fausse : *une série convergente n'est pas forcément absolument convergente*, nous le verrons. D'où une définition :

Définition 6. Une série est dite *semi-convergente* si elle est convergente mais pas absolument convergente.

Une série convergente est absolument convergente ou semi-convergente.

Le théorème suivant servira à l'étude de l'exponentielle complexe (voir le module II.5) et il sera important pour l'étude des séries entières en L2.

Théorème et définition 21.

1. Soient $(a_n)_{n\geqslant 0}$ et $(b_n)_{n\geqslant 0}$ deux suites complexes. Pour tout $n \in \mathbb{N}$, posons :
$$c_n = \sum_{h=0}^{n} a_h b_{n-h}\,. \tag{19}$$
On dit que la suite (c_n) est le *produit de Cauchy* des suites (a_n) et (b_n).
2. Supposons que les deux séries $\sum\limits_{n\geqslant 0} a_n$ et $\sum\limits_{n\geqslant 0} b_n$ soient emphabsolument convergentes. Alors la série $\sum\limits_{n\geqslant 0} c_n$ est *absolument convergente*, et l'on a :
$$\left(\sum_{n=0}^{+\infty} a_n\right)\left(\sum_{n=0}^{+\infty} b_n\right) = \sum_{n=0}^{+\infty} c_n\,. \tag{20}$$

Démonstration. Notons $A_n, A'_n, B_n, B'_n, C_n, C'_n$ les sommes partielles des séries respectives $\sum\limits_{n\geqslant 0} a_n$, $\sum\limits_{n\geqslant 0} |a_n|$, ... Soit $n \in \mathbb{N}$. L'inégalité triangulaire donne :
$$|c_n| = \left|\sum_{h=0}^{n} a_h b_{n-h}\right| \leqslant \sum_{h=0}^{n} |a_h|\,|b_{n-h}|\,.$$
Soit ensuite $N \in \mathbb{N}$. Sommons les inégalités précédentes pour $n = 0, \ldots, N$:
$$C'_N = \sum_{n=0}^{N} |c_n| \leqslant \sum_{n=0}^{N}\left(\sum_{h=0}^{n} |a_h|\,|b_{n-h}|\right).$$
Le membre de droite est la somme des $|a_h|\,|b_k|$ pour tous les $(h,k) \in [\![0,N]\!]^2$ tels que $h+k \leqslant N$. Cette somme est majorée par la somme des $|a_h|\,|b_k|$ pour *tous* les $(h,k) \in [\![0,N]\!]^2$, *i.e.* par le produit de $A'_N = \sum\limits_{h=0}^{N} |a_h|$ et de $B'_N = \sum\limits_{k=0}^{N} |b_k|$. Ainsi $C'_N \leqslant A'_N B'_N$ pour tout N.

Puisque $\sum\limits_{n\geqslant 0} |a_n|$ converge, la suite (A'_N) est majorée (théorème 9 de la page 721), de même (B'_N) est majorée, donc (C'_N) est majorée, et alors $\sum\limits_{n\geqslant 0} |c_n|$ converge. Ainsi $\sum\limits_{n\geqslant 0} c_n$ est absolument convergente, donc convergente.

Il reste à montrer que $C = AB$, en notant A la somme de la série $\sum\limits_{n\geqslant 0} a_n$, etc. Pour tout n, notons R_n (resp. R'_n) le reste d'indice n de $\sum\limits_{n\geqslant 0} |a_n|$ (resp. $\sum\limits_{n\geqslant 0} |b_n|$). Il suffit de montrer que, lorsque N tend vers $+\infty$, la différence $D_N = A_{2N}B_{2N} - C_{2N}$ tend vers 0. Comme ci-dessus, D_N est la somme des $a_h b_k$ pour tous les couples $(h,k) \in [\![0,2N]\!]^2$ tels que $h+k > 2N$, donc $|D_N|$ est majoré par la somme S_N des $|a_h||b_k|$ pour tous les couples (h,k) en question. On peut écrire $S_N = S'_N + S''_N$, la somme S'_N (resp. S''_N) correspondant aux couples (h,k) tels que $h \leqslant N$ (resp. $h > N$). Les suites (R_n) et (R'_n) étant décroissantes, on obtient les majorations :
$$0 \leqslant S'_N = \sum_{h=0}^{N} |a_h|\left(\sum_{k=2N-h+1}^{2N} |b_k|\right) \leqslant \sum_{h=0}^{N} |a_h| R'_{2N-h} \leqslant \sum_{h=0}^{N} |a_h| R'_N \leqslant R'_N \sum_{h=0}^{+\infty} |a_h|\,,$$
$$0 \leqslant S''_N \leqslant \sum_{k=0}^{2N} |b_k|\left(\sum_{h=N+1}^{2N} |a_h|\right) \leqslant R_N \sum_{k=0}^{2N} |b_k| \leqslant R_N \sum_{k=0}^{+\infty} |b_k|\,.$$
Vu la définition 4 de la page 717, les suites $(R_N)_{N\geqslant 0}$ et $(R'_N)_{N\geqslant 0}$ convergent vers 0.

Il en résulte que (S'_N) et (S''_N), puis (S_N), convergent vers 0, et ainsi (D_N) converge vers 0. ■

3.2 Séries alternées

On appelle *série alternée* toute série à termes *réels* dont les termes sont alternativement positifs et négatifs. Le terme général d'une telle série peut donc s'écrire $u_n = (-1)^n a_n$ ou $u_n = (-1)^{n+1} a_n$, où (a_n) est une suite de réels positifs. Voici des exemples de séries alternées :

$$1 - \frac{1}{2} + \frac{1}{3} + \cdots + \frac{(-1)^{n+1}}{n} + \cdots = \sum_{n \geqslant 1} \frac{(-1)^{n+1}}{n}, \tag{21}$$

$$1 - \frac{1}{3} + \frac{1}{5} + \cdots + \frac{(-1)^n}{2n+1} + \cdots = \sum_{n \geqslant 0} \frac{(-1)^n}{2n+1}. \tag{22}$$

La première de ces séries est appelée *série harmonique alternée*.

Théorème 22 (Critère spécial des séries alternées).
Soit (a_n) une suite de réels positifs, décroissante et convergeant vers 0.
La série alternée $\sum_{n \geqslant 0} (-1)^n a_n$ est alors convergente :
De plus, pour tout $n \geqslant 0$, notons R_n le reste d'indice n de cette série.

1. On a $(-1)^{n+1} R_n \geqslant 0$, autrement dit R_n a (au sens large) le signe de son premier terme $(-1)^{n+1} a_{n+1}$..
2. On a $|R_n| \leqslant a_{n+1}$.

Démonstration. Notons u_n le terme général de la série considérée et (U_n) la suite des sommes partielles de cette série. Pour tout entier $p \geqslant 0$, posons $x_p := U_{2p+1}$ et $y_p := U_{2p}$. Nous allons montrer que les deux suites (x_p) et (y_p) sont *adjacentes*. Si c'est le cas, chacune convergera, et ces deux suites auront une même limite S (théorème 45 de la page 560). Alors S sera aussi limite de la suite (U_n), ce qui prouvera déjà la convergence de la série donnée.
Soit $p \in \mathbb{N}$. Compte tenu des définitions, il vient :

$$y_p - x_p = U_{2p} - U_{2p+1} = -u_{2p+1} = a_{2p+1} \geqslant 0.$$

Ainsi $x_p \leqslant y_p$, et en outre la suite $(y_p - x_p)$ converge vers 0. Ensuite

$$x_{p+1} - x_p = U_{2p+3} - U_{2p+1} = u_{2p+2} + u_{2p+3} = a_{2p+2} - a_{2p+3} \geqslant 0,$$

puisque la suite (a_n) décroît. La suite (x_p) est donc croissante. De même

$$y_p - y_{p+1} = U_{2p} - U_{2p+2} = -u_{2p+1} - u_{2p+2} = a_{2p+1} - a_{2p+2} \geqslant 0,$$

i.e. la suite (y_p) est décroissante. Ainsi les suites (x_p) et (y_p) sont adjacentes.
Pour tout $p \geqslant 0$, nous avons $x_p \leqslant S \leqslant y_p$. Rappelons pourquoi. Pour tout $n \geqslant p$, on a $x_p \leqslant x_n \leqslant y_n \leqslant y_p$, et il suffit de faire tendre n vers $+\infty$ dans ces inégalités. Ainsi $U_{2p+1} \leqslant S \leqslant U_{2p}$. Or $S = U_{2p} + R_{2p} = U_{2p+1} + R_{2p+1}$, et par suite $R_{2p} \leqslant 0 \leqslant R_{2p+1}$. Cela démontre l'assertion 1.
Soit enfin $n \in \mathbb{N}$. Puisque $R_n = u_{n+1} + R_{n+1}$, l'assertion 1 donne :

$$|R_n| = (-1)^{n+1} R_n = (-1)^{n+1} u_{n+1} + (-1)^{n+1} R_{n+1} = a_{n+1} - |R_{n+1}| \leqslant a_{n+1},$$

ce qui prouve l'assertion 2. ∎

Exercice 6.

Soit a un nombre réel fixé. Étudier la série alternée suivante :

$$\sum_{n\geqslant 1} u_n, \quad \text{où} \quad u_n := \frac{(-1)^{n+1}}{n^a}\cdot \tag{23}$$

Solution. Notons U_n ses sommes partielles. Si $a < 0$, $|u_n| = n^{-a}$ tend vers $+\infty$ lorsque n tend vers $+\infty$, donc la série diverge : son terme général ne tend pas vers 0. Si $a = 0$, on a $|u_n| = 1$ pour tout n, donc la série diverge pour la même raison que précédemment.

Supposons $a > 0$. La suite (n^{-a}) est décroissante et converge vers 0. D'après le critère spécial des séries alternées, la série en (23) converge. Si $a > 1$, le théorème 14 de la page 724 montre que cette série est absolument convergente. En revanche, si $0 < a \leqslant 1$, le même théorème montre que cette série n'est pas absolument convergente, elle est donc *semi-convergente*.

Exercice 7.

Pour $a := 1$, on obtient la série harmonique alternée (21), convergente d'après l'exercice 6. Montrer que sa somme vaut :

$$\sum_{n=1}^{+\infty} \frac{(-1)^{n+1}}{n} = \ln 2. \tag{24}$$

Solution. Voici l'idée. Si $n \in \mathbb{N}^*$, l'égalité (18) de la page 727, appliquée à n et $2n$, donne :

$$\left\{\begin{array}{l} 1+\dfrac{1}{2}+\dfrac{1}{3}+\cdots+\dfrac{1}{2n} = \ln(2n)+C+\varepsilon_{2n} \\[2ex] 1+\dfrac{1}{2}+\dfrac{1}{3}+\cdots+\dfrac{1}{n} = \ln n + C + \varepsilon_n. \end{array}\right. \tag{$*$}$$

où ε_n tend vers 0 lorsque n tend vers $+\infty$.

Notons A_n (resp. B_n) la somme des inverses des entiers pairs (resp. impairs) entre 1 et $2n$. La première égalité de $(*)$ s'écrit :

$$A_n + B_n = \ln(2n) + C + \varepsilon_{2n}.$$

Si l'on divise la deuxième égalité de $(*)$ par 2, il vient :

$$A_n = \frac{1}{2}\big[\ln n + C + \varepsilon_n\big], \quad \text{d'où}$$

$$B_n - A_n = \Big(\ln(2n) + C + \varepsilon_{2n}\Big) - \Big(\ln n + C + \varepsilon_n\Big) = \ln 2 + \varepsilon_{2n} - \varepsilon_n.$$

Puisque les suites $(\varepsilon_n)_{n\geqslant 1}$ et $(\varepsilon_{2n})_{n\geqslant 1}$ convergent vers 0, la suite $(B_n - A_n)$ converge vers $\ln 2$. D'où la conclusion, car

$$B_n - A_n = \left(1+\frac{1}{3}+\frac{1}{5}+\cdots+\frac{1}{2n-1}\right) - \left(\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2n}\right),$$

$$\text{soit} \qquad B_n - A_n = \sum_{k=1}^{2n} \frac{(-1)^{k+1}}{k}\cdot$$

Exercice 8.

Pour tout $n \geqslant 2$, posons :

$$a_n := \frac{1}{n+(-1)^n} \quad \text{et} \quad u_n := (-1)^n a_n = \frac{(-1)^n}{n+(-1)^n}\cdot$$

Montrer que la série $\sum\limits_{n\geqslant 2} u_n$ converge.

Solution. La suite $(a_n)_{n\geqslant 2}$ est bien une suite de réels positifs et elle converge vers 0. Cependant, on ne peut pas appliquer directement le critère spécial des séries alternées, car cette suite n'est pas décroissante : pour tout entier $p \geqslant 1$, on a :

$$a_{2p+1} = \frac{1}{(2p+1)-1} = \frac{1}{2p} > \frac{1}{2p+1} = a_{2p}.$$

L'idée est alors de « remplacer » la série donnée par la série plus simple $\sum\limits_{n\geqslant 2} v_n$, en posant $v_n := (-1)^n/n$.

La série $\sum\limits_{n\geqslant 2} v_n$ converge : c'est la série harmonique alternée commençant à l'indice 2. D'après le théorème 2 de la page 717, il suffit de montrer que $\sum\limits_{n\geqslant 2}(v_n - u_n)$ converge. Si $n \geqslant 2$, on a :

$$\begin{aligned} v_n - u_n &= \frac{(-1)^n}{n} - \frac{(-1)^n}{n+(-1)^n} = \frac{(-1)^n}{n\big(n+(-1)^n\big)}\Big[\big(n+(-1)^n\big) - n\Big] \\ &= \frac{(-1)^{2n}}{n\big(n+(-1)^n\big)} = \frac{1}{n\big(n+(-1)^n\big)} \end{aligned}$$

d'où $n^2(v_n - u_n) = \dfrac{1}{1+\frac{(-1)^n}{n}}\cdot$

Ainsi $n^2(v_n - u_n)$ tend vers 1 lorsque n tend vers $+\infty$. D'après le théorème 15 de la page 725, la série $\sum\limits_{n\geqslant 2}(v_n - u_n)$ converge, d'où la conclusion.

Exercice 9.

Une série alternée peut évidemment diverger. Étudier l'exemple suivant.

Solution. Pour tout entier $n \geqslant 1$, posons :

$$a_n := \frac{2+(-1)^n}{n} \quad \text{et} \quad u_n := (-1)^n a_n = \frac{1+2(-1)^n}{n}\cdot$$

La série $\sum\limits_{n\geqslant 1} u_n$ est alternée (le numérateur de a_n est positif, il vaut 3 ou 1 suivant la parité de n), et la suite (a_n) converge vers 0. Cependant, la série $\sum\limits_{n\geqslant 1} u_n$ diverge.

En effet, la série $\sum\limits_{n\geqslant 1} 2(-1)^n/n$ converge (*cf.* la série harmonique alternée), mais la série $\sum\limits_{n\geqslant 1} 2/n$ diverge. Il suffit alors d'appliquer le corollaire 3 de la page 718.

Pourquoi le critère spécial des séries alternées ne s'applique-t-il pas ? La réponse est simple : la suite (a_n) n'est pas décroissante : $a_n > a_{n+1}$ si n est pair, mais $a_n < a_{n+1}$ si n est impair (le vérifier).

Remarque. Contrairement aux suites, les séries ne se « multiplient » pas simplement (voir cependant le théorème 21 de la page 731). Soit par exemple $u_n := \dfrac{(-1)^n}{\sqrt{n}}$ pour $n \geqslant 1$. La série $\sum\limits_{n\geqslant 1} u_n$ converge, *cf.* la série en (23). Cependant la série $\sum\limits_{n\geqslant 1} u_n^2$ diverge : c'est la série harmonique. Ainsi, lorsque deux séries $\sum\limits_n u_n$ et $\sum\limits_n v_n$ convergent, la série $\sum\limits_n u_n v_n$ ne converge pas nécessairement ; cette série est appelée *produit d'Hadamard* des deux séries données.

Exercice type corrigé

IV.4.0 Soient a et b deux entiers tels que $0 < a < b$.

Montrer que la série $\sum\limits_{n\geqslant 0}^{+\infty} 1/\big((n+a)(n+b)\big)$ converge, et calculer sa somme.

Solution. Notons u_n le terme général de la série. Lorsque n tend vers $+\infty$, u_n est équivalent à $1/n^2$. Comme la série $\sum\limits_{n\geqslant 1} 1/n^2$ converge, le théorème d'équivalence (théorème 11 de la page 722) montre que $\sum\limits_{n\geqslant 0} u_n$ converge. Notons U_n les sommes partielles et S la somme de la série donnée.

À l'aide de la décomposition en éléments simples de la fraction rationnelle $\dfrac{1}{(X+a)(X+b)}$, on a :

$$\frac{1}{(n+a)(n+b)} = \frac{1}{b-a}\left[\frac{1}{n+a} - \frac{1}{n+b}\right].$$

On en déduit, pour tout $n \in \mathbb{N}$:

$$(b-a)U_n = \sum_{k=0}^{n}\left[\frac{1}{k+a} - \frac{1}{k+b}\right] = \left(\sum_{h=a}^{n+a}\frac{1}{h}\right) - \left(\sum_{h=b}^{n+b}\frac{1}{h}\right),$$

la deuxième égalité venant de deux changements d'indices évidents (se rappeler que a et b sont des entiers naturels).

Prenons n tel que $n + a > b$. Dans les deux sommes à soustraire, les termes $1/h$ pour $h = b$, $h = b+1, \ldots, h = n+a$ s'éliminent. Il reste :

$$(b-a)U_n = \left(\sum_{h=a}^{b-1}\frac{1}{h}\right) - \left(\sum_{h=n+a+1}^{n+b}\frac{1}{h}\right).$$

La deuxième somme est formée de $b-a$ termes compris entre 0 et $1/(n+a+1)$, elle est donc comprise entre 0 et $(b-a)/(n+a+1)$, et par suite elle tend vers 0 lorsque n tend vers $+\infty$. D'où, en passant à la limite :

$$(b-a)S = \lim_{n\to+\infty}(b-a)U_n = \sum_{h=a}^{b-1}\frac{1}{h}, \quad \text{soit} \quad S = \frac{1}{b-a}\left(\sum_{h=a}^{b-1}\frac{1}{h}\right).$$

On remarquera que la série donnée converge dès que a et b sont des *réels* strictement positifs. Par contre, ce qui précède ne permet de calculer la somme de cette série que lorsque a et b sont deux *entiers* strictement posifis distincts.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

IV.4.1 Soit p un entier fixé, $p \geqslant 2$. Pour tout $n \in \mathbb{N}$, on pose :

$$u_n := \frac{1}{(n+1)(n+2)\cdots(n+p)}.$$

1. Montrer que la série $\sum_{n\geqslant 0} u_n$ converge.
2. Démontrer pour tout $n \in \mathbb{N}$ la formule suivante :

$$U_n = \frac{1}{p-1}\left[\frac{1}{(p-1)!} - \frac{1}{(n+2)(n+3)\cdots(n+p)}\right],$$

U_n ayant le sens habituel. En déduire $\sum_{n=0}^{+\infty} u_n$.

IV.4.2 Pour tout entier $n \geqslant 1$, on pose $u_n := 0$ si l'écriture décimale de n comporte un chiffre 8 et $u_n := 1/n$ sinon. Montrer que la série $\sum_{n\geqslant 1} u_n$ converge. Pour tout entier $p \geqslant 0$, on majorera convenablement la somme $T_p := \sum_{n=10^p}^{10^{p+1}-1} u_n$.

IV.4.3 Soit (u_n) une suite réelle décroissante, à termes positifs.
Montrer que les deux séries $\sum_{n\geqslant 1} u_n$ et $\sum_{k\geqslant 0} 2^k u_{2^k}$ sont de même nature. En prenant $u_n := n^{-a}$, où le réel $a > 0$ est fixé, retrouver le théorème 14 de la page 724.

IV.4.4 Soit (u_n) une suite réelle décroissante, à termes positifs. À l'aide du théorème 6 de la page 719 montrer que, si la série $\sum_n u_n$ converge, nu_n tend vers 0 lorsque n tend vers $+\infty$. Montrer que la réciproque est fausse ; on pourra considérer la suite $\big(1/(n\ln n)\big)_{n\geqslant 2}$.

IV.4.5 Prouver la convergence et calculer la somme de la série :

$$\sum_{n\geqslant 1} \frac{E\big(\sqrt{n+1}\big) - E\big(\sqrt{n}\big)}{n}.$$

On rappelle que $E(x)$ désigne la partie entière d'un nombre réel x.

IV.4.6 Montrer que les séries $\sum_{n\geqslant 1} 1/n^{\ln n}$ et $\sum_{n\geqslant 2} 1/(\ln n)^n$ convergent.

IV.4.7 Soit $a > 0$ un nombre réel fixé. Déterminer, suivant la valeur de a, la nature de la série de terme général a^{H_n} $(n \geqslant 1)$, où

$$H_n := 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

On pourra utiliser l'égalité (18) de la page 727.

IV.4.8 Soient a, b deux réels tels que $0 < a < b$. Soit $(u_n)_{n\geqslant 1}$ la suite définie par son premier terme $u_1 \in \mathbb{C}^*$ et la relation de récurrence :

$$u_{n+1} := \left(\frac{n+a}{n+b}\right) u_n \ \ (n \geqslant 1)\,.$$

1. Soit (U_n) la suite des sommes partielles de la série $\sum\limits_{n\geqslant 1} u_n$. Pour tout entier $n \geqslant 1$, prouver la formule :
$$(b-a-1)U_n \ = \ bu_1 - (n+b)u_{n+1}\,.$$
2. En déduire que la série $\sum\limits_{n\geqslant 1} u_n$ converge si, et seulement si, $b > a+1$, et déterminer dans ce cas la somme de la série.

IV.4.9 $*$ On pose $a := 2+\sqrt{5}$ et $b := 2-\sqrt{5}$.

1. Montrer que, pour tout $n \geqslant 0$, $a^n + b^n$ est un entier pair.
2. En déduire que la série de terme général $\sin\big(\pi a^n\big)$ converge (on rappelle que $|\sin t| \leqslant t$ pour tout réel $t \geqslant 0$).

IV.4.10 Soit f une fonction réelle continue et décroissante sur $]0, +\infty[$. On suppose que $f(x)$ tend vers 0 lorsque x tend vers $+\infty$. Soit F une primitive de f : F est dérivable, et sa dérivée est f. Montrer que la série $\sum\limits_{n\geqslant 1} f(n)$ converge si, et seulement si, $F(x)$ a une limite finie lorsque x tend vers $+\infty$. On appliquera la proposition 13 de la page 724.

Application : soit b un nombre réel. Montrer que la série de terme général $\dfrac{1}{n(\ln n)^b}$ $(n \geqslant 2)$ converge si, et seulement si, $b > 1$.

IV.4.11 Soit $(a_n)_{n\geqslant 1}$ une suite à termes réels, décroissante et convergeant vers 0.
Pour tout $n \geqslant 1$, on pose $b_n := n(a_n - a_{n+1})$. On note A_n, B_n les sommes partielles de ces séries.

1. Pour tout entier $n \geqslant 1$, prouver l'égalité $B_n = A_n - na_{n+1}$.
2. Montrer que les séries $\sum\limits_{n\geqslant 1} a_n$ et $\sum\limits_{n\geqslant 1} b_n$ sont de même nature et que, lorsque ces deux séries convergent, leurs sommes sont égales.

IV.4.12 $*$ Soit a un réel fixé. Montrer que la série

$$\sum_{n\geqslant 1} \frac{(-1)^n}{n^a + (-1)^n}$$

converge si, et seulement si, $a > \dfrac{1}{2}$ (s'inspirer de l'exercice 8 de la page 734).

IV.4.13 Soit a un nombre complexe fixé, on suppose que $-a \notin \mathbb{N}$.
Déterminer la nature de la série $\sum\limits_{n\geqslant 0} \dfrac{(-1)^n}{n+a}$.

IV.4.14 $*$ Soit $a > 0$ un nombre réel fixé. Le but de cet exercice est d'établir la formule suivante :

$$\sum_{n=0}^{+\infty} \frac{(-1)^n}{an+1} = \int_0^1 \frac{\mathrm{d}x}{1+x^a}. \qquad (*)$$

On note I l'intégrale, u_n le terme général et U_n la somme partielle d'indice n de la série.

1. Déduire la formule $(*)$ des inégalités $\mid U_n - I \mid \leqslant \dfrac{1}{a(n+1)+1}$, $n \in \mathbb{N}$.
2. Pour tout $n \in \mathbb{N}$, soit R_n le reste d'indice n de la série en $(*)$. À l'aide de la proposition 13 de la page 724 et du théorème 11 de la page 722, montrer que $2an|R_n|$ tend vers 1 lorsque n tend vers $+\infty$ (grouper les u_n par deux).
3. Démontrer la formule suivante, en faisant intervenir la fonction $\arctan$: $\displaystyle\sum_{n=0}^{+\infty} \frac{(-1)^n}{2n+1} = \frac{\pi}{4}$.

IV.4.15 Démontrer l'égalité suivante :

$$1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \frac{1}{5} - \frac{1}{10} - \frac{1}{12} + \frac{1}{7} + \cdots = \frac{1}{2}\ln 2.$$

Comparer avec l'égalité (24) de la page 733. Que peut-on en conclure ?

IV.4.16 $*$ Soit $(\varepsilon_n)_{n\geqslant 0}$ une suite dont chaque terme vaut 1 ou -1.
On pose pour tout $n \in \mathbb{N}$: $u_n := \dfrac{\varepsilon_0\varepsilon_1\cdots\varepsilon_n}{2^n}$.

1. Montrer que la série $\sum\limits_{n\geqslant 0} u_n$ converge, et que sa somme S appartient à $[-2, 2]$. On note (U_n) la suite des sommes partielles de cette série.
2. On définit une suite de nombres réels (x_n) comme suit :

$$x_0 := \varepsilon_0\sqrt{2}\,,\ \ x_1 := \varepsilon_0\sqrt{2+\varepsilon_1\sqrt{2}}\,,\ \ x_2 := \varepsilon_0\sqrt{2+\varepsilon_1\sqrt{2+\varepsilon_2\sqrt{2}}}\,,\ldots$$

Justifier cette définition, et prouver que $x_n = 2\sin(\pi U_n/4)$ pour tout $n \in \mathbb{N}$.
En déduire que la suite (x_n) converge, et préciser sa limite.

IV.4.17 $*$ Soit $(a_n)_{n\geqslant 0}$ une suite à termes réels, décroissante et convergeant vers 0. Par ailleurs, soient m un entier strictement positif et $(\varepsilon_n)_{n\geqslant 0}$ une suite à termes complexes, périodique de période m ($\varepsilon_{k+m} = \varepsilon_k$ pour tout $k \in \mathbb{N}$). Le but de cet exercice est d'étudier la nature de la série $\sum\limits_{n\geqslant 0} \varepsilon_n a_n$. On note A_n, T_n, U_n les sommes partielles des séries $\sum\limits_{n\geqslant 0} a_n$, $\sum\limits_{n\geqslant 0} \varepsilon_n$, $\sum\limits_{n\geqslant 0} \varepsilon_n a_n$ respectivement et S la somme $\varepsilon_0 + \varepsilon_1 + \cdots + \varepsilon_{m-1}$.
Montrer que l'on peut écrire, pour tout $n \geqslant 0$, $T_n = nS/m + w_n$, où (w_n) est une suite bornée.
Établir ensuite pour tout $n \in \mathbb{N}^*$ l'égalité :

$$U_n = \frac{S(A_n - a_0)}{m} + w_n a_{n+1} + \sum_{k=0}^{n} w_k(a_k - a_{k+1}).$$

On montrera alors que la série $\sum\limits_{n\geqslant 0} \varepsilon_n a_n$ converge si, et seulement si, ou bien $S = 0$, ou bien la série $\sum\limits_{n\geqslant 0} a_n$ converge (ou les deux).

Introduction à l'intégration

IV.5

Jusqu'à la fin de la section 3.4, a *et* b *désignent deux réels tels que* $a < b$ et toutes les fonctions sont supposées définies sur $[a, b]$ et à valeurs réelles.

Le but de l'intégration est de définir un nombre qui, pour une fonction f positive sur un segment $[a, b]$, mesure l'aire délimitée, dans un repère orthonormé, par sa courbe représentative, l'axe des abscisses et les deux droites d'équations $x = a$ et $x = b$.
Ce nombre sera appelé intégrale de f sur $[a, b]$ et sera noté $\displaystyle\int_{[a,b]} f$ ou $\displaystyle\int_a^b f(x)\,\mathrm{d}x$.

Si f est en escalier, son intégrale doit donc être la somme des aires des rectangles délimités par sa courbe représentative (son graphe). Par définition, cette aire est le produit de la hauteur par la largeur, l'aire d'un carré de côté 1 étant égale à 1.
Dans le cas général, s'il existe deux fonctions en escalier φ et ψ vérifiant $\varphi \leqslant f \leqslant \psi$, on doit avoir :

$$\int_{[a,b]} \varphi \leqslant \int_{[a,b]} f \leqslant \int_{[a,b]} \psi.$$

Dans les bons cas, et en particulier lorsque f est continue par morceaux, on peut prendre φ et ψ aussi proches que l'on veut l'une de l'autre, et avoir ainsi un encadrement aussi précis que l'on veut de l'intégrale de f.

1 Intégrale des fonctions en escalier

1.1 Subdivision d'un segment

Définition 1.

- On appelle *subdivision* de $[a,b]$, toute famille $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$ telle que $n \in \mathbb{N}^*$ et :
$$a = x_0 < x_1 < \cdots < x_n = b.$$
- On appelle *pas* ou *module* de la subdivision $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$, le réel :
$$\delta(\boldsymbol{u}) = \max_{i\in[\![1,n]\!]}(x_i - x_{i-1}).$$

Exemple. Étant donné un entier naturel n non nul, on définit une subdivision $(x_i)_{i\in[\![0,n]\!]}$ de $[a,b]$ par :
$$\forall i \in [\![0,n]\!]\,,\ x_i = a + i\,\frac{b-a}{n}\cdot$$
Elle est appelée *subdivision à pas constant* ; son module est $\dfrac{b-a}{n}$ et l'on peut le rendre aussi petit que l'on veut en choisissant n suffisamment grand..

Définition 2. Si $\boldsymbol{u}$ et $\boldsymbol{v}$ sont deux subdivisions de $[a,b]$, on dit que $\boldsymbol{u}$ est *plus fine* que $\boldsymbol{v}$ si tout élément de $\boldsymbol{v}$ est élément de $\boldsymbol{u}$.

Remarque. Se donner une subdivision de $[a,b]$, équivaut à se donner une partie finie de $[a,b]$ contenant a et b.

La relation définie ci–dessus est une relation d'ordre sur l'ensemble des subdivisions de $[a,b]$. Elle correspond à l'inclusion sur les parties finies de $[a,b]$ contenant a et b.
Comme la réunion de telles parties A et B est une partie finie C contenant a et b et telle que $A \subset C$ et $B \subset C$, on en déduit :

Proposition 1. Si $\boldsymbol{u}$ et $\boldsymbol{v}$ sont deux subdivisions de $[a,b]$, il existe une subdivision plus fine que $\boldsymbol{u}$ et $\boldsymbol{v}$.

1.2 Fonctions en escalier

Définition 3. Une fonction φ de $[a,b]$ dans $\mathbb{R}$ est *en escalier* si l'on peut trouver une subdivision $(x_i)_{i\in[\![0,n]\!]}$ de $[a,b]$ telle que φ soit constante sur chacun des intervalles $]\,x_{i-1}, x_i\,[$ $(1 \leqslant i \leqslant n)$.
Une telle subdivision est appelée *subdivision adaptée* à la fonction en escalier φ.

Notation. On note $\mathcal{E}\big([a,b]\big)$ l'ensemble des fonctions en escalier sur $[a,b]$.

Exemples.

1. Une fonction constante sur le segment $[a,b]$ est en escalier sur $[a,b]$.
2. La fonction partie entière est en escalier sur tout segment $[a,b]$, une subdivision adaptée étant constituée de a, b et de tous les entiers compris entre a et b.

Remarques.

- Une fonction en escalier prend un nombre fini de valeurs. En particulier, elle est bornée.
- Si $\boldsymbol{u}$ est une subdivision adaptée à une fonction φ en escalier, alors toute subdivision plus fine que $\boldsymbol{u}$ est adaptée à φ.
- Si φ et ψ sont deux fonctions en escalier sur $[a,b]$, alors il existe une subdivision adaptée à φ et à ψ.
  En effet, soit $\boldsymbol{u}$ (respectivement $\boldsymbol{v}$) une subdivision adaptée à φ (respectivement à ψ). D'après la proposition 1 de la page précédente, il existe une subdivision $\boldsymbol{w}$ plus fine que $\boldsymbol{u}$ et $\boldsymbol{v}$. Une telle subdivision $\boldsymbol{w}$ est alors adaptée à φ et à ψ.

Proposition 2. $\mathcal{E}\big([a,b]\big)$ est un sous–espace vectoriel de $\mathcal{F}([a,b],\mathbb{R})$.

Démonstration. Soient φ et ψ deux fonctions en escalier sur $[a,b]$, ainsi que λ et μ deux réels. Prenons une subdivision $(x_i)_{i\in[\![0,n]\!]}$ adaptée à φ et à ψ. Les fonctions φ et ψ sont constantes sur chacun des intervalles $]\,x_{i-1},x_i\,[$ et donc il en est de même de $\lambda\,\varphi+\mu\,\psi$. La fonction nulle étant en escalier, on en déduit le résultat. ∎

1.3 Intégrale d'une fonction en escalier

Proposition et définition 3. Soient $\varphi\in\mathcal{E}\big([a,b]\big)$ et $\boldsymbol{u}=(x_i)_{i\in[\![0,n]\!]}$ une subdivision adaptée à φ. Si, pour $1\leqslant i\leqslant n$, on note c_i la valeur de φ sur $]\,x_{i-1},x_i\,[$, alors le réel $\sum\limits_{i=1}^{n}(x_i-x_{i-1})\,c_i$ ne dépend pas de la subdivision $\boldsymbol{u}$ choisie.

On l'appelle *intégrale* de φ sur $[a,b]$ et on la note $\displaystyle\int_{[a,b]}\varphi$.

Démonstration. Si $\boldsymbol{u}$ est une subdivision adaptée à la fonction φ, notons :

$$I(\varphi,\boldsymbol{u})=\sum_{i=1}^{n}(x_i-x_{i-1})\,c_i.$$

On veut montrer que $I(\varphi,\boldsymbol{u})=I(\varphi,\boldsymbol{v})$ pour toutes subdivisions $\boldsymbol{u}$ et $\boldsymbol{v}$ adaptées à φ.

Si $\boldsymbol{v}$ est plus fine que $\boldsymbol{u}$, elle est obtenue en ajoutant un nombre fini d'éléments à la subdivision $\boldsymbol{u}$. Pour démontrer que $I(\varphi,\boldsymbol{u})=I(\varphi,\boldsymbol{v})$, il suffit donc de le démontrer dans le cas où $\boldsymbol{v}$ a un élément de plus que $\boldsymbol{u}$.

Supposons donc $\boldsymbol{u}=(x_i)_{i\in[\![0,n]\!]}$ et $\boldsymbol{v}=(x_0,x_1,\ldots,x_{p-1},y,x_p,\ldots,x_n)$. La fonction φ vaut c_p sur $]\,x_{p-1},y\,[$ et sur $]\,y,x_p\,[$, et :

$$\begin{aligned}I(\varphi,\boldsymbol{v})&=\sum_{i=1}^{p-1}(x_i-x_{i-1})\,c_i+(y-x_{p-1})\,c_p\\&\qquad+(x_p-y)\,c_p+\sum_{i=p+1}^{n}(x_i-x_{i-1})\,c_i\\&=\sum_{i=1}^{n}(x_i-x_{i-1})\,c_i=I(\varphi,\boldsymbol{u}).\end{aligned}$$

Cas général. Si $\boldsymbol{u}$ et $\boldsymbol{v}$ sont deux subdivision adaptées à φ, alors il existe une subdivision $\boldsymbol{w}$ plus fine que $\boldsymbol{u}$ et $\boldsymbol{v}$, qui est donc aussi adaptée à φ. D'après le premier cas, on a :

$$I(\varphi, \boldsymbol{u}) = I(\varphi, \boldsymbol{w}) = I(\varphi, \boldsymbol{v})$$

d'où le résultat. ∎

Remarques.

- L'intégrale de φ ne dépend pas des valeurs de φ aux points de la subdivision adaptée qui a servi à la calculer.
- Une fonction nulle sauf en un nombre fini de points est en escalier et son intégrale est nulle, comme on peut le vérifier en prenant une subdivision contenant les points où la fonction n'est pas nulle.

1.4 Propriétés de l'intégrale des fonctions en escalier

Proposition 4. L'intégrale est une forme linéaire sur $\mathcal{E}\big([a,b]\big)$.

Démonstration. Soient φ et ψ deux fonctions en escalier sur $[a,b]$ et $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$ une subdivision adaptée à φ et ψ. Si c_i (respectivement d_i) est la valeur de la fonction φ (respectivement ψ) sur l'intervalle $]\,x_{i-1}, x_i\,[$, alors, sur ce même intervalle, $\lambda\,\varphi + \mu\,\psi$ vaut $\lambda\,c_i + \mu\,d_i$, et :

$$\begin{aligned}\int_{[a,b]} (\lambda\,\varphi + \mu\,\psi) &= \sum_{i=1}^{n} (x_i - x_{i-1})\,(\lambda\,c_i + \mu\,d_i)\\ &= \lambda \sum_{i=1}^{n} (x_i - x_{i-1})\,c_i + \mu \sum_{i=1}^{n} (x_i - x_{i-1})\,d_i\\ &= \lambda \int_{[a,b]} \varphi + \mu \int_{[a,b]} \psi.\end{aligned}$$

∎

Proposition 5.

- Toute fonction en escalier positive a une intégrale positive.
- Si φ et ψ sont deux fonctions en escalier sur $[a,b]$ alors :

$$\varphi \leqslant \psi \Longrightarrow \int_{[a,b]} \varphi \leqslant \int_{[a,b]} \psi.$$

Démonstration.

- En utilisant les notations de la proposition et défintiion 3 de la page précédente, si φ est positive alors les c_i sont tous positifs et, puisque $x_i - x_{i-1} > 0$, on en déduit :

$$\int_{[a,b]} \varphi = \sum_{i=1}^{n} (x_i - x_{i-1})\,c_i \geqslant 0.$$

- Il suffit d'appliquer le résultat précédent à la fonction en escalier positive $\psi - \varphi$ et d'utiliser la linéarité de l'intégrale. ∎

Proposition 6. Soient $c \in \,]\,a,b\,[$ et φ une fonction définie sur $[a,b]$. La fonction φ est en escalier sur $[a,b]$ si, et seulement si, ses restrictions à $[\,a,c\,]$ et à $[\,c,b\,]$ sont en escalier, et l'on a alors :

$$\int_{[a,b]} \varphi = \int_{[\,a,c\,]} \varphi_{|_{[a,c]}} + \int_{[\,c,b\,]} \varphi_{|_{[c,b]}}.$$

Démonstration.

- Soit φ une fonction en escalier sur $[a,b]$. Prenons une subdivision adaptée à φ. En lui ajoutant éventuellement le point c, on obtient une subdivision $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$ encore adaptée à φ et contenant le point c.
  Désignons par p l'entier strictement compris entre 0 et n tel que $c = x_p$.
  - $(x_i)_{i\in[\![0,p]\!]}$ est une subdivision de $[\,a,c\,]$ et pour tout $i \in [\![1,p]\!]$, la fonction φ est constante sur $]\,x_{i-1},x_i\,[$. Donc $\varphi_{|_{[a,c]}}$ est en escalier sur $[\,a,c\,]$.
  - De même, $(x_i)_{i\in[\![p,n]\!]}$ est une subdivision de $[\,c,b\,]$ et la fonction $\varphi_{|_{[c,b]}}$ est en escalier sur $[\,c,b\,]$.

  Si c_i est la valeur de φ sur l'intervalle $]\,x_{i-1},x_i\,[$, on a alors :

$$\begin{aligned}\int_{[a,b]} \varphi &= \sum_{i=1}^{n} (x_i - x_{i-1})\, c_i \\ &= \sum_{i=1}^{p} (x_i - x_{i-1})\, c_i + \sum_{i=p+1}^{n} (x_i - x_{i-1})\, c_i \\ &= \int_{[\,a,c\,]} \varphi_{|_{[a,c]}} + \int_{[\,c,b\,]} \varphi_{|_{[c,b]}}.\end{aligned}$$

- Réciproquement, supposons $\varphi_{|_{[a,c]}}$ et $\varphi_{|_{[c,b]}}$ en escalier. Si :
  - $(x_i)_{i\in[\![0,p]\!]}$ est une subdivision de $[\,a,c\,]$ adaptée à $\varphi_{|_{[a,c]}}$,
  - $(y_j)_{j\in[\![0,q]\!]}$ est une subdivision de $[\,c,b\,]$ adaptée à $\varphi_{|_{[c,b]}}$,

  alors, $(x_0, x_1, \ldots, x_{p-1}, x_p = c = y_0, y_1, \ldots, y_q)$ est une subdivision de $[a,b]$ et φ est constante sur chacun des intervalles $]\,x_{i-1},x_i\,[$ et $]\,y_{j-1},y_j\,[$. Donc φ est en escalier sur $[a,b]$. ■

2 Fonctions continues par morceaux

2.1 Définition, exemples

Définition 4. Une application f de $[a,b]$ dans $\mathbb{R}$ est *continue par morceaux* s'il existe une subdivision $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$ de $[a,b]$ telle que pour tout $i \in [\![1,n]\!]$, la restriction de f à $]\,x_{i-1},x_i\,[$ soit continue et admette des limites finies en x_{i-1} et x_i.
Une telle subdivision $\boldsymbol{u}$ est dite adaptée à la fonction continue par morceaux f.

Remarques.

- Une fonction f de $[a,b]$ dans $\mathbb{R}$ est continue par morceaux si, et seulement si, l'on peut trouver une subdivision $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$ de $[a,b]$ telle que la restriction de f à chacun des intervalles $]\,x_{i-1}, x_i\,[$ se prolonge en une fonction f_i continue sur le segment $[\,x_{i-1}, x_i\,]$.
- Une fonction continue par morceaux sur $[a,b]$ admet une limite finie à droite en tout point de $[a,b[$ et une limite finie à gauche en tout point de $]a,b]$. C'est en effet vrai par définition en tout point d'une subdivision adaptée, et sinon parce que f est continue en un tel point.

Exemples.

1. Toute fonction en escalier est continue par morceaux.
2. Toute fonction continue sur un segment est continue par morceaux.
3. La fonction f définie sur $[\,-1,1\,]$ par :
$$f(0) = 0 \qquad \text{et} \qquad \forall x \neq 0\,,\ f(x) = \mathrm{e}^{1/x}$$
n'est pas continue par morceaux, car elle n'a pas de limite finie à droite en 0.
4. Si f est continue par morceaux sur $[a,b]$ et si $\alpha \in \mathbb{R}$, alors la fonction f_α définie sur $[\,a+\alpha, b+\alpha\,]$ par $f_\alpha(x) = f(x-\alpha)$ est continue par morceaux.
Réciproquement, toute fonction f continue par morceaux sur $[a+\alpha, b+\alpha]$ est ainsi obtenue à partir de la fonction $x \mapsto f(x+\alpha)$ continue par morceaux sur $[a,b]$.

Remarques. Comme pour les fonctions en escalier, on peut vérifier que :

- si $\boldsymbol{u}$ est une subdivision adaptée à une fonction f continue par morceaux, alors toute subdivision plus fine que $\boldsymbol{u}$ est adaptée à f,
- si f et g sont deux fonctions continues par morceaux sur $[a,b]$, alors il existe une subdivision adaptée à f et à g.

Proposition 7.
Toute fonction continue par morceaux sur le segment $[a,b]$ est bornée sur $[a,b]$.

Démonstration. Soit f une fonction continue par morceaux sur $[a,b]$. Prenons une subdivision $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$ adaptée à f. Pour $i \in [\![1,n]\!]$, la restriction de f à $]\,x_{i-1}, x_i\,[$ se prolonge en une fonction continue sur le segment $[\,x_{i-1}, x_i\,]$ qui est donc bornée. Par suite, la fonction f est bornée sur $]\,x_{i-1}, x_i\,[$ et l'on peut alors poser $M_i = \sup\limits_{]\,x_{i-1},x_i\,[} |f|$.

La fonction f est donc bornée sur $[a,b]$ par :
$$\max\Big(M_1, M_2, \ldots, M_n, \big|f(x_0)\big|, \big|f(x_1)\big|, \ldots, \big|f(x_n)\big|\Big).$$
■

Proposition 8. L'ensemble des fonctions continues par morceaux sur $[a,b]$ constitue un sous–espace vectoriel de $\mathcal{F}([a,b],\mathbb{R})$. Il est de plus stable par produit.

Démonstration. Soient f et g deux fonctions continues par morceaux sur $[a,b]$, ainsi que λ et μ deux réels. Prenons une subdivision $(x_i)_{i\in[\![0,n]\!]}$ adaptée à f et à g. Les restrictions des fonctions f et g à chacun des intervalles $]\,x_{i-1}, x_i\,[$ sont continues et admettent

des limites finies en x_{i-1} et x_i, donc il en est de même pour $\lambda f + \mu g$ et $f g$. Les fonctions $\lambda f + \mu g$ et $f g$ sont donc continues par morceaux sur $[a, b]$.

Comme de plus la fonction nulle est continue par morceaux, on en déduit le résultat. ■

2.2 Approximation des fonctions continues par morceaux

Pour définir l'intégrale des fonctions continues par morceaux, on va montrer que l'on peut approcher ou encadrer une fonction continue par morceaux d'aussi près que l'on veut par des fonctions en escalier. La démonstration de ce résultat peut, dans un premier temps, être sautée.

Théorème 9. Soit f une fonction continue par morceaux sur le segment $[a, b]$. Pour tout réel $\varepsilon > 0$:

- il existe une fonction θ en escalier sur $[a, b]$ telle que $|f - \theta| \leqslant \varepsilon$,
- il existe des fonctions φ et ψ en escalier sur $[a, b]$ telles que :
$$\varphi \leqslant f \leqslant \psi \qquad \text{et} \qquad \psi - \varphi \leqslant \varepsilon.$$

Démonstration.

- Soit $\varepsilon > 0$. Considérons l'ensemble J des réels $x \in [a, b]$ pour lesquels il existe $\theta \in \mathcal{E}\big([a, b]\big)$ telle que $\forall t \in [a, x]\,,\ \big|f(t) - \theta(t)\big| \leqslant \varepsilon$. C'est une partie non vide de $\mathbb{R}$ puisqu'elle contient a (la fonction constante $\theta = f(a)$ vérifie la propriété demandée sur $[a, a]$) et majorée par b. Elle possède donc une borne supérieure $c \in [a, b]$. Montrons que c appartient à J puis que $c = b$.
  - Si $c = a$, alors $c \in J$.
  - Supposons donc $c > a$. D'après une remarque de la page précédente, f admet une limite finie ℓ à gauche en c. On peut donc trouver $\eta > 0$ tel que $\forall t \in [c - \eta, c[\,,\ \big|f(t) - \ell\big| \leqslant \varepsilon$. D'autre part, par caractérisation de la borne supérieure c, il existe $x \in J$ tel que $c - \eta \leqslant x \leqslant c$. Prenons alors une fonction $\theta \in \mathcal{E}\big([a, b]\big)$ telle que $\forall t \in [a, x]\,,\ \big|f(t) - \theta(t)\big| \leqslant \varepsilon$. Il suffit alors de considérer la fonction en escalier $\hat{\theta}$ définie sur $[a, b]$ par :
$$\hat{\theta}(t) = \begin{cases} \theta(t) & \text{si } t \in [a, x] \\ \ell & \text{si } x < t < c \\ f(c) & \text{si } t \geqslant c. \end{cases}$$
pour avoir $\forall t \in [a, c]\,,\ \big|f(t) - \hat{\theta}(t)\big| \leqslant \varepsilon$. Ainsi $c \in J$.
  - Supposons $c < b$. La fonction f admet donc une limite finie ℓ à droite en c et l'on peut donc trouver $\eta > 0$ tel que $\forall t \in]c, c + \eta]\,,\ \big|f(t) - \ell\big| \leqslant \varepsilon$. En prenant une fonction $\theta \in \mathcal{E}\big([a, b]\big)$ telle que $\forall t \in [a, c]\,,\ \big|f(t) - \theta(t)\big| \leqslant \varepsilon$, on construit, comme précédemment une fonction $\hat{\theta}$ en escalier sur $[a, b]$ telle que $\forall t \in [a, c + \eta]\,,\ \big|f(t) - \hat{\theta}(t)\big| \leqslant \varepsilon$ ce qui donne $c + \eta \in J$ et contredit la définition de c.

  Donc $c = b$.
- En prenant une fonction en escalier θ telle que $|f - \theta| \leqslant \varepsilon/2$, les deux fonctions en escalier $\varphi = \theta - \varepsilon/2$ et $\psi = \theta + \varepsilon/2$ vérifient alors :
$$\varphi \leqslant f \leqslant \psi \qquad \text{et} \qquad \psi - \varphi \leqslant \varepsilon.$$ ■

2.3 Intégrale d'une fonction continue par morceaux

Notation. Si f est continue par morceaux sur $[a,b]$, on note :

- $\mathcal{E}^+(f)$ l'ensemble des fonctions en escalier sur $[a,b]$ plus grandes que f.
- $\mathcal{E}^-(f)$ l'ensemble des fonctions en escalier sur $[a,b]$ plus petites que f.

Proposition 10. Si f est continue par morceaux, alors :

- $\left\{\int_{[a,b]} \varphi \mid \varphi \in \mathcal{E}^-(f)\right\}$ admet une borne supérieure,
- $\left\{\int_{[a,b]} \psi \mid \psi \in \mathcal{E}^+(f)\right\}$ admet une borne inférieure,

et ces deux bornes sont égales.

Démonstration. Soit f une fonction continue par morceaux sur $[a,b]$. D'après la proposition 7 de la page 744, la fonction f est bornée ; notons $m = \inf\limits_{[a,b]} f$ et $M = \sup\limits_{[a,b]} f$.

- Les fonctions constantes m et M appartiennent respectivement à $\mathcal{E}^-(f)$ et $\mathcal{E}^+(f)$, donc $\mathcal{E}^-(f)$ et $\mathcal{E}^+(f)$ sont non vides.
  De plus, si $\varphi \in \mathcal{E}^-(f)$, alors $\varphi \leqslant f \leqslant M$, donc :
  $$\int_{[a,b]} \varphi \leqslant \int_{[a,b]} M = M(b-a).$$
  L'ensemble $\left\{\int_{[a,b]} \varphi \mid \varphi \in \mathcal{E}^-(f)\right\}$ est une partie non vide majorée de $\mathbb{R}$ et possède donc une borne supérieure α.
  De même, $\left\{\int_{[a,b]} \psi \mid \psi \in \mathcal{E}^+(f)\right\}$ est une partie non vide minorée de $\mathbb{R}$ et possède donc une borne inférieure β.
- Toute fonction φ de $\mathcal{E}^-(f)$ est inférieure à toute fonction ψ de $\mathcal{E}^+(f)$ et par suite :
  $$\int_{[a,b]} \varphi \leqslant \int_{[a,b]} \psi.$$
  Soit $\psi \in \mathcal{E}^+(f)$ fixée. L'ensemble $\left\{\int_{[a,b]} \varphi \mid \varphi \in \mathcal{E}^-(f)\right\}$ étant majoré par $\int_{[a,b]} \psi$, sa borne supérieure α est plus petite que cette intégrale.
  Le réel α est donc un minorant de l'ensemble $\left\{\int_{[a,b]} \psi \mid \psi \in \mathcal{E}^+(f)\right\}$, par conséquent il est plus petit que la borne inférieure de ce dernier. On en déduit donc $\alpha \leqslant \beta$.

- Soit $\varepsilon > 0$. D'après le théorème 9 de la page 745, on peut trouver des fonctions $\varphi \in \mathcal{E}^-(f)$ et $\psi \in \mathcal{E}^+(f)$ telles que $\psi - \varphi \leqslant \varepsilon$. On a alors :

$$\int_{[a,b]} \psi - \int_{[a,b]} \varphi = \int_{[a,b]} (\psi - \varphi) \leqslant \varepsilon\,(b-a)\,.$$

Or, par définition de α et β, on a :

$$\int_{[a,b]} \varphi \leqslant \alpha \leqslant \beta \leqslant \int_{[a,b]} \psi.$$

On a donc :

$$\forall \varepsilon > 0\,,\ 0 \leqslant \beta - \alpha \leqslant \varepsilon\,(b-a)$$

ce qui prouve $\alpha = \beta$. ■

Définition 5. Si f est une fonction continue par morceaux sur $[a,b]$, on appelle *intégrale* de f sur $[a,b]$ le réel :

$$\int_{[a,b]} f = \sup\left\{\int_{[a,b]} \varphi \mid \varphi \in \mathcal{E}^-(f)\right\} = \inf\left\{\int_{[a,b]} \psi \mid \psi \in \mathcal{E}^+(f)\right\}.$$

Cohérence. Si f est en escalier sur $[a,b]$, alors elle est continue par morceaux et son intégrale en tant que fonction continue par morceaux est la même que son intégrale en tant que fonction en escalier, ce qui permet de noter de la même façon ces deux intégrales.

En effet, $f \in \mathcal{E}^-(f) \cap \mathcal{E}^+(f)$ et donc son intégrale en tant que fonction en escalier est le plus grand élément de $\left\{\int_{[a,b]} \varphi \mid \varphi \in \mathcal{E}^-(f)\right\}$ et le plus petit élément de $\left\{\int_{[a,b]} \psi \mid \psi \in \mathcal{E}^+(f)\right\}$.

Proposition 11. Soit f continue par morceaux sur $[a,b]$. Pour toute suite (ε_n) à valeurs réelles strictement positives et tendant vers 0, il existe une suite de fonctions en escalier (θ_n) telles que :

$$\forall n \in \mathbb{N}\,,\ \forall t \in [a,b]\,,\ \left|f(t) - \theta_n(t)\right| \leqslant \varepsilon_n$$

et l'on a alors $\displaystyle\int_{[a,b]} f = \lim_{n\to+\infty} \int_{[a,b]} \theta_n\,.$

Démonstration. Soit $n \in \mathbb{N}$. Comme $\varepsilon_n > 0$, par le théorème 9 de la page 745, on peut trouver une fonction $\theta_n \in \mathcal{E}([a,b])$ telle que $\forall t \in [a,b]\,,\ \left|f(t) - \theta_n(t)\right| \leqslant \varepsilon_n$ et les fonctions $\theta_n - \varepsilon_n$ et $\theta_n + \varepsilon_n$ appartiennent respectivement à $\mathcal{E}^-(f)$ et $\mathcal{E}^+(f)$.

On a donc $\displaystyle\int_{[a,b]} (\theta_n - \varepsilon_n) \leqslant \int_{[a,b]} f \leqslant \int_{[a,b]} (\theta_n + \varepsilon_n)$, ce qui donne :

$$\left| \int_{[a,b]} f - \int_{[a,b]} \theta_n \right| \leqslant \int_{[a,b]} \varepsilon_n = \varepsilon_n\,(b-a) \underset{n\to+\infty}{\longrightarrow} 0.$$

3 Propriétés de l'intégrale

Au moyen de la proposition 11 de la page précédente, les propriétés de l'intégrale d'une fonction continue par morceaux se déduisent, par passage à la limite, de celles déjà établies pour les fonctions en escalier.

3.1 Linéarité, relation de Chasles

Proposition 12. L'intégrale est une forme linéaire sur l'espace vectoriel des fonctions continues par morceaux sur $[a,b]$.

Démonstration. Soient f et g deux fonctions continues par morceaux sur $[a,b]$, ainsi que λ et μ deux réels. Considérons (voir la proposition 11 de la page précédente) deux suites (φ_n) et (ψ_n) de fonctions en escalier telles que $|f-\varphi_n| \leqslant \varepsilon_n$, $|g-\psi_n| \leqslant \varepsilon_n$, et $\lim\limits_{n\to+\infty} \varepsilon_n = 0$. On a ainsi :

$$\int_{[a,b]} f = \lim_{n\to+\infty} \int_{[a,b]} \varphi_n \qquad \text{et} \qquad \int_{[a,b]} g = \lim_{n\to+\infty} \int_{[a,b]} \psi_n.$$

Alors, $\big|(\lambda\, f + \mu\, g) - (\lambda\, \varphi_n + \mu\, \psi_n)\big| \leqslant \big(|\lambda| + |\mu|\big)\, \varepsilon_n$ et donc, toujours par la proposition 11 :

$$\int_{[a,b]} (\lambda\, f + \mu\, g) = \lim_{n\to+\infty} \int_{[a,b]} (\lambda\, \varphi_n + \mu\, \psi_n).$$

La linéarité de l'intégrale des fonctions en escalier donne :

$$\forall n \in \mathbb{N}\,,\ \int_{[a,b]} (\lambda\, \varphi_n + \mu\, \psi_n) = \lambda \int_{[a,b]} \varphi_n + \mu \int_{[a,b]} \psi_n$$

puis $\displaystyle\int_{[a,b]} (\lambda\, f + \mu\, g) = \lambda \int_{[a,b]} f + \mu \int_{[a,b]} g$ par passage à la limite.

Exemple. Deux fonctions continues par morceaux sur $[a,b]$ qui sont égales sauf en un nombre fini de points ont la même intégrale car leur différence, qui est nulle sauf en un nombre fini de points, est une fonction en escalier dont l'intégrale est nulle.

Proposition 13 (Relation de Chasles). Soient $c \in\,]\,a,b\,[$ et f une fonction définie sur $[a,b]$.
La fonction f est continue par morceaux sur $[a,b]$ si, et seulement si, ses restrictions à $[\,a,c\,]$ et à $[\,c,b\,]$ sont continues par morceaux, et l'on a alors :

$$\int_{[a,b]} f = \int_{[\,a,c\,]} f_{|[a,c]} + \int_{[\,c,b\,]} f_{|[c,b]}.$$

Remarque. Pour simplifier, nous noterons $\displaystyle\int_{[a,c]} f$ et $\displaystyle\int_{[c,b]} f$ à la place respectivement de $\displaystyle\int_{[a,c]} f_{|[a,c]}$ et $\displaystyle\int_{[c,b]} f_{|[c,b]}$.

Démonstration.

- La démonstration de l'équivalence est similaire à celle correspondant au cas des fonctions en escalier.
- Pour montrer l'égalité, on procède de la même façon que dans la proposition précédente en considérant une suite de fonctions en escalier (φ_n) telle que :

$$\forall t \in [a,b]\,,\ \left|f(t) - \varphi_n(t)\right| \leqslant \varepsilon_n \qquad \text{avec} \qquad \lim_{n\to+\infty} \varepsilon_n = 0$$

et en passant à la limite dans la relation :

$$\int_{[a,b]} \varphi_n = \int_{[a,c]} \varphi_n + \int_{[c,b]} \varphi_n.$$

Remarque. Soit f une fonction continue par morceaux sur $[a,b]$.
Prenons $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$ une subdivision adaptée à f et notons f_i la fonction continue sur $[\,x_{i-1}, x_i\,]$ dont la restriction à $]\,x_{i-1}, x_i\,[$ est égale à celle de f.
La relation de Chasles nous donne alors :

$$\int_{[a,b]} f = \sum_{i=1}^{n} \int_{[\,x_{i-1},x_i\,]} f = \sum_{i=1}^{n} \int_{[\,x_{i-1},x_i\,]} f_i.$$

Ainsi, l'intégrale d'une fonction continue par morceaux est la somme d'intégrales de fonctions continues.

3.2 Inégalités

Proposition 14.

- Toute fonction positive et continue par morceaux a une intégrale positive.
- Si f et g sont deux fonctions continues par morceaux sur $[a,b]$ alors :

$$f \leqslant g \Longrightarrow \int_{[a,b]} f \leqslant \int_{[a,b]} g.$$

Démonstration.

- Si f est positive, alors la fonction nulle appartient à $\mathcal{E}^-(f)$ et son intégrale est plus petite que l'intégrale de f, qui est donc positive.
- Il suffit d'appliquer le résultat précédent à la fonction positive $g - f$ et d'utiliser la linéarité de l'intégrale.

Théorème 15 (Inégalité de la moyenne). Si f est continue par morceaux sur $[a,b]$, alors $|f|$ est continue par morceaux sur $[a,b]$ et :

$$\left|\int_{[a,b]} f\right| \leqslant \int_{[a,b]} |f| \leqslant (b-a) \sup_{[a,b]} |f|.$$

Démonstration. Soit $\boldsymbol{u} = (x_i)_{i\in[\![0,n]\!]}$ une subdivision adaptée à la fonction f. Alors la restriction de f à chacun des intervalles $]\,x_{i-1}, x_i\,[$ est continue et admet des limites en x_{i-1} et x_i. Il en est donc de même pour la fonction $|f|$ d'après les propriétés des limites, ce qui prouve que $|f|$ est continue par morceaux.

En intégrant la double inégalité $-|f| \leqslant f \leqslant |f|$, on obtient :

$$-\int_{[a,b]} |f| \leqslant \int_{[a,b]} f \leqslant \int_{[a,b]} |f|, \qquad \text{c'est-à-dire} \qquad \left|\int_{[a,b]} f\right| \leqslant \int_{[a,b]} |f|.$$

La dernière inégalité du théorème 15 s'obtient par intégration de la relation $|f| \leqslant \sup_{[a,b]} |f|$.

Interprétation géométrique.

De la façon dont l'intégrale a été construite (*cf.* introduction), la quantité $\int_{[a,b]} |f|$ représente la somme des aires des parties du plan comprises entre l'axe Ox et la courbe représentative de f, alors que $\int_{[a,b]} f$ représente la somme algébrique de ces aires, les parties situées sous l'axe Ox étant comptées négativement.

Si $m = \inf_{[a,b]} f$ et $M = \sup_{[a,b]} f$, alors m est une fonction en escalier inférieure à f, donc son intégrale $m(b-a)$ est inférieure à l'intégrale de f. En faisant le même type de raisonnement pour M, on obtient :

$$m(b-a) \leqslant \int_{[a,b]} f \leqslant M(b-a).$$

La quantité $\dfrac{1}{b-a}\displaystyle\int_{[a,b]} f$ est donc un réel compris entre m et M.

Définition 6. On appelle *valeur moyenne* de la fonction f continue par morceaux sur le segment $I = [a,b]$, le réel $\dfrac{1}{b-a}\displaystyle\int_{[a,b]} f$.

Exercice 1.
Soient f une fonction *continue* sur $[a,b]$ de valeur moyenne μ.
Montrer qu'il existe $c \in [a,b]$ tel que $f(c) = \mu$.
Solution. La fonction f étant continue, l'image du segment $[a,b]$ est le segment $[\min\limits_{[a,b]} f, \max\limits_{[a,b]} f]$ qui contient donc la valeur moyenne de f d'après ce qui précède.

Théorème 16 (Inégalité de Cauchy-Schwarz).
Si f et g sont deux fonctions continues par morceaux sur le segment $[a,b]$, on a :

$$\left(\int_{[a,b]} f\,g\right)^2 \leqslant \int_{[a,b]} f^2 \int_{[a,b]} g^2.$$

Démonstration. Posons :

$$P(\lambda) = \int_{[a,b]} (f + \lambda\, g)^2 = \lambda^2 \int_{[a,b]} g^2 + 2\lambda \int_{[a,b]} f\,g + \int_{[a,b]} f^2.$$

définissant ici une fonction polynomiale P de degré inférieur ou égal à 2 vérifiant :

$$\forall \lambda \in \mathbb{R}\,,\ P(\lambda) \geqslant 0.$$

- Si $\displaystyle\int_{[a,b]} g^2 = 0$, alors P est une fonction affine positive sur $\mathbb{R}$, et par suite une fonction constante. On en déduit $\displaystyle\int_{[a,b]} f\,g = 0$ et donc :

$$\left(\int_{[a,b]} f\,g\right)^2 = 0 = \int_{[a,b]} f^2 \int_{[a,b]} g^2.$$

- Sinon, P est un trinôme du second degré qui reste positif sur $\mathbb{R}$. Son discriminant :

$$\Delta = 4\left(\left(\int_{[a,b]} f\,g\right)^2 - \int_{[a,b]} f^2 \int_{[a,b]} g^2\right)$$

  est donc négatif, ce qui donne l'inégalité annoncée. ∎

3.3 Cas des fonctions continues

Théorème 17. Une fonction continue et positive sur un segment $[a,b]$ est nulle si, et seulement si, son intégrale sur $[a,b]$ est nulle.

Démonstration. Si f est nulle, alors son intégrale est évidemment nulle.
Pour la réciproque, supposons f continue, positive et non nulle.

On peut alors trouver un élément $c \in [a,b]$ tel que $f(c) > 0$.

- Si $a < c < b$: posons $\varepsilon = f(c)/2$. La continuité de f en c permet de trouver un réel $\eta > 0$ tel que :

$$\forall x \in [a,b]\,,\ |x-c| \leqslant \eta \Longrightarrow f(x) \geqslant \varepsilon.$$

  En posant $\alpha = \max\left(\dfrac{a+c}{2}, c-\eta\right)$ et $\beta = \min\left(\dfrac{b+c}{2}, c+\eta\right)$, on a :

$$\begin{aligned}
\int_{[a,b]} f &= \int_{[\,a,\alpha\,]} f + \int_{[\,\alpha,\beta\,]} f + \int_{[\,\beta,b\,]} f && \text{relation de Chasles}\\
&\geqslant \int_{[\,\alpha,\beta\,]} f && \text{puisque } f \text{ est positive}\\
&\geqslant (\beta-\alpha)\,\varepsilon > 0 && \text{puisque } \forall x \in [\,\alpha,\beta\,]\ ,\ f(x) \geqslant \varepsilon
\end{aligned}$$

  ce qui prouve que l'intégrale de f est non nulle.

- Si $c = a$ ou $c = b$: la fonction f est continue et $f(c) > 0$, ce qui prouve que f est strictement positive au voisinage de c. Il existe donc un réel $d \in\ ]\,a,b\,[$ tel que $f(d) > 0$, et l'on est ramené au cas précédent. ■

S'il manque l'une des deux hypothèses (continuité ou positivité), le résultat peut tomber en défaut.

- La fonction $\cos$ est continue et non nulle sur $[\,0,\pi\,]$ alors que son intégrale est nulle.
- La fonction définie sur $[\,0,1\,]$ par

$$f(x) = \begin{cases} 0 & \text{si } x \neq 0 \\ 1 & \text{si } x = 0 \end{cases}$$

  est positive et n'est pas la fonction nulle, mais son intégrale est nulle.

- Néanmoins, si f est positive et continue par morceaux, la démonstration précédente montre que si son intégrale est nulle, alors f ne peut pas être strictement positive en un point où elle est continue.

  Comme une fonction continue par morceaux n'a qu'un nombre fini de points de discontinuité, on en déduit qu'une fonction continue par morceaux positive qui a une intégrale nulle est nulle sauf éventuellement en un nombre fini de points.

Le théorème 17 de la page précédente permet de généraliser et de préciser le résultat de l'exercice 1 de la page précédente. C'est l'objet des deux exercices suivants.

Exercice 2 (Première formule de la moyenne).

Soient f et g deux fonctions *continues* sur $[a,b]$, g étant positive sur $[a,b]$ et non identiquement nulle. Montrer qu'il existe $d \in [a,b]$ tel que $\displaystyle\int_{[a,b]} fg = f(d)\int_{[a,b]} g$.

Solution. On note $m := [\min_{[a,b]} f$ et $M := \max_{[a,b]} f]$.

On a $mg \leqslant fg \leqslant Mg$, d'où $\displaystyle m\int_{[a,b]} g \leqslant \int_{[a,b]} fg \leqslant M\int_{[a,b]} g$ et, puisque, d'après le théorème 17 ci-dessus, $\displaystyle\int_{[a,b]} g > 0$, on en déduit $\displaystyle m \leqslant \frac{\int_{[a,b]} fg}{\int_{[a,b]} g} \leqslant M$.

Il existe donc $d \in [a,b]$ tel que $\displaystyle f(d) = \frac{\int_{[a,b]} fg}{\int_{[a,b]} g}$.

Exercice 3.

1. Montrer que dans l'exercice 1 de la page 751, le réel c peut être choisi dans l'intervalle *ouvert* $]a,b[$.
2. Dans l'exercice 2 ci-dessus, on suppose que, pour tout $x \in]a,b[$, $g(x) > 0$. Montrer qu'alors le réel d peut être choisi dans l'intervalle *ouvert* $]a,b[$.

Solution.

1. Soit f continue sur $[a,b]$. Dans l'exercice précité, nous avons montré l'existence d'un élément $c \in [a,b]$ tel que $\displaystyle\int_{[a,b]} f = (b-a)\,f(c)$. Si $c \in]a,b[$, il n'y a rien à montrer.

   Sinon, supposons par exemple $c = a$. Alors, ou bien, il existe $c' \in]a,b[$ tel que $f(c') = f(a)$ et le résultat est démontré, ou bien $\forall x \in]a,b[,\ f(x) \neq f(a)$. Dans ce cas la fonction $f - f(a)$ est continue et ne s'annule pas sur l'intervalle $]a,b[$, donc elle y garde un signe constant. Par continuité, elle est aussi de signe constant sur $[a,b]$. Elle n'est pas identiquement nulle, donc, d'après le théorème 17, son intégrale n'est pas nulle, ce qui contredit l'égalité $\displaystyle\int_{[a,b]} f = (b-a)\,f(a)$. Donc il existe $x \in]a,b[$ tel que $f(x) = f(a)$, ce qui prouve le résultat.
2. On reprend mutatis mutandis la preuve ci-dessus en notant que, si $x \in]a,b[$, alors $f(x) \neq f(a)$ si, et seulement si, $f(x)g(x) \neq f(a)g(x)$.

Corollaire 18. Si f et g sont deux fonctions continues sur le segment $[a,b]$, on a :

$$\left(\int_{[a,b]} f\,g\right)^2 = \int_{[a,b]} f^2 \int_{[a,b]} g^2$$

si, et seulement si, f et g sont proportionnelles.

Démonstration.

- Si f et g sont proportionnelles, il existe un réel λ tel que $f = \lambda\,g$ ou $g = \lambda\,f$. Supposons par exemple $f = \lambda\,g$.
  Alors :
  $$\left(\int_{[a,b]} f\,g\right)^2 = \lambda^2 \left(\int_{[a,b]} g^2\right)^2 = \int_{[a,b]} f^2 \int_{[a,b]} g^2.$$
- Supposons que l'égalité ait lieu et reprenons le polynôme P de la démonstration de l'inégalité de Cauchy–Schwarz :
  $$P(\lambda) = \int_{[a,b]} (f + \lambda\,g)^2 = \lambda^2 \int_{[a,b]} g^2 + 2\lambda \int_{[a,b]} f\,g + \int_{[a,b]} f^2.$$
  - Si $\displaystyle\int_{[a,b]} g^2 = 0$, alors g^2 est continue positive et a une intégrale nulle. La fonction g est donc nulle et f et g sont proportionnelles.
  - Sinon, le polynôme P a un discriminant nul, et l'on peut donc trouver un réel λ_0 tel que $P(\lambda_0) = 0$.
    La fonction $(f + \lambda_0\,g)^2$ est alors continue, positive et d'intégrale nulle. Par suite $f + \lambda_0\,g = 0$, ce qui prouve que f et g sont proportionnelles. ■

3.4 Sommes de Riemann

Définition 7. On appelle *somme de Riemann régulière à gauche* associée à f, à l'ordre n, sur l'intervalle $[a,b]$, la quantité :

$$S_n(f;[a,b]) := \frac{b-a}{n} \sum_{k=0}^{n-1} f\Big(a + \frac{k}{n}(b-a)\Big).$$

La *somme de Riemann régulière à droite* est obtenue en sommant k de 1 à n.

La somme de Riemann ainsi définie n'est pas autre chose que l'intégrale d'une fonction en escalier égale à $f(x_{k-1,n})$ (ou $f(x_{k,n})$ pour les sommes de Riemann régulières à droite) sur chacun des intervalles $]x_{k-1,n}, x_{k,n}[$, avec $x_{k,n} := a + k\,\frac{b-a}{n}$, pour $k \in [\![0,n]\!]$.
On « voit » donc qu'elles vont approcher l'intégrale de f d'autant mieux lorsque n augmente (voir la figure IV.5.1 de la page ci-contre).
On peut définir des sommes de Riemann plus générales attachées à la donnée d'une subdivision « pointée ». Soit $\boldsymbol{u} = (x_k)_{k\in[\![0,n]\!]}$ une subdivision de $[a,b]$. Pour tout $k \in [\![1,n]\!]$, on choisit $\xi_k \in [x_{k-1}, x_k]$ et l'on considère $\sum\limits_{k=1}^{n} f(\xi_k)(x_k - x_{k-1})$. Nous étudierons ces sommes de Riemann dans le livre L2.

FIGURE IV.5.1 – Sommes de Riemann régulières à gauche

Théorème 19. Pour toute fonction f continue par morceaux sur $[a,b]$, la suite $\big(S_n(f;[a,b])\big)$ converge vers l'intégrale de f sur $[a,b]$.

Ce théorème sera démontré dans le cas général (*cf.* L2), mais nous en proposons néanmoins ici une preuve dans le cas particulier des fonctions continues. Elle peut être sautée dans une première lecture.

Démonstration. Intéressons-nous, par exemple, aux sommes de Riemann régulières à gauche et notons φ_n la fonction en escalier égale à $f(x_{k-1,n})$ sur l'intervalle $]x_{k-1,n}, x_{k,n}[$, avec $x_{k,n} := a + k\,\frac{b-a}{n}$, pour $k \in [\![0,n]\!]$ et telle que $\varphi_n(x_{k,n}) = f(x_{k,n})$.

Soit $\varepsilon > 0$. Montrons par l'absurde qu'il existe un rang p tel que, pour tout $n \geqslant p$, on ait $\forall t \in [a,b]\,,\ \big|f(t) - \varphi_n(t)\big| \leqslant \varepsilon$. Supposons que ce ne soit pas le cas, on a donc :

$$\forall p \in \mathbb{N}^*\,,\ \exists n \geqslant p\,,\ \exists t \in [a,b]\ :\ \big|f(t) - \varphi_n(t)\big| > \varepsilon.$$

Prenons, pour chaque entier p un tel n, que l'on note n_p et un élément $t_p \in [a,b]$ tel que $\big|f(t_p) - \varphi_{n_p}(t_p)\big| > \varepsilon$. La suite (t_p) étant bornée, elle admet une suite extraite convergente et donc, quitte à remplacer (t_p) par une telle sous-suite et (φ_{n_p}) par la sous-suite correspondante, on peut supposer que la suite (t_p) converge vers un certain élément $c \in [a,b]$. Par définition, $\varphi_{n_p}(t_p)$ est la valeur de f en un point x_p tel que $x_p \leqslant t_p \leqslant x_p + \frac{1}{n_p}$, et donc (x_p) converge aussi vers c. Par continuité de f en c, on a ainsi $\lim\limits_{p\to+\infty} f(x_p) - f(t_p) = f(c) - f(c) = 0$, ce qui contredit la relation $\big|f(x_p) - f(t_p)\big| = \big|f(t_p) - \varphi_{n_p}(t_p)\big| > \varepsilon$.

Ainsi, il existe un entier p tel que, pour tout $n \geqslant p$, on ait $\forall t \in [a,b]\,,\ \big|f(t) - \varphi_n(t)\big| \leqslant \varepsilon$ et par conséquent :

$$\left|\int_{[a,b]} f - \int_{[a,b]} \varphi_n\right| \leqslant \int_{[a,b]} \big|f - \varphi_n\big| \leqslant \int_{[a,b]} \varepsilon = (b-a)\,\varepsilon.$$

On a ainsi montré la convergence de $\displaystyle\int_{[a,b]} \varphi_n$ vers $\displaystyle\int_{[a,b]} f$. ∎

Exercice 4.
Calculer $\displaystyle\int_0^{\pi} \sin(x)\,\mathrm{d}x$ en utilisant des sommes de Riemann.

Solution. La somme de Riemann régulière à gauche à l'ordre n associée à la fonction sinus est $S_n := \dfrac{\pi}{n}\displaystyle\sum_{k=0}^{n-1} \sin\left(\frac{k\pi}{n}\right)$.

La somme $\displaystyle\sum_{k=0}^{n-1} \sin\left(\frac{k\pi}{n}\right)$ est la partie imaginaire de :

$$\sum_{k=0}^{n-1} \mathrm{e}^{ik\pi/n} = \frac{1-\mathrm{e}^{i\pi}}{1-\mathrm{e}^{i\pi/n}} = \frac{2}{\mathrm{e}^{i\pi/2n}\,(-2i)\,\sin(\pi/2n)} = \frac{i\,\mathrm{e}^{-i\pi/2n}}{\sin(\pi/2n)},$$

ce qui donne $S_n = \dfrac{\pi}{n}\dfrac{\cos(\pi/2n)}{\sin(\pi/2n)}$.

En utilisant $\displaystyle\lim_{x\to 0}\frac{\sin x}{x} = 1$, on obtient $\displaystyle\lim_{n\to+\infty}\frac{\pi/n}{\sin(\pi/2n)} = 2$, d'où $\displaystyle\int_0^{\pi}\sin(x)\,\mathrm{d}x = 2$, résultat que l'on retrouvera par la suite en utilisant le calcul intégral (voir page 759).

4 Intégration et dérivation, calcul des intégrales

Dans cette section, I désigne un intervalle quelconque de $\mathbb{R}$.

4.1 Fonctions continues par morceaux sur un intervalle

Définition 8. On dit qu'une application définie sur I est *continue par morceaux* sur I si elle est continue par morceaux sur tout segment de I.

Remarque. On peut donc intégrer une fonction continue par morceaux sur I sur tout segment inclus dans I et contenant au moins deux points.

Exemples.

1. Toute fonction continue sur I est continue par morceaux sur I.
2. La fonction $x \mapsto x - E(x)$ est continue par morceaux sur $\mathbb{R}$.
3. La fonction f définie sur $\mathbb{R}$ par :

$$f(0) = 0 \qquad \text{et} \qquad \forall x \neq 0\,,\ f(x) = \mathrm{e}^{1/x}$$

n'est pas continue par morceaux sur $\mathbb{R}$ puisque sa restriction à $[-1,1]$ n'est pas continue par morceaux (voir l'exemple 3 de la page 744).

En revanche, ses restrictions à $\mathbb{R}_+^*$ et $\mathbb{R}_-^*$ sont continues par morceaux puisqu'elles sont continues.

Notation. Soient f une fonction continue par morceaux sur un intervalle I, ainsi que a et b deux éléments quelconques de I (on ne suppose plus à partir de maintenant $a < b$). On note :

- si $a < b$, $\displaystyle\int_a^b f(x)\,\mathrm{d}x = \int_{[a,b]} f$

- si $a > b$, $\displaystyle\int_a^b f(x)\ \mathrm{d}x = -\int_{[b,a]} f$
- si $a = b$, $\displaystyle\int_a^b f(x)\ \mathrm{d}x = 0$

Un peu d'histoire

Au XVIIème siècle, Leibniz interprétait l'intégrale, et donc l'aire sous la courbe, comme une somme infinie de rectangles infinitésimaux de hauteur $f(x)$ et de largeur infiniment petite Δx, et donc d'aire $f(x)\,\Delta x$. Le produit $f(x)\ \mathrm{d}x$ rappelle cette notation et le symbole d'intégrale le « S » de la somme.
À l'époque moderne, on reprend cette idée en introduisant la notion de *forme différentielle* : $f(x)\,\mathrm{d}x$ représente une forme différentielle que l'on intègre entre les deux points a et b. Sur le plan pratique la notation $\displaystyle\int_a^b f(x)\,\mathrm{d}x$ facilite le changement de variable dans une intégrale.

L'implication $f \leqslant g \Longrightarrow \displaystyle\int_a^b f(x)\ \mathrm{d}x \leqslant \int_a^b g(x)\,\mathrm{d}x$
n'est valable que lorsque $a \leqslant b$.

Dans le cas général, les inégalités du théorème 15 de la page 750 deviennent :

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \leqslant \int_J |f| \leqslant |b-a| \sup_J |f|$$

où J désigne le segment de $\mathbb{R}$ d'extrémités a et b, c'est-à-dire $[a,b]$ si $a \leqslant b$ et $[b,a]$ sinon.

Proposition 20 (Relation de Chasles).
Si f est continue par morceaux sur un intervalle I, alors :

$$\forall (a,b,c) \in I^3,\ \int_a^b f(x)\ \mathrm{d}x = \int_a^c f(x)\ \mathrm{d}x + \int_c^b f(x)\ \mathrm{d}x.$$

Démonstration.

- Le cas où $a < c < b$ est, avec la nouvelle notation, la traduction de la proposition 13.
- Si $a < b < c$, la même proposition 13 de la page 748 permet d'écrire :

$$\int_a^c f(x)\ \mathrm{d}x = \int_a^b f(x)\ \mathrm{d}x + \int_b^c f(x)\ \mathrm{d}x$$

  et donc :

$$\int_a^b f(x)\ \mathrm{d}x = \int_a^c f(x)\ \mathrm{d}x - \int_b^c f(x)\ \mathrm{d}x = \int_a^c f(x)\ \mathrm{d}x + \int_c^b f(x)\ \mathrm{d}x.$$

- Les cas :

$$b < a < c \qquad b < c < a \qquad c < a < b \qquad c < b < a$$

  se traitent de la même façon.
- Enfin, si a, b et c ne sont pas deux à deux distincts, le résultat est immédiat. ■

4.2 Le théorème fondamental de l'analyse

L'intégrale est bien plus qu'un concept géométrique, comme on a commencé à le voir. Pour le réaliser pleinement, il faut faire le lien (pas apparent du tout) avec la dérivation, ce que nous allons voir ci-après. Rappelons d'abord une définition reliée à la dérivation :

Définition 9. On dit qu'une fonction f admet une primitive sur un intervalle I s'il existe une application dérivable $F : I \to \mathbb{R}$ (appelée *primitive* de f) telle que $F'(x) = f(x)$ pour tout $x \in I$.

Rappelons aussi qu'en vertu de l'inégalité des accroissements finis, deux primitives d'une même fonction sur un intervalle diffèrent d'une constante.
Quel rapport y a-t-il entre le concept géométrique de droite tangente à une courbe, et le calcul d'aire ? Apparemment aucun, et si l'on s'en tenait à la vision purement géométrique de l'intégrale, on en resterait là. Le théorème qui suit, et qui est justement appelé fondamental, va pourtant nous donner un lien absolument inattendu : le concept de tangente est relié à celui de dérivée (qui est la pente de la tangente), et celui d'aire à l'intégrale, comme expliqué à la section précédente. Or le théorème affirme que l'intégration est l'opération contraire de la dérivation (on parle parfois d'antidérivation) :

Théorème 21 (Cauchy). Toute fonction continue sur un intervalle I admet une primitive. Plus précisément, la primitive de f qui s'annule en un point $a \in I$ est l'application $F : I \to \mathbb{R}$ définie par $F(x) := \displaystyle\int_a^x f(t)\,\mathrm{d}t$.

Démonstration. Soit $c \in I$; par continuité de f, pour tout $\varepsilon > 0$ il existe δ tel que $\big|f(x) - f(c)\big| \leqslant \varepsilon$ dès que $x \in I$ et $|x - c| \leqslant \delta$. Prenons un tel x et posons $h := x - c$.
Il vient alors, en notant J le segment d'extrémités c et $c + h$:

$$\big|F(c+h) - F(c) - h\,f(c)\big| = \Big|\int_c^{c+h} \big(f(t) - f(c)\big)\,\mathrm{d}t\Big| \leqslant \int_J \big|f(t) - f(c)\big|\,\mathrm{d}t \leqslant |h|\,\varepsilon$$

d'où la dérivabilité de F en c avec $F'(c) = f(c)$. ∎

Remarque. En utilisant la relation de Chasles, on voit que $x \mapsto \displaystyle\int_b^x f(t)\,\mathrm{d}t$, avec $b \in I$ quelconque, est aussi une primitive de f, car sa différence avec F est constante. On prendra néanmoins garde que toutes les primitives de f ne sont pas, *a priori*, de cette forme. Mais elles sont toutes de la forme $\lambda + F$, avec $\lambda \in \mathbb{R}$, car I est un intervalle.

En raison du théorème de Cauchy, on note parfois la primitive d'une fonction f sous la forme $x \mapsto \displaystyle\int f(x)\,\mathrm{d}x$ (sans bornes sur l'intégrale) ou sous la forme $x \mapsto \displaystyle\int^x f(t)\,\mathrm{d}t$ (sans borne basse, supposée connue, mais non précisée).

Exercice 5.
Redémontrer, en utilisant le théorème de Cauchy, le résultat du théorème 17 de la page 751.
Solution. Soit f une fonction continue et positive sur un segment $[a,b]$. La fonction $F : x \mapsto \displaystyle\int_a^x f(t)\,\mathrm{d}t$ en est alors une primitive et elle est croissante, par positivité de sa dérivée f. Donc, si $\displaystyle\int_a^b f(t)\,\mathrm{d}t = 0$, alors $F(b) - F(a) = 0$ et la fonction croissante F est constante sur $[a,b]$. Par conséquent, sa dérivée est nulle, ce qui donne $f = 0$.

Théorème 22 (Théorème fondamental de l'analyse). Soit f une fonction continûment dérivable sur un intervalle I. Alors, pour tous α, β dans I, on a :

$$\int_\alpha^\beta f'(x)\,\mathrm{d}x = f(\beta) - f(\alpha).$$

Le membre de droite est appelé la variation de f entre α et β et est souvent noté $[f(x)]_\alpha^\beta$ ou $[f]_\alpha^\beta$.

Démonstration. Le théorème de Cauchy appliqué à f' (supposée continue), nous dit que cette dernière admet f et $G : x \mapsto \displaystyle\int_a^x f'(t)\,\mathrm{d}t$ comme primitives. Elles diffèrent donc d'une constante, ce qui donne :

$$f(\beta) - f(\alpha) = G(\beta) - G(\alpha) = \int_\alpha^\beta f'(t)\,\mathrm{d}t$$

et justifie le théorème fondamental. ∎

Ce théorème nous permet donc de calculer certaines intégrales, si l'on parvient à identifier l'intégrande (la fonction que l'on intègre) comme une dérivée.

Corollaire 23. Soit f continue sur un intervalle I. Si F est une primitive de f, alors :

$$\forall (a,b) \in I^2\,,\ \int_a^b f(t)\,\mathrm{d}t = F(b) - F(a).$$

Exemple. Ainsi, l'exercice 4 de la page 756 est-il maintenant bien plus simple à résoudre, la fonction $-\cos$ étant une primitive de $\sin$:

$$\int_0^\pi \sin(x)\,\mathrm{d}x = \big[-\cos(x)\big]_0^\pi = -\cos(\pi) + \cos(0) = 2.$$

On peut utiliser un tableau de dérivation pour retrouver certaines primitives importantes, en lisant son contenu « à l'envers ». Nous présentons dans la table de la page suivante les primitives déduites directement des formules de dérivation ou des définitions et propriétés

élémentaires des fonctions transcendantes.

Fonction	*Primitive*	*Fonction*	*Primitive*
$x^\alpha \qquad \alpha \in \mathbb{R} \setminus \{-1\}$	$\dfrac{x^{\alpha+1}}{\alpha+1}$	$\cosh x$	$\sinh x$
$\dfrac{1}{x}$	$\ln\lvert x\rvert$	$\sinh x$	$\cosh x$
e^x	e^x	$\tanh x$	$\ln\cosh x$
$\ln x$	$x\,\ln x - x$	$\coth x$	$\ln\lvert\sinh x\rvert$
$\cos x$	$\sin x$	$\dfrac{1}{a^2+x^2} \qquad a \neq 0$	$\dfrac{1}{a}\arctan\left(\dfrac{x}{a}\right)$
$\sin x$	$-\cos x$	$\dfrac{1}{a^2-x^2} \qquad a \neq 0$	$\dfrac{1}{2a}\ln\left\lvert\dfrac{a+x}{a-x}\right\rvert$
$\tan x$	$-\ln\lvert\cos x\rvert$	$\dfrac{1}{\sqrt{a^2-x^2}} \qquad a > 0$	$\arcsin\left(\dfrac{x}{a}\right)$
$\operatorname{cotan} x$	$\ln\lvert\sin x\rvert$	$\dfrac{1}{\sqrt{x^2+h}} \qquad h \neq 0$	$\ln\bigl(x+\sqrt{x^2+h}\bigr)$

TABLE IV.5.1 – Primitives de fonctions élémentaires

Notons d'une part que certaines fonctions manquent (les fonctions $\arcsin$ et $\arctan$, par exemple ne figurent pas dans la colonne de gauche, car on ne reconnaît pas immédiatement la dérivée d'une fonction connue). Il en est de même pour la fonction logarithme, mais on peut vérifier par simple dérivation que la fonction proposée en est bien une primitive. Nous verrons dans l'exercice 8 de la page 763 comment, par intégration par parties, trouver une telle primitive, méthode qui peut être adaptée pour les fonctions $\arctan$ et $\arcsin$ (voir les exercices 9 et 10 à la page 763).
D'autre part, nous donnons seulement une primitive : les autres se déduisent en ajoutant une constante sur chaque intervalle de définition de la fonction qu'il faut bien délimiter.
Par exemple, la fonction $x \mapsto 1/x$ est définie et continue sur $\mathbb{R}^*$ qui n'est pas un intervalle mais la réunion de deux demi-droites ; la fonction $x \mapsto \ln\lvert x\rvert$ en est une primitive à la fois sur $\mathbb{R}_+^*$ et sur $\mathbb{R}_-^*$; on dit parfois – par abus de langage – qu'elle en est aussi une primitive sur $\mathbb{R}^*$. Plus précisément toutes ses primitives sont de la forme $x \mapsto a + \ln x$ sur $\mathbb{R}_+^*$ et $x \mapsto b + \ln(-x)$ sur $\mathbb{R}_-^*$ où a et b sont arbitraires.

5 Méthodes de calcul d'intégrales

Dans bien des situations, on cherche à intégrer une fonction, souvent donnée par une expression explicite. Lorsque cette dernière est une combinaison linéaire d'autres fonctions, on est ramené au calcul des intégrales de celles-ci (si elles existent). Cela permet notamment d'intégrer des sommes ou des différences, mais qu'en est-il des trois autres opérations importantes sur les fonctions : la multiplication, la composition et la division ?

La réponse courte est : en général, on ne peut pas. Même des fonctions apparemment très simples comme $x \mapsto \mathrm{e}^{x^2}$ n'ont pas de primitives exprimables avec les autres fonctions usuelles. Attention : il y a quand même une primitive, car il s'agit d'une fonction continue, mais celle-ci ne peut s'écrire à l'aide des fonctions transcendantes que nous avons vues. Dans ce cas, on peut introduire de nouvelles fonctions à cette fin. Ainsi, si nous considérons la classe des fractions rationnelles (quotient de deux polynômes), certaines n'ont pas de primitives dans cette classe, comme par exemple $x \mapsto 1/(1+x^2)$. On peut alors définir de nouvelles fonctions (l'arc-tangente dans notre exemple) pour enrichir cet ensemble (c'est l'une des définitions classiques de cette fonction, même si ce n'est pas ainsi que nous l'avons introduite dans ce cours). Mais la nouvelle classe contiendra alors d'autres fonctions dont les primitives ne s'expriment pas dans cette classe. Un théorème (difficile) montre même que, quel que soit le nombre (fini) de nouvelles fonctions que nous introduisons, on aura toujours une classe « incomplète pour l'intégration ».

La réponse longue est : dans certains cas, on peut quand même faire le calcul, en utilisant deux outils essentiels que nous allons développer ci-dessous, l'intégration par parties, et les changements de variables.

5.1 Intégration par parties

Théorème 24. Soient f, g deux fonctions continûment dérivables sur $[a,b]$. Alors :

$$\int_a^b f(x)g'(x)\,\mathrm{d}x = \big[f(x)g(x)\big]_a^b - \int_a^b f'(x)g(x)\,\mathrm{d}x.$$

On rappelle que les crochets désignent la variation de fg, soit ici $f(b)g(b) - f(a)g(a)$.

Démonstration. Le produit $h := fg$ est continûment dérivable de dérivée $h' = fg' + f'g$. On a donc, d'après le théorème fondamental :

$$\big[h(x)\big]_a^b = \int_a^b h'(x)\,\mathrm{d}x = \int_a^b f(x)g'(x)\,\mathrm{d}x + \int_a^b f'(x)g(x)\,\mathrm{d}x$$

ce qui donne le résultat. ∎

La formule d'intégration par parties permet de transformer une intégrale d'un produit fg' en une autre (celle de gf'). Donnons quatre exemples.

Exercice 6.

Calculer $\displaystyle\int_0^1 x\,\mathrm{e}^x\,\mathrm{d}x$.

Solution. Il suffit de poser $f(x) := x$ et $g'(x) := \mathrm{e}^x$, d'où par exemple $g(x) := \mathrm{e}^x$. Par suite :

$$\int_0^1 x\,\mathrm{e}^x\,\mathrm{d}x = [x\,\mathrm{e}^x]_0^1 - \int_0^1 \mathrm{e}^x\,\mathrm{d}x = \mathrm{e} - (\mathrm{e}-1) = 1.$$

Le succès vient du fait que la deuxième intégrale est maintenant une intégrale « simple » (primitive évidente). Cette méthode peut se généraliser en effectuant plusieurs intégrations par parties successives, comme indiqué dans l'exercice suivant.

Exercice 7.

Soit P un polynôme de degré inférieur ou égal à n.

1. Montrer que l'on peut calculer $\displaystyle\int_a^b P(x)\,\mathrm{e}^x\,\mathrm{d}x$ en effectuant au plus n intégrations par parties. En déduire que toute primitive de $P(x)\,\mathrm{e}^x$ est de la forme $Q(x)\,\mathrm{e}^x + c$, où Q est un polynôme.
2. Même question avec $P(x)\,\cos x$, $P(x)\,\sin x$, $P(x)\,\cosh x$ et $P(x)\,\sinh x$. Dans ces cas, quelle est la forme des primitives ?

Solution.

1. Démontrons, par récurrence sur $n := \deg P$, qu'une des primitives de $x \mapsto P(x)\,\mathrm{e}^x$ est de la forme $x \mapsto Q(x)\,\mathrm{e}^x$ où Q est un polynôme de degré au plus n. C'est clair pour $n = 0$; si c'est vrai pour $n \in \mathbb{N}$, c'est alors vrai pour $n+1$ puisque :
$$\int_a^b P(x)\,\mathrm{e}^x\,\mathrm{d}x = [P(x)\,\mathrm{e}^x]_a^b - \int_a^b P'(x)\,\mathrm{e}^x\,\mathrm{d}x.$$
Une suite de n intégrations par parties de ce genre conduit donc à une expression sans signe intégral.
2. Démontrons, par récurrence sur le degré n de P, qu'une des primitives de $x \mapsto P(x)\,\cos x$ et une des primitives de $x \mapsto P(x)\,\sin x$ s'écrivent toutes deux sous la forme $Q(x)\,\cos x + R(x)\,\sin x$ où Q et R sont des polynômes de degré au plus n. C'est clair pour $n = 1$; si c'est vrai pour $n > 0$, c'est alors vrai pour $n+1$ puisque :
$$\int_a^b P(x)\,\cos x\,\mathrm{d}x = [P(x)\,\sin x]_a^b - \int_a^b P'(x)\,\sin x\,\mathrm{d}x.$$
$$\int_a^b P(x)\,\sin x\,\mathrm{d}x = -\,[P(x)\,\cos x]_a^b + \int_a^b P'(x)\,\cos x\,\mathrm{d}x.$$
De la même façon, l'une des primitives de $x \mapsto P(x)\,\cosh x$ et l'une des primitives de $P(x)\,\sinh x$ s'écrivent sous la forme $x \mapsto Q(x)\,\cosh x + R(x)\,\sinh x$ où Q et R sont des polynômes de degré au plus égal à celui de P (on peut aussi introduire e^x et e^{-x}).

Exercice 8.

On cherche une primitive de la fonction logarithme sur son intervalle $\mathbb{R}_+^* =]0, +\infty[$ de définition. Nous savons que $t \mapsto \displaystyle\int_1^t \ln x \, \mathrm{d}x$ est une telle primitive. Calculer cette intégrale.

Solution. Il suffit de la considérer comme le produit de $f(x) := \ln x$ par la fonction constante $g'(x) := 1$, d'où par exemple $g(x) := x$. On a alors $f'(x) = 1/x$, d'où :

$$\int_1^t \ln x \, \mathrm{d}x = [fg]_1^t - \int_1^t x \, \frac{1}{x} \, \mathrm{d}x = (t \ln t - 0) - \int_1^t \mathrm{d}x = t \ln t - (t-1) = t\,(\ln t - 1) + 1$$

et toutes les primitives de $t \mapsto \ln t$ sont donc de la forme $t \mapsto t \, \ln t - t + c$ avec $c \in \mathbb{R}$.

Exercice 9.

On cherche une primitive de la fonction arc-tangente sur $\mathbb{R}$.

Nous savons que $t \mapsto \displaystyle\int_0^t \arctan x \, \mathrm{d}x$ est une telle primitive. Calculer cette intégrale.

Solution. Il suffit de la considérer comme le produit de $f(x) := \arctan x$ par la fonction constante $g'(x) := 1$, d'où par exemple $g(x) := x$. On a alors $f'(x) = 1/(1+x^2)$, d'où :

$$\int_1^t \arctan x \, \mathrm{d}x = [fg]_1^t - \int_1^t x \, \frac{1}{1+x^2} \, \mathrm{d}x.$$

Puisque x est la demi-dérivée de $1 + x^2$, nous pouvons écrire :

$$\int_0^t \frac{x}{1+x^2} \, \mathrm{d}x = \left[\frac{1}{2} \, \ln(1+x^2)\right]_0^t = \frac{1}{2} \, \ln(1+t^2) = \ln \sqrt{1+t^2}$$

et les primitives de $t \mapsto \arctan t$ sont de la forme $t \mapsto t \arctan t - \ln \sqrt{1+t^2} + c$ avec $c \in \mathbb{R}$.

Exercice 10.

Donner une primitive de la fonction arc-sinus sur $]-1, 1[$.

Solution. Pour $x \in]-1, 1[$, écrivons $\arcsin x = f(x)\, g'(x)$, avec $f(x) := \arcsin x$ et $g(x) := x$. On a alors $f'(x) = 1/\sqrt{1-x^2}$, d'où :

$$\begin{aligned}\int_0^t \arcsin x \, \mathrm{d}x &= [fg]_0^t - \int_0^t \frac{x}{\sqrt{1-x^2}} \, \mathrm{d}x \\ &= t \, \arcsin t + \sqrt{1-t^2}.\end{aligned}$$

Les primitives de $t \mapsto \arcsin t$ sur $]-1, 1[$ sont de la forme $t \mapsto t \, \arcsin t + \sqrt{1-t^2} + c$ avec $c \in \mathbb{R}$.

Remarque. La fonction $F : t \mapsto t \, \arcsin t + \sqrt{1-t^2}$ est continue sur $[-1, 1]$ et est une primitive de $f : t \mapsto \arcsin t$ sur $]-1, 1[$. Or, f est continue sur $[-1, 1]$, donc admet des primitives sur ce segment. Notons G celle de ses primitives qui s'annule en 0. Comme $F(0) = 0$, les fonctions F et G coïncident sur $]-1, 1[$. Par continuité, elles sont donc égales sur $[-1, 1]$, ce qui prouve que F est une primitive de f sur $[-1, 1]$ et en particulier dérivable en ± 1, ce qui n'est pas évident au vu de son expression.

5.2 Changement de variable

Théorème 25. Soit φ une fonction continûment dérivable sur un intervalle I, à valeurs dans un intervalle J. Si a et b sont deux points de I et f une fonction continue sur J, alors :

$$\int_{\varphi(a)}^{\varphi(b)} f(x)\ \mathrm{d}x = \int_a^b \big(f \circ \varphi(t)\big)\, \varphi'(t)\, \mathrm{d}t.$$

Démonstration. Soient F une primitive de f sur J et $G := F \circ \varphi$. Alors G est bien définie et dérivable, de dérivée $G' = \varphi'(F' \circ \varphi) = (f \circ \varphi)\,\varphi'$. On a donc, d'après le théorème fondamental :

$$F\big(\varphi(b)\big) - F\big(\varphi(a)\big) = G(b) - G(a) = \int_a^b G'(t)\,\mathrm{d}t = \int_a^b \big(f \circ \varphi(t)\big)\, \varphi'(t)\,\mathrm{d}t$$

ce qui donne le résultat. ∎

Remarque. Soit θ une fonction continûment dérivable sur un intervalle J et telle que $\forall x \in J\,,\ \theta'(x) \neq 0$. Alors θ définit une bijection de J sur un intervalle I dont la réciproque est aussi continûment dérivable d'après le théorème 56 de la page 657. On peut donc appliquer le théorème de changement de variable à $\varphi = \theta^{-1}$. Ainsi, si α et β sont deux points de J et f une fonction continue sur J, alors :

$$\int_\alpha^\beta f(x)\,\mathrm{d}x = \int_{\theta(\alpha)}^{\theta(\beta)} f\big(\theta^{-1}(t)\big)\,(\theta^{-1})'(t)\ \mathrm{d}t.$$

Ces deux formules sont appelées « formules de changement de variable ».
Un moyen mnémotechnique, remontant à Leibniz, consiste à poser $x := \varphi(t)$ $\big($respectivement $x := \theta^{-1}(t)$, c'est-à-dire $t := \theta(x)\big)$ et à remplacer systématiquement x par sa valeur en fonction de t et $\mathrm{d}x$ par $\varphi'(t)\,\mathrm{d}t$ $\big($respectivement $(\theta^{-1})'(t)\,\mathrm{d}t\big)$.
Un cas simple, mais important en pratique, est celui des changements de variables affines, c'est-à-dire de la forme $x := \gamma t + \delta$ avec $\gamma \neq 0$, d'où $t = (x - \delta)/\gamma$. On obtient alors :

$$\int_\alpha^\beta f(x)\ \mathrm{d}x = \gamma \int_{\frac{\alpha-\delta}{\gamma}}^{\frac{\beta-\delta}{\gamma}} f(\gamma t + \delta)\ \mathrm{d}t.$$

Par exemple : $\displaystyle\int_0^4 \cos(3x+1)\ \mathrm{d}x = \frac{1}{3}\int_1^{13} \cos t\ \mathrm{d}t = \frac{\sin 13 - \sin 1}{3}\cdot$

Les deux formules ne s'utilisent pas exactement de la même façon en pratique.
Notons qu'en général, on les applique toutes deux de droite à gauche.

Exercice 11.

Calculer l'intégrale $I = \displaystyle\int_0^{4\pi} \cos t\, \mathrm{e}^{\sin t}\,\mathrm{d}t$.

Solution. Nous « reconnaissons » une formule semblable au membre de droite de la première formule de changement de variable avec $\varphi := \sin$ (d'où $\varphi' = \cos$) et $f := \exp$ pour $a := 0$ et $b := 4\pi$, continue sur l'image $[-1, 1]$ de la fonction $\sin$. On a donc :

$$I = \int_{\varphi(a)}^{\varphi(b)} f = \int_{\sin 0}^{\sin 4\pi} \exp = \int_0^0 \exp = 0.$$

Exercice 12.

Calculer l'aire d'un disque de rayon $R > 0$.

Solution. On se place dans un système de coordonnées centré sur le centre du disque, qui a donc pour équation $x^2 + y^2 \leqslant R^2$. Si nous ne conservons que les valeurs positives de y, on obtiendra un demi-disque, ensemble des points :

$$\{(x, y) \in \mathbb{R}^2 \mid -R \leqslant x \leqslant R \qquad \text{et} \qquad 0 \leqslant y \leqslant \sqrt{R^2 - x^2}\}.$$

C'est la surface sous la courbe de la fonction $x \mapsto f(x) := \sqrt{R^2 - x^2}$ dans l'intervalle $[-R, R]$. Comme c'est une fonction positive, l'aire correspondante A est égale à l'intégrale $\displaystyle\int_{-R}^{R} \sqrt{R^2 - x^2}\ \mathrm{d}x$.

Pour calculer celle-ci, effectuons le changement de variable $x = R\cos t$. On souhaite que x varie dans l'intervalle $[-R, R]$, on prend donc t variant dans l'intervalle $[0, \pi]$ par exemple. Là, la fonction $\varphi : t \mapsto R\cos t$ est continûment dérivable sur $[0, \pi]$. De plus $\sqrt{R^2 - x^2} = R\sin t$ car le sinus est positif ou nul sur $[0, \pi]$. On a donc, en effectuant ce changement de variable et en utilisant la formule de trigonométrie $\cos 2t = 1 - 2\sin^2 t$:

$$\begin{aligned} A = \int_{-R}^{R} \sqrt{R^2 - x^2}\ \mathrm{d}x &= \int_{\pi}^{0} (R\sin t)(-R\sin t)\ \mathrm{d}t \\ &= R^2 \int_0^{\pi} \frac{1 - \cos 2t}{2}\ \mathrm{d}t = R^2 \left[\frac{t}{2} - \frac{\sin 2t}{4}\right]_0^{\pi} = \frac{\pi R^2}{2}. \end{aligned}$$

C'est l'aire d'un demi-disque, donc l'aire du disque entier est πR^2.

Nous avons ainsi retrouvé une formule classique de collège (et montré que les intégrales permettent vraiment de calculer des aires !). Mais nous n'avons pas utilisé de géométrie : le nombre π qui apparaît ici est la demi-période du sinus ; c'est aussi le demi-périmètre du cercle unité. Son apparition dans une formule d'aire est donc *a priori* un peu « surprenante » (même quand on connaît déjà la formule).

5.3 Quel changement de variable choisir ?

Il n'y a pas de règle absolue, mais certains motifs reviennent fréquemment. En voici quelques-uns. Nous utiliserons ici parfois l'écriture $\displaystyle\int f(x)\ \mathrm{d}x$ pour représenter une primitive arbitraire de f.

Fonction f faisant intervenir seulement e^x

Poser $t := \mathrm{e}^x$, de sorte que $\mathrm{d}x = \mathrm{d}t/t$.

Exercice 13.

Rechercher une primitive de $x \mapsto \dfrac{1}{\cosh x}$.

Solution. Par définition $f : x \mapsto 1/\cosh x$ vérifie les égalités :

$$f(x) = \frac{2}{\mathrm{e}^x + \mathrm{e}^{-x}} = \frac{2\,\mathrm{e}^x}{\mathrm{e}^{2x} + 1} = \frac{2t}{1 + t^2}$$

et en posant $t = \mathrm{e}^x$, d'où $\mathrm{d}t = t\,\mathrm{d}x$, on obtient :

$$\int f(x)\ \mathrm{d}x = \int \frac{2}{1+t^2}\ \mathrm{d}t = 2\arctan t = 2\arctan \mathrm{e}^x.$$

Fonction f faisant intervenir seulement $\cos x$, $\sin x$ et $\tan x$

Puisque $\tan$ est le quotient $\sin / \cos$, on peut se limiter à des fonctions où x n'apparaît que par l'intermédiaire de cosinus et de sinus, ainsi que sa différentielle.
Une bonne règle est d'essayer d'abord l'un des trois changements de variables $t := \cos x$, $t := \sin x$, $t := \tan x$. Si cela ne fonctionne pas, on peut alors utiliser le changement de variable $t := \tan(x/2)$. Ce changement de variable convient toujours, mais donne des calculs plus compliqués que les trois précédents (lorsque l'un de ceux-ci convient).

Utilisation de $\cos x$ *et* $\sin x$.

On peut éliminer les termes en $\cos^2 x$ ou $\sin^2 x$ à l'aide de la relation $\cos^2 x + \sin^2 x = 1$, mais il faut éviter en général d'introduire des racines carrées, en tenant compte que $\mathrm{d}t$ contient, lui aussi, un $\cos x$ ou un $\sin x$.
De manière plus précise, si l'on veut utiliser le changement de variable $t := \cos x$, d'où $\mathrm{d}t = -\sin x\ \mathrm{d}x$, on essaie d'écrite $f(x) := g(\cos x)\ \sin x$, ce qui donne :

$$\int f(x)\ \mathrm{d}x = \int g(\cos x)\ \sin x\,\mathrm{d}x = -\int g(t)\ \mathrm{d}t.$$

De le même façon, pour utiliser le changement de variable $t := \sin x$, d'où $\mathrm{d}t = \cos x\ \mathrm{d}x$, on essaie d'écrire $f(x) := g(\sin x)\ \cos x$.

Exercice 14.

Calculer la primitive :

$$\int (4\cos^3 x - 3\cos x)\ \mathrm{d}x.$$

Solution. On peut poser ici $t := \sin x$, d'où les égalités (à une constante près) :

$$\begin{aligned}\int (4\cos^3 x - 3\cos x)\ \mathrm{d}x &= \int (4\,(1-\sin^2 x) - 3)\ \cos x\ \mathrm{d}x = \int (4\,(1-t^2) - 3)\ \mathrm{d}t \\ &= \int (1-4t^2)\ \mathrm{d}t = \frac{t}{3}\,(3-4t^2) = \frac{\sin x}{3}\,(3-4\sin^2 x).\end{aligned}$$

Utilisation de $\tan x$.

Dans le cas $t := \tan x$, on a $\mathrm{d}t = (1+t^2)\ \mathrm{d}x$.
Or, $\cos^2 x = 1/(1+t^2)$ et $\sin^2 x = t^2/(1+t^2)$, donc ce changement de variable fonctionne bien dans le cas où $f(x)$ peut s'exprimer uniquement à l'aide de $\tan x$, $\cos^2 x$ et $\sin^2 x$.

Exercice 15.

Calculer une primitive de la fonction $x \mapsto \tan^3 x$ sur l'intervalle $\left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$.

Solution. On posera ici $t := \tan x$ puis $u := t^2$. Il vient :

$$\begin{aligned}\int \tan^3 x \,\mathrm{d}x = \int t^3 \frac{\mathrm{d}t}{1+t^2} &= \frac{1}{2}\int \frac{u\,\mathrm{d}u}{1+u} \\ &= \frac{1}{2}\int \left(1 - \frac{1}{1+u}\right)\mathrm{d}u \\ &= \frac{1}{2}\left(u - \ln|1+u|\right) \\ &= \frac{1}{2}\left(\tan^2 x - \ln\left(1+\tan^2 x\right)\right) \\ &= \frac{1}{2}\left(\tan^2 x + 2\ln|\cos x|\right) = \frac{1}{2}\tan^2 x + \ln(\cos x)\end{aligned}$$

puisque sur l'intervalle de définition $0 < \cos x = \dfrac{1}{\sqrt{1+\tan^2 x}}$.

Utilisation de $\tan(x/2)$.

En travaillant sur un intervalle de la forme $]-\pi + 2k\pi, \pi + 2k\pi[$, on peut exprimer $\sin x$, $\cos x$, $\tan x$ et $\mathrm{d}x$ en fonction de $t := \tan(x/2)$ et $\mathrm{d}t$:

$$\cos x = \frac{1-t^2}{1+t^2}, \quad \sin x = \frac{2t}{1+t^2}, \quad \tan x = \frac{2t}{1-t^2}, \quad \mathrm{d}x = \frac{2\,\mathrm{d}t}{1+t^2}.$$

Exercice 16.

Déterminer une primitive de $1/\sin$. En déduire une primitive de $1/\cos$.

Solution.

- On trouve : $\displaystyle\int \frac{\mathrm{d}x}{\sin x} = \int \frac{1+t^2}{2t}\,\frac{2}{1+t^2}\,\mathrm{d}t = \int \frac{1}{t}\,\mathrm{d}t = \ln|t| = \ln\left|\tan\frac{x}{2}\right|$.
- On peut faire le même type de calcul :

$$\int \frac{\mathrm{d}x}{\cos x} = \int \frac{1+t^2}{1-t^2}\,\frac{2}{1+t^2}\,\mathrm{d}t = \int \frac{2}{1-t^2}\,\mathrm{d}t = \ln\left|\frac{1+t}{1-t}\right| = \ln\left|\frac{1+\tan\frac{x}{2}}{1-\tan\frac{x}{2}}\right|,$$

  mais un changement de variable affine $u = x + \pi/2$ nous ramène aussi au calcul précédent et donc à :

$$\int \frac{\mathrm{d}x}{\cos x} = \ln\left|\tan\left(\frac{x}{2} + \frac{\pi}{4}\right)\right|.$$

Fonction f faisant intervenir seulement $\cosh x$, $\sinh x$ et $\tanh x$

Les mêmes idées que dans le cas des fonctions trigonométriques ordinaires fonctionnent. Mais on peut aussi tout écrire en fonction de $t := \mathrm{e}^x$.

Fonction f faisant intervenir $\sqrt{ax^2+bx+c}$

Posons $y := \sqrt{ax^2+bx+c}$ lorsque cette expression a un sens. Il s'agit donc de calculer une intégrale de la forme $\displaystyle\int f\big(x, y(x)\big)\ \mathrm{d}x$ en faisant disparaître la racine carrée. On ne peut pas, en général, prendre y comme nouvelle variable, car pour pouvoir exprimer x en fonction de y (au moins pour transformer le $\mathrm{d}x$), cela ferait intervenir une autre racine carrée.
Si $a = 0$, il suffit de poser $t := \sqrt{bx+c}$. Sinon, la racine porte sur un polynôme du second degré, dont on calcule le discriminant $\Delta := b^2 - 4ac$.

- Si $\Delta < 0$ et $a < 0$, il n'existe aucune valeur de x pour laquelle $\sqrt{ax^2+bx+c}$ soit défini. Le problème n'a donc pas de sens dans ce cas.

- Si $\Delta < 0$ et $a > 0$, le réel y est défini pour tout x. L'égalité :
$$y = \frac{1}{2\sqrt{a}}\sqrt{(2ax+b)^2-\Delta} = \frac{1}{2}\sqrt{\frac{-\Delta}{a}}\sqrt{1+\Big(\frac{2ax+b}{\sqrt{-\Delta}}\Big)^2}$$
conduit à poser $2ax+b = \sqrt{-\Delta}\,\sinh u$, d'où :
$$\mathrm{d}x = \frac{\sqrt{-\Delta}}{2a}\cosh u\ \mathrm{d}u, \qquad y = \frac{1}{2}\sqrt{\frac{-\Delta}{a}}\cosh u.$$
On est donc ramené à calculer une intégrale de la forme $\displaystyle\int g(\cosh u, \sinh u)\ \cosh u\ \mathrm{d}u$.

- Si $\Delta = 0$, il n'y a qu'une racine réelle α, et $y = \sqrt{ax^2+bx+c}$ est un multiple de $|x-\alpha|$: il suffit donc de séparer l'intégrale au point α avec la formule de Chasles.

- Si $\Delta > 0$ et $a < 0$, il faut se placer entre les deux racines de l'équation $y^2 = 0$. L'égalité :
$$y = \frac{1}{2\sqrt{-a}}\sqrt{\Delta-(2ax+b)^2} = \frac{1}{2}\sqrt{\frac{\Delta}{-a}}\sqrt{1-\Big(\frac{2ax+b}{\sqrt{\Delta}}\Big)^2}$$
conduit à poser $2ax+b = \sqrt{\Delta}\,\sin u$ avec par exemple $u \in \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big]$, d'où :
$$\mathrm{d}x = \frac{\sqrt{\Delta}}{-2a}\cos u\ \mathrm{d}u, \qquad y = \frac{1}{2}\sqrt{\frac{\Delta}{-a}}\cos u.$$
On est donc ramené à calculer une intégrale de la forme $\displaystyle\int h(\cos u, \sin u)\ \cos u\ \mathrm{d}u$.

- Si $\Delta > 0$ et $a > 0$, il faut se placer à l'extérieur des racines de l'équation $y^2 = 0$. L'égalité :
$$y = \frac{1}{2\sqrt{a}}\sqrt{(2ax+b)^2-\Delta} = \frac{1}{2}\sqrt{\frac{\Delta}{a}}\sqrt{\Big(\frac{2ax+b}{\sqrt{\Delta}}\Big)^2-1}$$
conduit à poser $2ax+b = \pm\sqrt{\Delta}\,\cosh u$ avec $u \geqslant 0$ (le signe $\pm$ dépend de l'inter-

valle sur lequel on cherche à intégrer), d'où :

$$\mathrm{d}x = \pm\frac{\sqrt{\Delta}}{2a}\,\sinh u\ \mathrm{d}u, \qquad y = \frac{1}{2}\sqrt{\frac{-\Delta}{a}}\,\sinh u.$$

On est donc ramené à calculer une intégrale de la forme :

$$\int k(\cosh u, \sinh u)\ \sinh u\ \mathrm{d}u.$$

Dans tous ces cas, on est donc ramené à l'un des problèmes précédents.

Exercice 17.

Calculer une primitive de $x \mapsto x\sqrt{x^2+x+1}$.

Solution. On note que $x^2+x+1 = \left(x+\frac{1}{2}\right)^2 + \frac{3}{4}$, d'où le changement de variable $x = \frac{\sqrt{3}}{2}\,\sinh u - \frac{1}{2}$.

On sait que le réel u défini par une égalité de la forme $\sinh u = v$ vérifie l'égalité $u = \ln\left(v+\sqrt{1+v^2}\right)$. Sachant que $v = (2x+1)/\sqrt{3}$, tous calculs faits, cela donne :

$$u = \ln\left(\frac{2x+1+2\sqrt{x^2+x+1}}{\sqrt{3}}\right).$$

Il résulte du changement de variable annoncé que :

$$y := \sqrt{x^2+x+1} = \frac{\sqrt{3}}{2}\,\cosh u$$

et que l'intégrale à calculer I s'écrit :

$$\int x\sqrt{x^2+x+1}\ \mathrm{d}x = \int xy\ \mathrm{d}x = \frac{3}{8}\int(\sqrt{3}\sinh u - 1)\ \cosh^2 u\,\mathrm{d}u.$$

On obtient ainsi une égalité de la forme $I = J - K$, avec :

$$J := \frac{3\sqrt{3}}{8}\int \cosh^2 u\ \sinh u\ \mathrm{d}u = \frac{\sqrt{3}}{8}\cosh^3 u = \frac{y^3}{3} = \frac{y\,(x^2+x+1)}{3},$$

$$K := \frac{3}{8}\int \cosh^2 u\ \mathrm{d}u = \frac{3}{16}\int(1+\cosh 2u)\ \mathrm{d}u$$

$$= \frac{3u}{16} + \frac{3}{32}\sinh 2u = \frac{3u}{16} + \frac{3}{16}\sinh u\ \cosh u = \frac{3u}{16} + \frac{y\,(2x+1)}{8}.$$

Tous calculs faits, une primitive, égale à I augmentée d'un logarithme de constante pour simplifier, est :

$$x \mapsto \frac{1}{48}\left(2y\,(8x^2+2x+5) - 9\,\ln(2x+1+2y)\right) \qquad \text{avec} \qquad y = \sqrt{x^2+x+1}.$$

5.4 Intégration des fractions rationnelles

Il n'y aucune relation simple entre des intégrales de fonctions et celle de leur quotient (ce n'est pas très étonnant, car déjà la formule pour la dérivée d'un quotient est relativement compliquée). Quand on cherche à intégrer un quotient f/g, on peut parfois le considérer comme un produit (de f par l'inverse de g, par exemple), et essayer une intégration par parties.

Mais il existe un cas dans lequel le calcul est toujours possible, c'est celui du quotient de deux polynômes, c'est-à-dire d'une fraction rationnelle. En effet, on peut démontrer (c'est un cas particulier du théorème 67 de la page 320) qu'une telle fraction peut s'écrire comme somme d'« éléments simples », de l'une des trois formes suivantes :

- un polynôme ;
- un « élément simple de première espèce » de la forme $\dfrac{\alpha}{(x-a)^n}$ avec $n \in \mathbb{N}^*$;
- un « élément simple de seconde espèce » de la forme $\dfrac{\alpha x+\beta}{\big((x-u)^2+v^2\big)^n}$ avec $n \in \mathbb{N}^*$ et $(u,v) \in \mathbb{R} \times \mathbb{R}_+^*$.

Cette décomposition peut être obtenue de manière explicite (et même algorithmique), si l'on sait factoriser le dénominateur de la fraction (voir cours d'algèbre) ; la méthode des coefficients indéterminés peut aider. On pourra se reporter au lemme 65 de la page 319.

Il suffit donc de savoir intégrer les « éléments simples ». Dans le cas des polynômes, c'est trivial. Pour les éléments simples de première espèce, on trouve aisément une primitive, à savoir $\ln|x-a|$ si $n=1$ et, sinon, $-\dfrac{1}{(n-1)(x-a)^{n-1}}$ sinon.

Pour les éléments simples de seconde espèce, on fait le changement de variable affine $t := (x-u)/v$, et l'on doit rechercher des primitives de $t/(t^2+1)^n$ et $1/(t^2+1)^n$:

- Une primitive de $t \mapsto t/(t^2+1)^n$ est obtenue simplement à l'aide du nouveau changement de variable $s = t^2+1$: on trouve $\ln\sqrt{t^2+1}$ si $n=1$, et $-1/\big((2n-2)(t^2+1)^{n-1}\big)$ si $n \geqslant 2$.
- Une primitive de $t \mapsto 1/(t^2+1)^n$ est connue dans le cas $n=1$ (c'est l'arc tangente). Pour les valeurs plus grandes de n, il faut procéder par étapes récurrentes. Intégrons :

$$R_n(x) := \int_0^x \frac{1}{(t^2+1)^n}\,\mathrm{d}t$$

par parties en introduisant le produit de t par $1/(t^2+1)^n$, ce qui donne :

$$\begin{aligned} R_n(x) = \int_0^x \frac{1}{(t^2+1)^n}\,\mathrm{d}t &= \Big[\frac{t}{(t^2+1)^n}\Big]_0^x + 2n\int_0^x \frac{t^2}{(t^2+1)^{n+1}}\,\mathrm{d}t \\ &= \frac{x}{(x^2+1)^n} + 2n\,R_n(x) - 2n\,R_{n+1}(x) \end{aligned}$$

puisque $t^2 = (t^2+1)-1$. Par suite :

$$R_{n+1}(x) = \frac{2n-1}{2n}\,R_n(x) + \frac{1}{2n}\,\frac{x}{(x^2+1)^n}.$$

On peut ainsi trouver de proche en proche, par récurrence, tous les $R_n(x)$ en partant de $R_1(x) = \arctan(x)$. Par exemple :

$$R_2(x) = \frac{1}{2}\arctan x + \frac{1}{2}\,\frac{x}{(x^2+1)}\cdot$$

Exercice 18.

Calculer une primitive de $f : x \mapsto \dfrac{60}{(x-3)(x^2-8x+25)}\cdot$

Solution. Pour le pôle 3 (c'est-à-dire la racine du dénominateur égale à 3), il existe un élément simple de la forme $a/(x-3)$. Soit $g(x) := f(x) - \dfrac{a}{x-3}$; multipliant les deux membres par $x-3$ puis remplaçant x par 3 (ce qui est possible car par définition 3 n'est pas un pôle de g), il vient :

$$a = \frac{60}{3^2 - 8.3 + 25} = 6, \qquad g(x) = f(x) - \frac{6}{x-3} = \frac{6(5-x)}{x^2-8x+25}$$

après simplification par $x-3$.
Le trinôme $x^2 - 8x + 25 = (x-4)^2 + 9 = (x-4)^2 + 3^2$ n'a pas de racines réelles. On est donc conduit à poser $t := (x-4)/3$, soit $x = 4 + 3t$, $\mathrm{d}x = 3\,\mathrm{d}t$. Il vient alors :

$$g(x)\ \mathrm{d}x = \frac{2(1-3t)}{t^2+1}\ \mathrm{d}t = \frac{2\,\mathrm{d}t}{t^2+1} - \frac{6t\,\mathrm{d}t}{t^2+1}$$

d'où enfin :

$$\int f(x)\ \mathrm{d}x = 6\int \frac{\mathrm{d}x}{x-3} + 2\int \frac{\mathrm{d}t}{1+t^2} - 6\int \frac{t\,\mathrm{d}t}{t^2+1}\cdot$$

Tout cela s'intègre désormais à vue ; on trouve :

$$\int f(x)\ \mathrm{d}x = 6\ \ln\,|x-3| + 2\ \arctan t - 3\ \ln\,(1+t^2)$$

c'est-à-dire en revenant à x :

$$\int f(x)\ \mathrm{d}x = 6\ \ln\,|x-3| + 2\ \arctan\frac{x-4}{3} - 3\ \ln\,\Big(\frac{x^2-8x+25}{9}\Big)$$

soit enfin, puisqu'une primitive n'est définie qu'à une constante près :

$$\int \frac{60}{(x-3)(x^2-8x+25)}\ \mathrm{d}x = 6\ \ln\,|x-3| + 2\ \arctan\frac{x-4}{3}\ -\ 3\ \ln\,(x^2-8x+25)$$

(des parenthèses suffisent dans le dernier logarithme, car la quantité $x^2-8x+25$ est toujours strictement positive). Bien entendu, une telle intégrale n'a de sens que sur un intervalle inclus dans l'une des deux demi-droites $]-\infty, 3[$ et $]3, +\infty[$.

Exercice 19.

Calculer une primitive de la fonction définie par :

$$f(x) := \frac{1}{x\,(x^2+1)^7}.$$

Solution. On pourrait appliquer à la lettre les méthodes décrites ci-dessus, mais le résultat serait pénible à obtenir.

Au contraire, poser $t := x^2+1$ dès le départ conduit à une solution beaucoup plus rapide. Multipliant en haut et en bas par x, on est ramené à intégrer par rapport à t l'expression $g(t) := \dfrac{1}{2(t-1)t^7}$ dont 1 est un pôle.

On doit chercher un réel a tel que $g(t) - a/(t-1)$ se simplifie ; multipliant par $t-1$ puis remplaçant t par 1, il vient $a = 1/2$. Effectivement :

$$g(t) - \frac{1}{2(t-1)} = \frac{1-t^7}{2(t-1)t^7} = -\frac{1+t+t^2+t^3+t^4+t^5+t^6}{t^7}$$

ce qui s'intègre à vue. On trouve en effet :

$$\int g(t)\ \mathrm{d}t = \frac{1}{2}\left(\ln(t-1) - \ln t\right) + \frac{1}{t} + \frac{1}{2t^2} + \frac{1}{3t^3} + \frac{1}{4t^4} + \frac{1}{5t^5} + \frac{1}{6t^6}.$$

Remplaçant enfin t par sa valeur en fonction de x, on trouve tous calculs faits :

$$\ln\frac{|x|}{\sqrt{x^2+1}} + \frac{147+522x^2+855x^4+740x^6+330x^8+60x^{10}}{120(x^2+1)^6}.$$

Remarque. Les calculs de primitive sont délicats et doivent être menés avec soin ; il faut souvent procéder par « essais et erreurs »... et se montrer patient ! Les logiciels de calcul formel ont pratiquement rendu ces techniques obsolètes pour le calculateur humain, car ils ont très bien su les intégrer dans leurs algorithmes, bien que parfois les résultats qu'ils donnent — mathématiquement corrects — puissent surprendre un œil non averti. Il y a souvent des problèmes de simplification des expressions trouvées, notamment lorsqu'il s'agit de fonctions trigonométriques.

Exemple. Le logiciel `Maple` donne (après simplification) :

$$\int \sqrt{\frac{x-1}{2-x}}\,\frac{\mathrm{d}x}{x} = \arcsin(-3+2\,x) - \frac{\sqrt{2}}{2}\arctan\left(\frac{(-4+3\,x)\sqrt{2}}{4\sqrt{-2-x^2+3\,x}}\right).$$

Pour calculer cette primitive sur l'intervalle $[1,2[$, on peut faire le changement de variable :

$$t = \sqrt{\frac{x-1}{2-x}} \quad \text{soit} \quad x = \frac{1+2t^2}{1+t^2} \quad \text{et} \quad \mathrm{d}x = \frac{2t\,\mathrm{d}t}{(1+t^2)^2}.$$

On obtient finalement $\displaystyle\int \sqrt{\frac{x-1}{2-x}}\,\frac{\mathrm{d}x}{x} = \int \frac{2t^2}{(1+2\,t^2)\,(1+t^2)}\mathrm{d}t.$

La décomposition en éléments simples $\frac{2u}{(1+u)(1+2u)} = \frac{2}{1+u} - \frac{2}{1+2u}$ conduit après calculs au résultat :

$$\int \sqrt{\frac{x-1}{2-x}}\,\frac{\mathrm{d}x}{x} = 2\arctan(t) - \sqrt{2}\arctan\left(t\sqrt{2}\right)$$
$$= 2\arctan\left(\sqrt{\frac{x-1}{2-x}}\right) - \sqrt{2}\arctan\left(\sqrt{\frac{2x-2}{2-x}}\right).$$

Il est un peu difficile de voir que les deux expressions trouvées correspondent à la même primitive (à une constante additive près). Pour s'en convaincre, on peut tracer les deux courbes représentatives (en trait continu, l'expression donnée par `Maple` et en pointillé celle trouvée à la main).

Exercice type corrigé

IV.5.0 Pour tout $n \in \mathbb{N}$, on pose $I_n := \displaystyle\int_0^1 (1-t^2)^n\,\mathrm{d}t$.

1. L'intégrale I_n est-elle bien définie ?
2. Montrer que la suite (I_n) est décroissante.
3. Montrer que pour tout $u \in [0,1]$, on a $0 \leqslant 1-u \leqslant \mathrm{e}^{-u}$.
4. En déduire une majoration de I_n à l'aide de $J_n := \displaystyle\int_0^{\sqrt{n}} \mathrm{e}^{-x^2}\,\mathrm{d}x$.
5. Montrer que la suite (J_n) est bornée. On pourra remarquer que $\forall x \geqslant 1\,,\ \mathrm{e}^{-x^2} \leqslant \mathrm{e}^{-x}$.
   En déduire que I_n tend vers 0 lorsque n tend vers $+\infty$.
6. Soit $n \in \mathbb{N}^*$. Calculer $I_{n-1} - I_n = \displaystyle\int_0^1 t \times t\,(1-t^2)^{n-1}\,\mathrm{d}t$ à l'aide d'une intégration par parties pour en déduire la relation $I_n = \frac{2n}{2n+1}\,I_{n-1}$.
7. En déduire une expression de I_n à l'aide de factorielles.
8. En déduire une expression de $\displaystyle\int_0^{\pi/2} \cos^{2n+1}\theta\,\mathrm{d}\theta$ pour $n \in \mathbb{N}$ (intégrale de Wallis).

Solution.

1. La fonction polynomiale $t \mapsto (1-t^2)^n$ est continue sur le segment $[0,1]$, ce qui justifie l'existence de son intégrale.
2. En intégrant la relation évidente $\forall t \in [0,1]\,,\ (1-t^2)^{n+1} \leqslant (1-t^2)^n$, on obtient $I_{n+1} \leqslant I_n$. D'où la décroissance de (I_n).
3. La première inégalité est évidente.
   Pour la seconde, on peut étudier la fonction $\varphi : u \mapsto \mathrm{e}^{-u} - 1 + u$. Sa dérivée $\varphi'(u) = 1 - \mathrm{e}^{-u}$ reste positive sur $[0,1]$, donc φ est croissante sur $[0,1]$. Comme $\varphi(0) = 0$, on en déduit que φ est positive, d'où le résultat.

4. On a donc :
$$I_n \leqslant \int_0^1 (\mathrm{e}^{-t^2})^n \,\mathrm{d}t = \int_0^1 \mathrm{e}^{-nt^2} \,\mathrm{d}t.$$
En effectuant le changement de variable affine $x = t\sqrt{n}$, $\mathrm{d}x = \sqrt{n}\,\mathrm{d}t$, cela donne :
$$\int_0^1 \mathrm{e}^{-nt^2} \,\mathrm{d}t = \frac{1}{\sqrt{n}} \int_0^{\sqrt{n}} \mathrm{e}^{-x^2} \,\mathrm{d}x$$
et donc $I_n \leqslant \dfrac{J_n}{\sqrt{n}}$.

5. En majorant e^{-x^2} par 1 si $x \leqslant 1$ et par e^{-x} si $x \geqslant 1$, on obtient, par la relation de Chasles :
$$0 \leqslant J_n \leqslant \int_0^1 \mathrm{d}x + \int_1^{\sqrt{n}} \mathrm{e}^{-x} \,\mathrm{d}x = 1 + \left[-\mathrm{e}^{-x}\right]_1^{\sqrt{n}} = 1 + \mathrm{e}^{-1} - \mathrm{e}^{-\sqrt{n}} \leqslant 1 + \mathrm{e}^{-1} \leqslant 2.$$
Donc (J_n) est bornée par 2.
On en déduit $0 \leqslant I_n \leqslant \frac{2}{\sqrt{n}} \underset{n\to+\infty}{\longrightarrow} 0$.

6. L'intégration par parties se fait en dérivant $u(t) := t$ et en primitivant $v'(t) := t\,(1-t^2)^{n-1}$, soit par exemple, $v(t) = -\frac{1}{2n}(1-t^2)^n$.
On obtient ainsi :
$$I_{n-1} - I_n = \left[-t\frac{(1-t^2)^n}{2n}\right]_0^1 + \int_0^1 \frac{(1-t^2)^n}{2n} \,\mathrm{d}t = \frac{1}{2n} I_n$$
ce qui donne $I_{n-1} = \left(1 + \dfrac{1}{2n}\right) I_n$ et donc $I_n = \dfrac{2n}{2n+1} I_{n-1}$.

7. Comme $I_0 = 1$, on en déduit :
$$I_n = \frac{(2n)(2n-2)\cdots 2}{(2n+1)(2n-1)\cdots 1} I_0 = \frac{\big((2n)(2n-2)\cdots 2\big)^2}{(2n+1)(2n)\cdots 1} = \frac{\big(2^n\, n!\big)^2}{(2n+1)!}$$
(on a multiplié en haut et en bas par le produit des entiers pairs pour former le $(2n+1)!$ du dénominateur).

8. Réécrivons l'intégrale proposée :
$$\int_0^{\pi/2} \cos^{2n+1}\theta \,\mathrm{d}\theta = \int_0^{\pi/2} \cos^{2n}\theta \,\cos\theta \,\mathrm{d}\theta = \int_0^{\pi/2} (1-\sin^2\theta)^n \,\cos\theta \,\mathrm{d}\theta$$
et le changement de variable $t = \sin\theta$ permet de conclure :
$$\int_0^{\pi/2} \cos^{2n+1}\theta \,\mathrm{d}\theta = I_n = \frac{\big(2^n\, n!\big)^2}{(2n+1)!}.$$

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

IV.5.1 Calculer les intégrales suivantes :

$$A := \int_1^3 \left(x + \frac{1}{x}\right)^2 \mathrm{d}x \qquad \text{et} \qquad B := \int_0^1 (2x+3)\,(x^2-3x+5)^3 \;\mathrm{d}x.$$

IV.5.2 Calculer les primitives suivantes :

$$C(x) := \int \frac{\ln|x|}{x}\;\mathrm{d}x \qquad \text{et} \qquad D(x) := \int \frac{1}{x\,\ln|x|}\;\mathrm{d}x.$$

IV.5.3
Calculer les primitives suivantes (on donne la forme de la décomposition en éléments simples) :

$$E := \int \frac{x+2}{x^2+5x-6}\;\mathrm{d}x \;\text{ en calculant }\; a \;\text{ et }\; b \;\text{ tels que }\; \frac{x+2}{x^2+5x-6} = \frac{a}{x-1} + \frac{b}{x+6}\cdot$$

$$F := \int \frac{x^3-2}{x^3-x^2}\;\mathrm{d}x \;\text{ en calculant }\; a,\; b,\; c \;\text{ et }\; d \;\text{ tels que }\; \frac{x^3-2}{x^3-x^2} = a + \frac{b}{x^2} + \frac{c}{x} + \frac{d}{x-1}\cdot$$

$$G := \int \frac{3x^4+2x^2+1}{x^3(x^2+1)^2}\;\mathrm{d}x \;\text{ en calculant }\; a \;\text{ et }\; b \;\text{ tels que }\; \frac{3x^4+2x^2+1}{x^3(x^2+1)^2} = \frac{a}{x^3} + \frac{bx}{(x^2+1)^2}\cdot$$

IV.5.4 Calculer les primitives suivantes en utilisant la méthode d'intégration par parties :

$$H := \int x^2\,\ln x\;\mathrm{d}x, \qquad I := \int \arctan x\;\mathrm{d}x.$$

IV.5.5 Calculer les intégrales suivantes :

$$J := \int_1^2 \frac{1}{\sinh x}\;\mathrm{d}x, \quad K := \int_1^2 \frac{1}{1+\mathrm{e}^x}\;\mathrm{d}x, \quad L := \int_{1/\sqrt{2}}^1 \mathrm{e}^{-\arcsin t}\,\mathrm{d}t$$

en utilisant les changements de variable indiqués ci-dessous :

$$u := \cosh x, \quad t := \mathrm{e}^x, \quad t := \sin\theta.$$

IV.5.6 Calculer les primitives suivantes :

$$M := \int \sin^3 x\,\cos^2 \mathrm{d}x, \qquad N := \int \frac{\sin 2x}{\cos^2 x\,(1+\sin^2 x)}\;\mathrm{d}x$$

(pour cette dernière, on pourra poser $t := \tan x$ en délimitant les intervalles sur lesquels il existe une primitive).

IV.5.7 Démontrer pour tout entier positif n l'égalité :

$$\int_0^{\pi/2} \sin^n x \, \mathrm{d}x = \int_0^{\pi/2} \cos^n x \, \mathrm{d}x.$$

IV.5.8 Soit f une fonction continue sur $\mathbb{R}$. Déterminer :

$$\lim_{x \to 0} \frac{1}{x} \int_{1-x}^{1+x} f(t) \, \mathrm{d}t.$$

IV.5.9 Étudier la limite quand x tend vers 0 de $F(x) = \dfrac{1}{x} \displaystyle\int_0^x \frac{\sin t}{t} \, \mathrm{d}t$ (l'intégrale est en fait celle de la fonction f définie par $f(t) := \sin(t)/t$ pour $t \neq 0$ et $f(0) := 1$).

IV.5.10 Pour tous les réels a, b tels que $0 < a < b$, montrer que :

$$\ln b - \ln a \leqslant \frac{b-a}{\sqrt{ab}}.$$

On pourra par exemple exprimer $\ln b - \ln a$ *comme une intégrale et appliquer l'inégalité de Cauchy-Schwarz.*

IV.5.11

1. Calculer la limite suivante : $\displaystyle\lim_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} f\left(\frac{k}{n}\right)$, où f est une fonction continue sur $[0,1]$.
2. En déduire ;

$$\lim_{n \to +\infty} \frac{1}{n^2} \sum_{k=1}^{n} k\mathrm{e}^{-k/n} \qquad \text{et} \qquad \lim_{n \to +\infty} \frac{\sqrt{1} + \sqrt{2} + \cdots + \sqrt{n}}{n\sqrt{n}}.$$

IV.5.12

1. Vérifier les inégalités $0 \leqslant \ln(1+x) \leqslant x$ pour tout réel $x \geqslant 0$.
2. Calculer la limite lorsque n tend vers $+\infty$ de $\displaystyle\int_0^1 \ln(1+x^n) \, \mathrm{d}x$.
3. Montrer que, pour tous les entiers strictement positifs, on dispose de l'égalité :

$$u_n := \int_0^1 \frac{x^n}{1+x^n} \, \mathrm{d}x = \frac{\ln 2}{n} - \frac{1}{n} \int_0^1 \ln(1+x^n) \, \mathrm{d}x.$$

4. Étudier les limites des suites (u_n) et (nu_n).

IV.5.13

1. Soit g une fonction continûment dérivable sur un intervalle $[a, b]$. Montrer que g est aussi bornée sur $[a, b]$.
2. En déduire que $n \displaystyle\int_a^b g(t)\, \sin nt \, \mathrm{d}t$ est bornée par une constante indépendante de l'entier n.

   Quelle est la limite de $\displaystyle\int_a^b \sin nt \, \mathrm{d}t$ quand n tend vers l'infini ?
3. Pour un entier fixé $p \geqslant 1$, étudier la limite lorsque n tend vers $+\infty$ de :

$$\int_0^{\pi/2} t^p \, \frac{\sin nt}{\sin t} \, \mathrm{d}t$$

   où l'on aura, bien entendu, prolongé la fonction $t \mapsto \sin(nt)/\sin(t)$ par n en 0 de façon à intégrer une fonction continue sur $[0, \pi/2]$.
4. En déduire la limite, quand n tend vers $+\infty$ de

$$A_{p,n} := \int_0^{\pi/2} t^p \cos(n+1)t \, \frac{\sin nt}{\sin t} \, \mathrm{d}t.$$

IV.5.14

Calculer $\displaystyle\int_0^1 x^a (1-x)^b \, \mathrm{d}x$ pour $(a, b) \in \mathbb{N}^2$.

Introduction aux fonctions vectorielles d'une variable réelle

IV.6

1 Suites vectorielles

Tout au long de ce module, nous allons manipuler des éléments de $\mathbb{R}^m$ avec $m \in \mathbb{N}^*$. Afin de les distinguer des réels, nous les noterons en gras :

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$$

avec $x_1 \in \mathbb{R}$, $x_2 \in \mathbb{R}$, ..., $x_m \in \mathbb{R}$. Nous penserons généralement à $\boldsymbol{x} \in \mathbb{R}^m$ comme à un élément de l'espace vectoriel $\mathbb{R}^m$.
Il est possible de munir $\mathbb{R}^m$ de sa structure d'espace affine canonique (voir module III.1) et de penser à $\boldsymbol{x}$ comme à un point. Dans le cas où $n = 2$ ou $n = 3$, nous renvoyons le lecteur au module III.2 pour une interprétation géométrique des résultats que nous établissons ici.
Dans le module IV.1, nous avons introduit la notion de suite. Nous avons mis l'accent sur les suites numériques, celles dont les termes appartiennent à $\mathbb{R}$ ou à $\mathbb{C}$. Nous allons maintenant nous intéresser aux suites vectorielles $(\boldsymbol{x}_n)$ dont les termes $\boldsymbol{x}_n \in \mathbb{R}^m$ sont des vecteurs.

Exemples.

1. L'ensemble $\mathbb{C}$ des nombres complexes a été défini comme l'ensemble produit $\mathbb{R} \times \mathbb{R}$ muni de deux lois de composition internes $+$ et $\times$. On peut donc naturellement identifier une suite numérique de nombre complexes $(z_n = x_n + iy_n)_{n\in\mathbb{N}}$ avec la suite vectorielle $\big((x_n, y_n)\big)_{n\in\mathbb{N}}$ de vecteurs de $\mathbb{R}^2$. Par exemple, la suite de nombres complexes $(1 + in)_{n\in\mathbb{N}}$ est identifiée avec la suite de vecteurs $\big((1, n)\big)_{n\in\mathbb{N}}$.

2. L'ensemble $M_{m,n}(\mathbb{R})$ des matrices $m \times n$ peut être identifié avec $\mathbb{R}^{mn}$. Par exemple, la matrice $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{R})$ peut être identifiée avec le vecteur $(a, b, c, d) \in \mathbb{R}^4$.
   L'étude des suites vectorielles couvre donc l'étude des suites de matrices.

1.1 Distance entre deux vecteurs

Dans le module IV.1, nous avons introduit la notion de suite numérique convergente, en particulier pour pouvoir parler de fonction continue au module IV.2. Nous allons maintenant étendre la notion de convergence pour les suites vectorielles. Pour cela, il faut que les termes de la suite se rapprochent de la limite. Dans le cas d'une suite numérique (x_n) qui converge vers une limite ℓ, nous avons traduit cela par le fait que la suite réelle positive $|x_n - \ell|$ tend vers 0. La quantité $|x_n - \ell|$ représente la distance entre x_n et ℓ. Il nous faut donc apprendre à mesurer la distance entre deux vecteurs $\boldsymbol{x} \in \mathbb{R}^m$ et $\boldsymbol{y} \in \mathbb{R}^m$.
Dans $\mathbb{R}$, nous mesurons la distance entre deux points x et x' par la valeur absolue $|x - x'|$, et dans $\mathbb{C}$, par le module $|x - x'|$. Dans $\mathbb{R}^2$, la norme du vecteur $\boldsymbol{x} - \boldsymbol{x}'$ où $\boldsymbol{x} := (x, y) \in \mathbb{R}^2$ et $\boldsymbol{x}' := (x', y') \in \mathbb{R}^2$ est donnée par le théorème de Pythagore :

$$\|\boldsymbol{x} - \boldsymbol{x}'\| = \sqrt{(x - x')^2 + (y - y')^2}.$$

Dans $\mathbb{R}^3$, la norme du vecteur $\boldsymbol{x} - \boldsymbol{x}'$ où $\boldsymbol{x} = (x, y, z) \in \mathbb{R}^3$ et $\boldsymbol{x}' = (x', y', z') \in \mathbb{R}^3$ est donnée par le théorème de Pythagore :

$$\|\boldsymbol{x} - \boldsymbol{x}'\| = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}.$$

Nous généralisons ces formules à $\mathbb{R}^m$ de la manière suivante.

Définition 1. La norme du vecteur $\boldsymbol{x} := (x_1, \ldots, x_m) \in \mathbb{R}^m$ est définie par :

$$\|\boldsymbol{x}\| := \sqrt{x_1^2 + \cdots + x_m^2}.$$

La distance entre deux éléments $\boldsymbol{x} \in \mathbb{R}^m$ et $\boldsymbol{y} \in \mathbb{R}^m$ est égale à $\|\boldsymbol{x} - \boldsymbol{y}\|$.

Exemples.

1. Dans $\mathbb{R}^m$, la norme des vecteurs de la base canonique est égale à 1.
2. La norme de la matrice $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{R})$ est :

$$\|A\| = \sqrt{a^2 + b^2 + c^2 + d^2}.$$

Notons que si $\|\boldsymbol{x} - \boldsymbol{y}\| = 0$, alors $\boldsymbol{x} = \boldsymbol{y}$. En effet, si :

$$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = (x_1 - y_1)^2 + \cdots + (x_m - y_m)^2 = 0,$$

chaque terme de la somme doit être nul (car les termes sont tous positifs) et les coordonnées des deux vecteurs sont donc égales.

Proposition 1. Pour tout $\boldsymbol{x} := (x_1, \ldots, x_m) \in \mathbb{R}^m$, on a $\|\boldsymbol{x}\| \leqslant |x_1| + \cdots + |x_m|$.

Démonstration. On compare les carrés des deux termes :

$$\|\boldsymbol{x}\|^2 = x_1^2 + \cdots + x_m^2 \leqslant \big(|x_1| + \cdots + |x_m|\big)^2$$

car dans le terme de droite, il faut rajouter les double produits $2|x_i||x_j|$ qui sont positifs. ■

Le produit scalaire de deux vecteurs joue un rôle particulier.

Définition 2.
Le produit scalaire des vecteurs $\boldsymbol{x} := (x_1, \ldots, x_m)$ et $\boldsymbol{y} := (y_1, \ldots, y_m)$ est :

$$\big(\boldsymbol{x} \mid \boldsymbol{y}\big) := x_1 y_1 + \cdots + x_m y_m.$$

Proposition 2. Pour tout $\boldsymbol{x}$, $\boldsymbol{y}$ et $\boldsymbol{z}$ dans $\mathbb{R}^m$ et pour tout $\lambda \in \mathbb{R}$, on a :

1. $(\boldsymbol{x} \mid \boldsymbol{x}) = \|\boldsymbol{x}\|^2$,
2. $(\boldsymbol{x} \mid \boldsymbol{y}) = (\boldsymbol{y} \mid \boldsymbol{x})$,
3. $(\lambda\boldsymbol{x} + \boldsymbol{y} \mid \boldsymbol{z}) = \lambda(\boldsymbol{x} \mid \boldsymbol{z}) + (\boldsymbol{y} \mid \boldsymbol{z})$.

Démonstration. Les points 1 et 2 sont évidents. Le point 3 consiste à observer que pour tout $\boldsymbol{z} \in \mathbb{R}^m$, l'application de $\mathbb{R}^m$ dans $\mathbb{R}$ qui à $\boldsymbol{x}$ associe $(\boldsymbol{x} \mid \boldsymbol{z})$ est une application linéaire. Cela est immédiat puisque l'expression de $(\boldsymbol{x} \mid \boldsymbol{z})$ est une combinaison linéaire des x_i. ■

Proposition 3 (Inégalité de Cauchy-Schwarz).
Pour tout $\boldsymbol{x}$ et tout $\boldsymbol{y}$ dans $\mathbb{R}^m$, on a

$$\big|(\boldsymbol{x} \mid \boldsymbol{y})\big| \leqslant \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$$

avec égalité si, et seulement si, $\boldsymbol{x}$ et $\boldsymbol{y}$ sont colinéaires.

Démonstration. Pour tout $\boldsymbol{x}$ et tout $\boldsymbol{y}$ dans $\mathbb{R}^m$ et pour tout $\lambda \in \mathbb{R}$, on pose :

$$Q(\lambda) := (\boldsymbol{x} + \lambda\boldsymbol{y} \mid \boldsymbol{x} + \lambda\boldsymbol{y}) = \|\boldsymbol{x} + \lambda\boldsymbol{y}\|^2 \geqslant 0.$$

Notons que :

$$Q(\lambda) = (\boldsymbol{x} \mid \boldsymbol{x} + \lambda\boldsymbol{y}) + \lambda(\boldsymbol{y} \mid \boldsymbol{x} + \lambda\boldsymbol{y}) = \|\boldsymbol{x}\|^2 + 2\lambda(\boldsymbol{x} \mid \boldsymbol{y}) + \lambda^2\big\|\boldsymbol{y}\big\|^2.$$

Si $\|\boldsymbol{y}\| = 0$, on a les égalités $\big|(\boldsymbol{x} \mid \boldsymbol{y})\big| = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\| = 0$.
Si $\|\boldsymbol{y}\| \neq 0$, alors Q est un polynôme de degré 2 qui ne prend que des valeurs positives. Son discriminant :

$$\Delta = 4\left((\boldsymbol{x} \mid \boldsymbol{y})^2 - \|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2\right)$$

est donc négatif ou nul, ce qui donne l'inégalité requise. Le discriminant s'annule si, et seulement s'il y a une racine double donc, si, et seulement s'il existe $\lambda_0 \in \mathbb{R}$ tel que $\|\boldsymbol{x} + \lambda_0\boldsymbol{y}\|^2 = 0$, c'est-à-dire $\boldsymbol{x} + \lambda_0\boldsymbol{y} = (0, \ldots, 0)$. ■

Proposition 4 (Inégalité triangulaire). Si $\boldsymbol{x} \in \mathbb{R}^m$ et $\boldsymbol{y} \in \mathbb{R}^m$, alors :

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$$

avec égalité si, et seulement s'il existe $\lambda \geqslant 0$ tel que $\boldsymbol{x} = \lambda\boldsymbol{y}$ ou $\boldsymbol{y} = \lambda\boldsymbol{x}$.

Démonstration. En reprenant la démonstration précédente, on voit que :

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + 2(\boldsymbol{x} \mid \boldsymbol{y}) + \|\boldsymbol{y}\|^2.$$

D'après l'inégalité de Cauchy-Schwarz, on a donc :

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 \leqslant \|\boldsymbol{x}\|^2 + 2\|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\| + \|\boldsymbol{y}\|^2 = \big(\|\boldsymbol{x}\| + \|\boldsymbol{y}\|\big)^2.$$

S'il existe $\lambda \geqslant 0$ tel que $\boldsymbol{x} = \lambda \boldsymbol{y}$, alors $\|\boldsymbol{x}+\boldsymbol{y}\| = (\lambda+1)\|\boldsymbol{y}\| = \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$. De même, s'il existe $\lambda \geqslant 0$ tel que $\boldsymbol{y} = \lambda \boldsymbol{x}$, alors $\|\boldsymbol{x}+\boldsymbol{y}\| = (1+\lambda)\|\boldsymbol{x}\| = \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.
Réciproquement, pour avoir l'égalité, il faut que $(\boldsymbol{x} \mid \boldsymbol{y}) = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$, ce qui impose d'une part que $\boldsymbol{x}$ et $\boldsymbol{y}$ sont colinéaires (pour avoir le cas d'égalité dans l'inégalité de Cauchy-Schwarz), d'autre part que $(\boldsymbol{x} \mid \boldsymbol{y}) \geqslant 0$, ce qui impose que $\boldsymbol{x} = \lambda \boldsymbol{y}$ ou $\boldsymbol{y} = \lambda \boldsymbol{x}$ avec $\lambda \geqslant 0$. ■

Corollaire 5. Si $\boldsymbol{x} \in \mathbb{R}^m$ et $\boldsymbol{y} \in \mathbb{R}^m$, alors :

$$\|\boldsymbol{x}-\boldsymbol{y}\| \geqslant \big|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|\big|.$$

Démonstration. On a $\|\boldsymbol{x}\| = \|\boldsymbol{x}-\boldsymbol{y}+\boldsymbol{y}\| \leqslant \|\boldsymbol{x}-\boldsymbol{y}\| + \|\boldsymbol{y}\|$, donc $\|x\| - \|y\| \leqslant \|x-y\|$. En échangeant les rôles de x et y, on en déduit $\|y\| - \|x\| \leqslant \|y-x\| = \|x-y\|$, d'où le résultat. ■

1.2 Convergence de suites

Nous sommes maintenant en mesure de définir la limite d'une suite vectorielle.

Définition 3. Une suite vectorielle $(\boldsymbol{x}_n)$ de $\mathbb{R}^m$ *converge* vers $\boldsymbol{a} \in \mathbb{R}^m$ si la suite de terme général $\|\boldsymbol{x}_n - \boldsymbol{a}\|$ tend vers 0, autrement dit :

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \geqslant N \quad \|\boldsymbol{x}_n - \boldsymbol{a}\| \leqslant \varepsilon.$$

On dit que $\boldsymbol{a}$ est la *limite* de la suite $(\boldsymbol{x}_n)$. Nous noterons $\boldsymbol{x}_n \underset{n\to+\infty}{\longrightarrow} \boldsymbol{a}$ ou bien plus simplement $\boldsymbol{x}_n \to \boldsymbol{a}$.

Exercice 1.

Soit $M \in M_m(\mathbb{R})$ une matrice $m \times m$; pour $n \geqslant 1$ posons : $M_n := \mathrm{I}_m + \dfrac{1}{n}M$.
Montrer que la suite $(M_n)_{n\geqslant 1}$ converge et déterminer la limite.

Solution. Comme $\|M_n - I_m\| = \dfrac{1}{n}\|M\| \underset{n\to+\infty}{\longrightarrow} 0$, la suite $(M_n)_{n\geqslant 1}$ converge vers I_m.

Dans de nombreux cas, nous ne vérifierons pas que la suite $(\boldsymbol{x}_n)$ converge vers $\boldsymbol{a}$ en montrant que $\|\boldsymbol{x}_n - \boldsymbol{a}\|$ tend vers 0. Nous montrerons plutôt que la suite converge coordonnée par coordonnée.

Théorème 6.
La suite vectorielle $\big(\boldsymbol{x}_n := (x_{1,n}, \ldots, x_{m,n})\big)_{n\in\mathbb{N}}$ converge vers $\boldsymbol{a} := (a_1, \ldots, a_m)$ si, et seulement si, pour tout entier $i \in [\![1, m]\!]$, la suite $(x_{i,n})_{n\in\mathbb{N}}$ converge vers a_i.

Démonstration. Pour tout entier $i \in [\![1, m]\!]$, on a :

$$0 \leqslant |x_{i,n} - a_i| \leqslant \|\boldsymbol{x}_n - \boldsymbol{a}\| \leqslant |x_{1,n} - a_1| + \cdots + |x_{m,n} - a_m|.$$

Si la suite $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ converge vers $\boldsymbol{a}$, la suite réelle $\big(\|\boldsymbol{x}_n - \boldsymbol{a}\|\big)_{n\in\mathbb{N}}$ converge vers 0, et donc la suite réelle $\big(|x_{i,n} - a_i|\big)_{n\in\mathbb{N}}$ converge aussi vers 0 d'après le théorème des gendarmes.

Cela démontre que la suite $(x_{i,n})_{n\in\mathbb{N}}$ converge vers a_i.
Réciproquement, si pour tout $i \in [\![1, m]\!]$ la suite $(x_{i,n})_{n\in\mathbb{N}}$ converge vers a_i, alors la suite réelle :

$$\big(|x_{1,n} - a_1| + \cdots + |x_{m,n} - a_m|\big)_{n\in\mathbb{N}}$$

converge vers 0. En effet, c'est la somme de m suites convergeant vers 0. De nouveau, d'après le théorème des gendarmes, la suite $\big(\|\boldsymbol{x}_n - \boldsymbol{a}\|\big)_{n\in\mathbb{N}}$ converge vers 0 et donc, la suite $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ converge vers $\boldsymbol{a}$. ■

Nous avons défini les notions de convergence et de limite d'une suite à l'aide des coordonnées des vecteurs dans la base canonique de $\mathbb{R}^m$. Nous allons maintenant montrer que les notions de convergence et de limite ne dépendent pas de la base dans laquelle nous travaillons. Ce résultat n'est pas immédiat puisque la définition même de la norme d'un vecteur est donnée à l'aide des coordonnées de ce vecteur dans la base canonique de $\mathbb{R}^m$. Nous reparlerons de tout cela dans le volume de L2. Le résultat suivant généralise le théorème 6 de la page précédente.

Théorème 7. Soit $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$ une base de $\mathbb{R}^m$. La suite vectorielle $(\boldsymbol{x}_n := x_{1,n}\boldsymbol{v}_1 + \cdots + x_{m,n}\boldsymbol{v}_m)_{n\in\mathbb{N}}$ converge vers $\boldsymbol{a} := a_1\boldsymbol{v}_1 + \cdots + a_m\boldsymbol{v}_m$ si, et seulement si, pour tout entier $i \in [\![1, m]\!]$, la suite $(x_{i,n})_{n\in\mathbb{N}}$ converge vers a_i.

Démonstration. Soient $x'_{i,n}$ les coordonnées de $x_{i,n}$ dans la base canonique $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m)$ de $\mathbb{R}^m$ et a'_i les coordonnées de $\boldsymbol{a}$. Alors, pour chaque entier $i \in [\![1, m]\!]$, les $x_{i,n}$ s'expriment comme combinaison linéaire des $x'_{i,n}$ et vice-versa :

$$x_{i,n} = A_1 x'_{1,n} + \cdots + A_m x'_{m,n}, \quad a_i = A_1 a'_1 + \cdots + A_m a'_n,$$
$$x'_{i,n} = B_1 x_{1,n} + \cdots + B_m x_{m,n} \quad \text{et} \quad a'_i = B_1 a_1 + \cdots + B_m a_n.$$

Les coefficients A_1, …, A_m, B_1,…, B_m ne dépendent pas de n. Ils sont obtenus à partir des matrices de passage de la base $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m)$ à la base $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$ et de son inverse.
On voit donc que pour tout i la suite $(x'_{i,n})_{n\in\mathbb{N}}$ converge vers a'_i, si, et seulement si, pour tout i la suite $(x_{i,n})_{n\in\mathbb{N}}$ converge vers a_i. ■

En travaillant coordonnée par coordonnée, on montre facilement que comme dans le cas des suites numériques, il y a unicité de la limite d'une suite vectorielle. On obtient également le résultat suivant.

Théorème 8. Soient $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ et $(\boldsymbol{y}_n)_{n\in\mathbb{N}}$ deux suites vectorielles qui convergent respectivement vers $\boldsymbol{x}$ et $\boldsymbol{y}$ et soit $(\lambda_n)_{n\in\mathbb{N}}$ une suite réelle qui converge vers λ. Alors

1. la suite $(\boldsymbol{x}_n + \boldsymbol{y}_n)_{n\in\mathbb{N}}$ converge vers $\boldsymbol{x} + \boldsymbol{y}$,
2. la suite $(\lambda_n \boldsymbol{x}_n)_{n\in\mathbb{N}}$ converge vers $\lambda\boldsymbol{x}$ et
3. la suite de produits scalaires $\big((\boldsymbol{x}_n \mid \boldsymbol{y}_n)\big)_{n\in\mathbb{N}}$ converge vers $(\boldsymbol{x} \mid \boldsymbol{y})$.

Démonstration. Pour (1) et (2), il suffit de travailler coordonnée par coordonnée. Chaque coordonnée de $\boldsymbol{x}_n + \boldsymbol{y}_n$ est la somme de deux suites réelles convergentes. Chaque co-

ordonnée de $\lambda_n \boldsymbol{x}_n$ est le produit de deux suites réelles convergentes. Pour (3), il suffit d'observer que le produit scalaire $(\boldsymbol{x}_n \mid \boldsymbol{y}_n)$ est une somme de produits de suites réelles convergentes. ■

Venons en maintenant au cas des suites bornées.

Définition 4 (Suites bornées).
La suite vectorielle $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ est bornée si la suite réelle $\big(\|\boldsymbol{x}_n\|\big)_{n ßn\mathbb{N}}$ est bornée.

Proposition 9. La suite vectorielle $\big(\boldsymbol{x}_n := (x_{1,n}, \ldots, x_{m,n})\big)_{n\in\mathbb{N}}$ est bornée si, et seulement si, pour tout entier $i \in [\![1, m]\!]$, la suite $(x_{i,n})_{n\in\mathbb{N}}$ est bornée.

Démonstration. Si la suite $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ est bornée, alors les suites $(x_{i,n})_{n\in\mathbb{N}}$ le sont puisque :

$$|x_{i,n}| \leqslant \|\boldsymbol{x}_n\|.$$

Réciproquement, si pour tout entier $i \in [\![1, m]\!]$, la suite $(x_{i,n})_{n\in\mathbb{N}}$ est bornée, alors il existe M tel que pour tout entier $i \in [\![1, m]\!]$ et pour tout $n \in \mathbb{N}$, $|x_{i,n}| \leqslant M$. On a alors :

$$\|\boldsymbol{x}_n\| \leqslant |x_{1,n}| + \cdots + |x_{m,n}| \leqslant m \cdot M.$$ ■

Proposition 10. Toute suite vectorielle convergente est bornée.

Démonstration. Si la suite vectorielle $\big(\boldsymbol{x}_n := (x_{1,n}, \ldots, x_{m,n})\big)_{n\in\mathbb{N}}$ converge, alors chaque suite $(x_{i,n})_{n\in\mathbb{N}}$ converge et donc, est bornée. La suite $(\boldsymbol{x}_n)$ est donc bornée. ■

Le théorème de Bolzano-Weierstraß, qui est en quelque sorte un énoncé réciproque, reste valide pour les suites vectorielles.

Théorème 11 (Bolzano-Weierstraß). Toute suite vectorielle bornée $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ contient une suite extraite $(\boldsymbol{x}_{n_k})_{k\in\mathbb{N}}$ convergente.

Démonstration. On peut travailler coordonnée par coordonnée. Pour cela, on écrit $\boldsymbol{x}_n = (x_{1,n}, \ldots, x_{m,n})$. La suite réelle $(x_{1,n})_{n\in\mathbb{N}}$ est bornée. On peut donc extraire une sous-suite $(\boldsymbol{x}_{n_k})_{k\in\mathbb{N}}$ de sorte que la suite $(x_{1,n_k})_{k\in\mathbb{N}}$ soit convergente.

De manière similaire, de la suite $(\boldsymbol{x}_{n_k})_{k\in\mathbb{N}}$, on peut extraire une sous-suite $(\boldsymbol{x}_{n_{k_j}})_{j\in\mathbb{N}}$ de sorte que la suite $(x_{2,n_{k_j}})_{j\in N}$ converge. La suite $(x_{1,n_{k_j}})_{j\in N}$ converge également puisque c'est une sous-suite d'une suite convergente.

En répétant le procédé, on peut extraire une sous-suite de la suite $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ de sorte que chaque suite de coordonnées soit convergente. La sous-suite extraite est alors convergente. ■

1.3 Suites vectorielles définies par une récurrence linéaire

Rappelons qu'une suite géométrique $(u_n)_{n\in\mathbb{N}}$ de premier terme a et de raison r est définie par récurrence par $u_0 := a$ et $u_{n+1} := ru_n$. Nous avons vu que dans ce cas $u_n = ar^n$ pour tout $n \in \mathbb{N}$.

Nous allons maintenant étudier une généralisation de cette notion dans le cas des suites vectorielles. Considérons une suite $(\boldsymbol{u}_n)_{n\in\mathbb{N}}$ définie par récurrence par :

$$\boldsymbol{u}_0 \in \mathbb{R}^m \quad \text{et} \quad \boldsymbol{u}_{n+1} := A \cdot \boldsymbol{u}_n$$

où $A \in M_m(\mathbb{R})$ est une matrice à m lignes et m colonnes.

Exercice 2.

Déterminer l'expression de la suite $\big((x_n, y_n)\big)_{n\in\mathbb{N}}$ définie par :

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \in \mathbb{R}^2 \quad \text{et} \quad \begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} := \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \cdot \begin{pmatrix} x_n \\ y_n \end{pmatrix}.$$

Solution. On peut écrire la relation de récurrence sous la forme :

$$\begin{cases} x_{n+1} = 2x_n \\ y_{n+1} = 3y_n. \end{cases}$$

La suite (x_n) est une suite géométrique de premier terme x_0 et de raison 2. On a donc $x_n = 2^n x_0$ pour tout $n \in \mathbb{N}$. La suite (y_n) est une suite géométrique de premier terme y_0 et de raison 3. On a donc $y_n = 3^n y_0$ pour tout $n \in \mathbb{N}$. On peut écrire cela sous la forme :

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = \begin{pmatrix} 2^n & 0 \\ 0 & 3^n \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}^n \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$

De manière générale, on a le résultat suivant.

Théorème 12.
Si $\boldsymbol{u}_0 \in \mathbb{R}^m$ et $\boldsymbol{u}_{n+1} := A{\cdot}\boldsymbol{u}_n$, alors $\boldsymbol{u}_n = A^n{\cdot}\boldsymbol{u}_0$ où A^n est la puissance $n^{\text{ème}}$ de A.

Démonstration. La démonstration se fait par récurrence. D'abord, pour $n = 0$, le résultat est vrai puisque, par définition, A^0 est la matrice identité. Ensuite, si $\boldsymbol{u}_n = A^n \cdot \boldsymbol{u}_0$, alors :

$$\boldsymbol{u}_{n+1} = A \cdot (A^n \cdot \boldsymbol{u}_0) = A^{n+1} \cdot \boldsymbol{u}_0.$$

Si le résultat est vrai au rang n, alors il est vrai au rang $n+1$. ∎

Le problème consiste alors à déterminer A^n. Nous renvoyons pour cela le lecteur au module II.7. Nous proposons en exercice des exemples dans le cas $m = 2$.

Exercice 3.

Soit $(\boldsymbol{u}_n)$ une suite définie par récurrence par :

$$\boldsymbol{u}_0 \in \mathbb{R}^2 \quad \text{et} \quad \boldsymbol{u}_{n+1} := A \cdot \boldsymbol{u}_n$$

avec $A \in M_2(\mathbb{R})$ une matrice 2×2.

1. On suppose que A est diagonalisable dans une base $(\boldsymbol{v}, \boldsymbol{w})$ de vecteurs propres. Soient λ et μ les valeurs propres associées (éventuellement égales). Montrer que pour tout $n \in \mathbb{N}$:
$$\boldsymbol{u}_n = a\lambda^n \boldsymbol{v} + b\mu^n \boldsymbol{w}$$
a et b étant les coefficients de la décomposition de $\boldsymbol{u}_0$ dans la base $(\boldsymbol{v}, \boldsymbol{w})$.
2. Déterminer la forme générale de la suite définie par récurrence par :
$$\boldsymbol{u}_0 := \begin{pmatrix} 3 \\ 7 \end{pmatrix} \quad \text{et} \quad \boldsymbol{u}_{n+1} := \begin{pmatrix} 0 & 1 \\ -6 & 5 \end{pmatrix} \cdot \boldsymbol{u}_n.$$
3. On suppose que A a une valeur propre double λ, et que $(\boldsymbol{v}, \boldsymbol{w})$ est une base de $\mathbb{R}^2$ telle que :
$$A\boldsymbol{v} = \lambda \boldsymbol{v} \quad \text{et} \quad A\boldsymbol{w} = \lambda(\boldsymbol{w} + \boldsymbol{v}).$$
Montrer que pour tout $n \in \mathbb{N}$:
$$\boldsymbol{u}_n = (a + bn)\lambda^n \boldsymbol{v} + b\lambda^n \boldsymbol{w}$$
a et b étant les coefficients de la décomposition de $\boldsymbol{u}_0$ dans la base $(\boldsymbol{v}, \boldsymbol{w})$.
4. Déterminer la forme générale de la suite définie par récurrence par :
$$\boldsymbol{u}_0 := \begin{pmatrix} 1 \\ -2 \end{pmatrix} \quad \text{et} \quad \boldsymbol{u}_{n+1} := \begin{pmatrix} 0 & 1 \\ -4 & 4 \end{pmatrix} \cdot \boldsymbol{u}_n.$$

Solution.

1. On veut déterminer $\boldsymbol{u}_n = A^n \boldsymbol{u}_0$.
On sait que $A\boldsymbol{v} = \lambda \boldsymbol{v}$ et $A\boldsymbol{w} = \mu \boldsymbol{w}$. Donc, $A^n \boldsymbol{v} = \lambda^n \boldsymbol{v}$ et $A^n \boldsymbol{w} = \mu^n \boldsymbol{w}$.
Puisque $\boldsymbol{u}_0 = a\boldsymbol{v} + b\boldsymbol{w}$, et puisque l'application $\boldsymbol{u} \mapsto A^n \boldsymbol{u}$ est une application linéaire, on a :
$$A^n \boldsymbol{u}_0 = A^n \cdot (a\boldsymbol{v} + b\boldsymbol{w}) = aA^n \boldsymbol{v} + bA^n \boldsymbol{w} = a\lambda^n \boldsymbol{v} + b\mu^n \boldsymbol{w}.$$
2. Posons $A := \begin{pmatrix} 0 & 1 \\ -6 & 5 \end{pmatrix}$. Alors :
$$A \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{et} \quad A \cdot \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 9 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$
Si l'on définit $\boldsymbol{v} := (1, 2)$ et $\boldsymbol{w} := (1, 3)$, alors $(\boldsymbol{v}, \boldsymbol{w})$ est une base de $\mathbb{R}^2$ formée de vecteurs propres. D'après la question précédente :
$$\boldsymbol{u}_0 = \begin{pmatrix} 3 \\ 7 \end{pmatrix} = 2 \cdot \boldsymbol{v} + 1 \cdot \boldsymbol{w} \quad \Longrightarrow \quad \boldsymbol{u}_n = 2 \cdot 2^n \boldsymbol{v} + 1 \cdot 3^n \boldsymbol{w} = \begin{pmatrix} 2^{n+1} + 3^n \\ 2^{n+2} + 3^{n+1} \end{pmatrix}.$$
3. On procède essentiellement comme dans 1. On a :
$$A\boldsymbol{v} = \lambda \boldsymbol{v} \quad \text{et} \quad A\boldsymbol{w} = \lambda(\boldsymbol{v} + \boldsymbol{w}).$$
On en déduit que pour tout $n \in \mathbb{N}$, $A^n \boldsymbol{v} = \lambda^n \boldsymbol{v}$, puis par récurrence que :
$$\forall n \in \mathbb{N}, \ A^n \boldsymbol{w} = \lambda^n (\boldsymbol{w} + n\boldsymbol{v}).$$

En effet, cette propriété est vraie pour $n = 0$ puisque $A^0\boldsymbol{w} = \boldsymbol{w} = \lambda^0(\boldsymbol{w} + 0\boldsymbol{v})$. Et si elle est vraie au rang n, alors elle est vraie au rang $n+1$:

$$\begin{aligned} A^{n+1}\boldsymbol{w} = A \cdot A^n\boldsymbol{w} &= A \cdot \big(\lambda^n(\boldsymbol{w} + n\boldsymbol{v})\big) \\ &= \lambda^n(A\boldsymbol{w} + nA\boldsymbol{v}) \\ &= \lambda^n(\lambda\boldsymbol{w} + \lambda\boldsymbol{v} + n\lambda\boldsymbol{v}) = \lambda^{n+1}\big(\boldsymbol{w} + (n+1)\boldsymbol{v}\big). \end{aligned}$$

Par conséquent :

$$\boldsymbol{u}_n = A^n(a\boldsymbol{v} + b\boldsymbol{w}) = a\lambda^n\boldsymbol{v} + b\lambda^n(\boldsymbol{w} + n\boldsymbol{v}) = (a + bn)\lambda^n\boldsymbol{v} + b\lambda^n\boldsymbol{w}.$$

4. Posons $A := \begin{pmatrix} 0 & 1 \\ -4 & 4 \end{pmatrix}$. Alors :

$$A \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{et} \quad A \cdot \begin{pmatrix} 1 \\ 4 \end{pmatrix} = 2\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 \\ 4 \end{pmatrix}\right).$$

Si l'on définit $\boldsymbol{v} := (1, 2)$ et $\boldsymbol{w} := (1, 4)$, alors $(\boldsymbol{v}, \boldsymbol{w})$ est une base de $\mathbb{R}^2$ vérifiant les hypothèse de la question précédente. On a :

$$\boldsymbol{u}_0 = \begin{pmatrix} 1 \\ -2 \end{pmatrix} = 3\boldsymbol{v} - 2\boldsymbol{w}.$$

D'après la question précédente, on a donc :

$$\boldsymbol{u}_n = (3 - 2n) \cdot 2^n\boldsymbol{v} - 2 \cdot 2^n\boldsymbol{w} = \begin{pmatrix} 2^n(1 - 2n) \\ 2^{n+1}(1 - 2(n+1)) \end{pmatrix}.$$

1.4 Suites réelles définies par une récurrence d'ordre 2

Définition 5.
Une suite réelle $(u_n)_{n\in\mathbb{N}}$ est définie par une récurrence linéaire d'ordre 2 si :

$$u_0 \in \mathbb{R}, \quad u_1 \in \mathbb{R} \quad \text{et} \quad u_{n+2} := au_{n+1} + bu_n$$

avec $a \in \mathbb{R}$ et $b \in \mathbb{R}$.

Remarquons que la relation de récurrence relie trois termes consécutifs. On pourrait étudier plus généralement les suites définies par une récurrence linéaire d'ordre k, c'est-à-dire reliant $k+1$ termes consécutifs :

$$u_0 \in \mathbb{R}, \ldots, u_{k-1} \in \mathbb{R} \quad \text{et} \quad u_{n+k} := a_{k-1}u_{n+k-1} + \cdots + a_0u_n$$

avec $a_0 \in \mathbb{R}$, …, $a_{k-1} \in \mathbb{R}$. Les techniques sont essentiellement les mêmes et nous nous limiterons au cas des récurrences d'ordre 2.

Il est très important de comprendre l'astuce que nous allons utiliser maintenant pour étudier de telles suites. Une astuce similaire sera utilisée plus loin pour étudier les équations différentielles linéaires d'ordre 2.

Théorème 13. Si (u_n) est une suite définie par récurrence par :

$$u_0 \in \mathbb{R}, \quad u_1 \in \mathbb{R} \quad \text{et} \quad u_{n+2} := au_{n+1} + bu_n,$$

alors la suite $(\boldsymbol{x}_n)$ définie par $\boldsymbol{x}_n := \begin{pmatrix} u_n \\ u_{n+1} \end{pmatrix}$ vérifie la récurrence :

$$\boldsymbol{x}_0 = \begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \quad \text{et} \quad \boldsymbol{x}_{n+1} = \begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix} \cdot \boldsymbol{x}_n.$$

Démonstration. Il suffit d'écrire :

$$\begin{pmatrix} u_{n+1} \\ u_{n+2} \end{pmatrix} = \begin{pmatrix} 0 \cdot u_n + 1 \cdot u_{n+1} \\ b \cdot u_n + a \cdot u_{n+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ b & a \end{pmatrix} \cdot \begin{pmatrix} u_n \\ u_{n+1} \end{pmatrix}.$$

L'étude d'une suite réelle définie par une récurrence linéaire d'ordre 2 se ramène donc à l'étude d'une suite vectorielle de $\mathbb{R}^2$ définie par une récurrence linéaire d'ordre 1.

Exercice 4.

Étudier la suite définie par récurrence par :

$$u_0 := 3, \; u_1 := 7 \qquad \text{et} \qquad u_{n+1} := 5u_n - 6u_{n-1}.$$

Solution. On peut écrire :

$$\begin{pmatrix} u_n \\ u_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \cdot u_{n-1} + 1 \cdot u_n \\ -6 \cdot u_{n-1} + 5 \cdot u_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -6 & 5 \end{pmatrix} \cdot \begin{pmatrix} u_{n-1} \\ u_n \end{pmatrix}.$$

Considérons la suite vectorielle $(\boldsymbol{x}_n)$ définie par $\boldsymbol{x}_n := \begin{pmatrix} u_n \\ u_{n+1} \end{pmatrix}$. Alors :

$$\boldsymbol{x}_0 = \begin{pmatrix} 3 \\ 7 \end{pmatrix} \qquad \text{et} \qquad \boldsymbol{x}_{n+1} = \begin{pmatrix} 0 & 1 \\ -6 & 5 \end{pmatrix} \cdot \boldsymbol{x}_n.$$

L'étude de cette suite est proposée dans l'exercice 3 de la page 785. On a :

$$\boldsymbol{x}_n = \begin{pmatrix} 2^{n+1} + 3^n \\ 2^{n+2} + 3^{n+1} \end{pmatrix}.$$

Puisque u_n est la première coordonnée de $\boldsymbol{x}_n$, on en déduit que :

$$u_n = 2^{n+1} + 3^n.$$

Exercice 5.

Étudier la suite définie par récurrence par :

$$u_0 := 1, \; u_1 := -2 \qquad \text{et} \qquad u_{n+1} := 4u_n - 4u_{n-1}.$$

Solution. On peut écrire :

$$\begin{pmatrix} u_n \\ u_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \cdot u_{n-1} + 1 \cdot u_n \\ -4 \cdot u_{n-1} + 4 \cdot u_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -4 & 4 \end{pmatrix} \cdot \begin{pmatrix} u_{n-1} \\ u_n \end{pmatrix}.$$

Considérons la suite vectorielle $(\boldsymbol{x}_n)$ définie par $\boldsymbol{x}_n := \begin{pmatrix} u_n \\ u_{n+1} \end{pmatrix}$. Alors :

$$\boldsymbol{x}_0 = \begin{pmatrix} 1 \\ -2 \end{pmatrix} \qquad \text{et} \qquad \boldsymbol{x}_{n+1} = \begin{pmatrix} 0 & 1 \\ -4 & 4 \end{pmatrix} \cdot \boldsymbol{x}_n.$$

L'étude de cette suite est proposée dans l'exercice 3 de la page 785. Puisque u_n est la première coordonnée de $\boldsymbol{x}_n$, on obtient :

$$u_n = 2^n(1-2n).$$

---

2 Fonctions vectorielles

Définition 6. On appelle fonction vectorielle de la variable réelle toute application $\boldsymbol{f}$ d'un sous-ensemble non vide de $\mathbb{R}$, et à valeurs dans $\mathbb{R}^m$. L'ensemble de départ est appelé ensemble de définition de $\boldsymbol{f}$. On le note $D_{\boldsymbol{f}}$.

Une fonction réelle sera notée $f : D_f \subset \mathbb{R} \to \mathbb{R}$ tandis qu'une fonction vectorielle sera notée $\boldsymbol{f} : D_{\boldsymbol{f}} \subset \mathbb{R} \to \mathbb{R}^m$. Une telle fonction vectorielle est donnée par m fonctions réelles $f_1 : D_{\boldsymbol{f}} \to \mathbb{R}$, $f_2 : D_{\boldsymbol{f}} \to \mathbb{R}$, ..., $f_m : D_{\boldsymbol{f}} \to \mathbb{R}$.
Ce sont les m coordonnées de $\boldsymbol{f} = (f_1, f_2, \ldots, f_m)$.

2.1 Continuité

On peut étendre les notions de limite et continuité des fonctions réelles au cas des fonctions vectorielles.

Définition 7. Une fonction vectorielle $\boldsymbol{f} : D_{\boldsymbol{f}} \to \mathbb{R}^m$ admet une limite $\boldsymbol{\ell} \in \mathbb{R}^m$ au point a, si pour toute suite $(x_n) \subset D_{\boldsymbol{f}} \setminus \{a\}$ de limite a, la suite $\big(\boldsymbol{f}(x_n)\big)$ converge vers ℓ. On écrit alors $\ell = \lim\limits_{x \to a} \boldsymbol{f}(x)$ ou $\ell = \lim\limits_{a} \boldsymbol{f}$.

Comme dans le cas des fonctions réelles, on peut donner une définition équivalente.

Définition 8. Une fonction vectorielle $\boldsymbol{f} : D_{\boldsymbol{f}} \to \mathbb{R}^m$ admet une limite $\boldsymbol{\ell} \in \mathbb{R}^m$ au point a, si :

$$\forall \varepsilon > 0,\ \exists \eta > 0 \quad \big(x \in D_{\boldsymbol{f}} \ \text{ et } \ 0 < |x-a| \leqslant \eta\big) \quad \Longrightarrow \quad \big\|\boldsymbol{f}(x) - \boldsymbol{f}(a)\big\| \leqslant \varepsilon.$$

Pour déterminer la limite d'une fonction vectorielle, on travaille généralement coordonnée par coordonnée, grâce au théorème suivant.

Théorème 14. Soient $\boldsymbol{f} := (f_1, \ldots, f_m) : D_{\boldsymbol{f}} \to \mathbb{R}^m$ une fonction vectorielle et $\boldsymbol{\ell} := (\ell_1, \ldots, \ell_m)$ un vecteur de $\mathbb{R}^m$. Alors, $\lim\limits_{x \to a} \boldsymbol{f}(x) = \ell$ si, et seulement si, pour tout entier $i \in [\![1, m]\!]$, $\lim\limits_{x \to a} f_i(x) = \ell_i$.

Démonstration. Soit $(x_k)_{k \in \mathbb{N}}$ une suite qui converge vers a.

D'après le théorème 6 de la page 782, la suite $\big(\boldsymbol{f}(x_k)\big)_{k\in\mathbb{N}}$ converge vers ℓ si, et seulement si, pour tout entier $i \in [\![1, m]\!]$, la suite $\big(f_i(x_k)\big)_{k\in\mathbb{N}}$ converge vers ℓ_i. ■

Comme dans le cas des fonctions réelles, on peut alors introduire la notion de continuité d'une fonction vectorielle en un point a.

Définition 9. Une fonction vectorielle $\boldsymbol{f} : D_{\boldsymbol{f}} \to \mathbb{R}^m$ est continue en a si $a \in D_{\boldsymbol{f}}$ et si $\lim\limits_{x\to a} \boldsymbol{f}(x) = \boldsymbol{f}(a)$.

On a la définition équivalente suivante.

Définition 10. La fonction $\boldsymbol{f} : D_{\boldsymbol{f}} \to \mathbb{R}^m$ est continue en a si $a \in D_{\boldsymbol{f}}$ et si :

$$\forall \varepsilon > 0,\ \exists \eta > 0, \quad |x - a| \leqslant \eta \quad \Longrightarrow \quad \big\|\boldsymbol{f}(x) - \boldsymbol{f}(a)\big\| \leqslant \varepsilon.$$

Définition 11. La fonction $\boldsymbol{f}$ est continue sur un sous-ensemble $I \subset \mathbb{R}$ si $\boldsymbol{f}$ est continue en tout point de I.

La continuité d'une fonction vectorielle se teste généralement coordonnée par coordonnée, grâce à la propriété suivante.

Théorème 15. Une fonction $\boldsymbol{f} := (f_1, \ldots, f_m) : D_{\boldsymbol{f}} \to \mathbb{R}^m$ est continue en $a \in D_{\boldsymbol{f}}$ si, et seulement si, chaque fonction $f_i : D_{\boldsymbol{f}} \to \mathbb{R}$ est continue en a.

Démonstration.
C'est une conséquence immédiate du théorème 14 de la page précédente. ■

Théorème 16 (Opérations élémentaires). Soient $\boldsymbol{f} : D \to \mathbb{R}^m$, $\boldsymbol{g} : D \to \mathbb{R}^m$ et $\lambda : D \to \mathbb{R}$ des fonctions continues en $a \in D$. Alors,

1. $\boldsymbol{f} + \boldsymbol{g}$ est continue en a ;
2. $\lambda\boldsymbol{f}$ est continue en a ;
3. la fonction produit scalaire $(\,\boldsymbol{f} \mid \boldsymbol{g}\,) : \boldsymbol{x} \mapsto (\,\boldsymbol{f}(\boldsymbol{x}) \mid \boldsymbol{g}(\boldsymbol{x})\,)$ est continue en a.

Démonstration. Soit (x_k) une suite qui converge vers a.
On a $\boldsymbol{f}(x_k) \underset{k\to+\infty}{\longrightarrow} \boldsymbol{f}(a)$, $\boldsymbol{g}(x_k) \underset{k\to+\infty}{\longrightarrow} \boldsymbol{g}(a)$ et $\lambda(x_k) \underset{k\to+\infty}{\longrightarrow} \lambda(a)$ par continuité de $\boldsymbol{f}$, $\boldsymbol{g}$ et λ. Par conséquent, d'après le théorème 8 de la page 783 :

1. la suite $\big(\boldsymbol{f}(x_k) + \boldsymbol{g}(x_k)\big)$ converge vers $\boldsymbol{f}(a) + \boldsymbol{g}(a)$;
2. la suite $\big(\lambda(x_k)\boldsymbol{f}(x_k)\big)$ converge vers $\lambda(a)\boldsymbol{f}(a)$;
3. la suite $\big(\big(\,\boldsymbol{f}(x_k) \mid \boldsymbol{g}(x_k)\,\big)\big)$ converge vers $\big(\,\boldsymbol{f}(a) \mid \boldsymbol{g}(a)\,\big)$. ■

2.2 Dérivabilité

Nous pouvons étendre la notion de dérivabilité d'une fonction réelle au cas des fonctions vectorielles.

Définition 12. Soient $\boldsymbol{f} : D_{\boldsymbol{f}} \to \mathbb{R}^m$ une fonction vectorielle et a un point de $D_{\boldsymbol{f}}$. On dit que $\boldsymbol{f}$ est dérivable au point a si l'accroissement fini $\dfrac{\boldsymbol{f}(x) - \boldsymbol{f}(a)}{x - a}$ a une limite lorsque x tend vers a. Cette limite est alors notée $\boldsymbol{f}'(a)$.

Définition 13. La fonction $\boldsymbol{f}$ est dérivable sur un sous-ensemble $I \subset \mathbb{R}$ si $\boldsymbol{f}$ est dérivable en tout point de I.

Comme dans le cas de la continuité, la dérivabilité d'une fonction vectorielle se teste généralement coordonnée par coordonnée.

Théorème 17. Une fonction $\boldsymbol{f} := (f_1, \ldots, f_m) : D_{\boldsymbol{f}} \to \mathbb{R}^m$ est dérivable en $a \in D_{\boldsymbol{f}}$ si, et seulement si, chaque fonction $f_i : D_{\boldsymbol{f}} \to \mathbb{R}$ est dérivable en a. On a alors :

$$\boldsymbol{f}'(a) = \big(f_1'(a), \ldots, f_m'(a)\big).$$

Démonstration. Appliquer le théorème 14 de la page 789 à la fonction $\boldsymbol{g} : D_{\boldsymbol{f}} \setminus \{a\} \to \mathbb{R}^m$ définie par $\boldsymbol{g}(x) := \dfrac{\boldsymbol{f}(x) - \boldsymbol{f}(a)}{x - a}$. ■

En travaillant coordonnée par coordonnée, on voit facilement que si $\boldsymbol{f}$ est dérivable en a, alors $\boldsymbol{f}$ est continue en a et que, comme dans le cas des fonctions réelles, la réciproque est fausse.

On peut également définir par récurrence une notion de fonction k fois dérivable sur un intervalle $I \subset \mathbb{R}$.

Définition 14. Soit $\boldsymbol{f} : I \to \mathbb{R}^m$ une fonction vectorielle définie sur un intervalle $I \subset \mathbb{R}$. On définit par récurrence la dérivée $k^{\text{ème}}$ de $\boldsymbol{f}$, notée $\boldsymbol{f}^{(k)}$.

Pour $k = 0$, on pose $\boldsymbol{f}^{(0)} := \boldsymbol{f}$. Soit $k \geqslant 1$; supposons que $\boldsymbol{f}^{(k-1)} : I \to \mathbb{R}^m$ soit définie. Alors, $\boldsymbol{f}$ admet une dérivée $k^{\text{ème}}$ sur I si $\boldsymbol{f}^{(k-1)}$ est dérivable sur I. On note alors $\boldsymbol{f}^{(k)} := \big(\boldsymbol{f}^{(k-1)}\big)'$.

Définition 15. Soit $k \geqslant 0$ un entier et $I \subset \mathbb{R}$ un intervalle. Une fonction $\boldsymbol{f} : I \to \mathbb{R}^m$ est de classe $\mathcal{C}^k$ sur I si $\boldsymbol{f}^{(k)}$ existe et est continue sur I.

Définition 16. Une fonction $\boldsymbol{f} : I \to \mathbb{R}^m$ est de classe $\mathcal{C}^\infty$ sur I si elle est de classe $\mathcal{C}^k$ sur I pour tout entier $k \geqslant 0$.

2.3 Opérations sur les dérivées

Théorème 18. Si $\boldsymbol{f} : I \to \mathbb{R}^m$ est constante, alors $\boldsymbol{f}$ est dérivable en tout point $x \in I$ et $\boldsymbol{f}'(x)$ est le vecteur nul de $\mathbb{R}^m$.

Démonstration.
Pour tout $x \neq a$, le taux d'accroissement $\dfrac{\boldsymbol{f}(x) - \boldsymbol{f}(a)}{x - a}$ est égal au vecteur nul de $\mathbb{R}^m$. ■

Théorème 19. Si $\boldsymbol{f} : I \to \mathbb{R}^m$ et $\boldsymbol{g} : I \to \mathbb{R}^m$ sont dérivables en $a \in I$, alors $\boldsymbol{f} + \boldsymbol{g}$ est dérivable en a et :

$$(\boldsymbol{f} + \boldsymbol{g})'(a) = \boldsymbol{f}'(a) + \boldsymbol{g}'(a).$$

Démonstration. Si $\boldsymbol{f} = (f_1, \ldots, f_m)$ et $\boldsymbol{g} = (g_1, \ldots, g_m)$, alors :

$$\boldsymbol{f} + \boldsymbol{g} = (f_1 + g_1, \ldots, f_m + g_m).$$

On peut alors appliquer le théorème 17 de la page précédente qui permet de travailler coordonnée par coordonnée. Si $\boldsymbol{f}$ et $\boldsymbol{g}$ sont dérivables en a, alors les f_i et les g_i sont dérivables en a. Pour tout entier $i \in [\![1, m]\!]$, $f_i + g_i$ est dérivable en a. La fonction $\boldsymbol{f} + \boldsymbol{g}$ est donc dérivable en a et :

$$(\boldsymbol{f} + \boldsymbol{g})'(a) = \begin{pmatrix} (f_1 + g_1)'(a) \\ \vdots \\ (f_m + g_m)'(a) \end{pmatrix} = \begin{pmatrix} f_1'(a) + g_1'(a) \\ \vdots \\ f_m'(a) + g_m'(a) \end{pmatrix} = \boldsymbol{f}'(a) + \boldsymbol{g}'(a). \quad ■$$

Théorème 20. Si $\lambda : I \to \mathbb{R}$ et $\boldsymbol{f} : I \to \mathbb{R}^m$ sont dérivables en $a \in I$, alors $\lambda \boldsymbol{f}$ est dérivable en a et :

$$(\lambda \boldsymbol{f})'(a) = \underbrace{\lambda(a)}_{\in \mathbb{R}} \cdot \underbrace{\boldsymbol{f}'(a)}_{\in \mathbb{R}^m} + \underbrace{\lambda'(a)}_{\in \mathbb{R}} \cdot \underbrace{\boldsymbol{f}(a)}_{\in \mathbb{R}^m}.$$

Démonstration. Si $\boldsymbol{f} = (f_1, \ldots, f_m)$, alors :

$$\lambda \boldsymbol{f} = (\lambda f_1, \ldots, \lambda f_m).$$

De nouveau, il suffit d'appliquer le théorème 17 de la page précédente qui permet de travailler coordonnée par coordonnée. Si λ et $\boldsymbol{f}$ sont dérivables en a, les fonctions λf_i le sont et $(\lambda f_i)'(a) = \lambda(a) f_i'(a) + \lambda'(a) f_i(a)$. Donc $\lambda \boldsymbol{f}$ est dérivable en a et :

$$\begin{aligned}(\lambda \boldsymbol{f})'(a) &= \begin{pmatrix} \lambda(a) f_1'(a) + \lambda'(a) f_1(a) \\ \vdots \\ \lambda(a) f_m'(a) + \lambda'(a) f_m(a) \end{pmatrix} \\ &= \lambda(a) \begin{pmatrix} f_1'(a) \\ \vdots \\ f_m'(a) \end{pmatrix} + \lambda'(a) \begin{pmatrix} f_1(a) \\ \vdots \\ f_m(a) \end{pmatrix}. \end{aligned}$$ ■

Exercice 6.

Considérons les fonctions $\lambda : \mathbb{R} \to \mathbb{R}$ et $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^3$ définies par $\lambda(x) := x^2 + 3$ et $\boldsymbol{f}(x) := \begin{pmatrix} 1 \\ \sin(x) \\ x \end{pmatrix}$. Déterminer $(\lambda \boldsymbol{f})'$.

Solution. La fonction $\lambda \boldsymbol{f}$ est dérivable sur $\mathbb{R}$ et :

$$\begin{aligned}(\lambda \boldsymbol{f})'(x) &= (x^2+3)\begin{pmatrix} 0 \\ \cos(x) \\ 1 \end{pmatrix} + 2x \begin{pmatrix} 1 \\ \sin(x) \\ x \end{pmatrix} \\ &= \begin{pmatrix} 2x \\ (x^2+3)\cos(x) + 2x\sin(x) \\ (x^2+3) + 2x^2 \end{pmatrix}.\end{aligned}$$

Théorème 21. Si $\boldsymbol{f} : I \to \mathbb{R}^m$ et $\boldsymbol{g} : I \to \mathbb{R}^m$ sont dérivables en $a \in I$, alors le produit scalaire $(\boldsymbol{f} \mid \boldsymbol{g})$ est dérivable en a et :

$$(\boldsymbol{f} \mid \boldsymbol{g})'(a) = \big(\boldsymbol{f}(a) \,\big|\, \boldsymbol{g}'(a)\big) + \big(\boldsymbol{f}'(a) \,\big|\, \boldsymbol{g}(a)\big).$$

Démonstration. Si $\boldsymbol{f} = (f_1, \ldots, f_m)$ et $\boldsymbol{g} = (g_1, \ldots, g_m)$, alors :

$$(\boldsymbol{f} \mid \boldsymbol{g}) = f_1 g_1 + \ldots f_m g_m.$$

On voit donc que si $\boldsymbol{f}$ et $\boldsymbol{g}$ sont dérivables en a, le produit scalaire $(\boldsymbol{f} \mid \boldsymbol{g})$ l'est également et :

$$\begin{aligned}(\boldsymbol{f} \mid \boldsymbol{g})'(a) &= f_1(a)g_1'(a) + f_1'(a)g_1(a) + \cdots + f_m(a)g_m'(a) + f_m'(a)g_m(a) \\ &= \big(f_1(a)g_1'(a) + \cdots + f_m(a)g_m'(a)\big) + \big(f_1'(a)g_1(a) + \cdots + f_m'(a)g_m(a)\big) \\ &= \big(\boldsymbol{f}(a) \,\big|\, \boldsymbol{g}'(a)\big) + \big(\boldsymbol{f}'(a) \,\big|\, \boldsymbol{g}(a)\big).\end{aligned}$$

■

Exercice 7.

Considérons les fonctions $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^3$ et $\boldsymbol{g} : \mathbb{R} \to \mathbb{R}^3$ définies par $\boldsymbol{f}(x) := \begin{pmatrix} x \\ \cos(x) \\ \mathrm{e}^x \end{pmatrix}$ et $\boldsymbol{g}(x) := \begin{pmatrix} 1 \\ \sin(x) \\ x \end{pmatrix}$. Déterminer $(\boldsymbol{f} \mid \boldsymbol{g})'$.

Solution. La fonction $(\boldsymbol{f} \mid \boldsymbol{g})$ est dérivable sur $\mathbb{R}$ et :

$$\begin{aligned}(\boldsymbol{f} \mid \boldsymbol{g})'(x) &= \left(\begin{pmatrix} x \\ \cos(x) \\ \mathrm{e}^x \end{pmatrix} \,\middle|\, \begin{pmatrix} 0 \\ \cos(x) \\ 1 \end{pmatrix}\right) + \left(\begin{pmatrix} 1 \\ -\sin(x) \\ \mathrm{e}^x \end{pmatrix} \,\middle|\, \begin{pmatrix} 1 \\ \sin(x) \\ x \end{pmatrix}\right) \\ &= \cos^2(x) + \mathrm{e}^x + 1 - \sin^2(x) + x\mathrm{e}^x.\end{aligned}$$

Exercice 8.
Montrer que si $A : I \to M_m(\mathbb{R})$ et $\boldsymbol{y} : I \to \mathbb{R}^m$ sont des applications dérivables sur I, alors l'application $\boldsymbol{f} : I \to \mathbb{R}^m$ définie par $\boldsymbol{f}(t) := A(t) \cdot \boldsymbol{y}(t)$ est dérivable sur I et :

$$\forall t \in I\,,\ \boldsymbol{f}'(t) = A'(t) \cdot \boldsymbol{y}(t) + A(t) \cdot \boldsymbol{y}'(t).$$

Solution. Soit L_k la $k^{ème}$ ligne de A. Alors $f_k = (L_k \mid \boldsymbol{y})$ est dérivable et $f_k' = (L_k' \mid y) + (L_k \mid y')$. De plus, L_k' est la $k^{ème}$ ligne de A'.

2.4 Inégalité des accroissements finis

Le théorème des accroissements finis n'est pas valide pour les fonctions vectorielles comme le montre l'exemple suivant.

Exemple.

Considérons la fonction vectorielle $\boldsymbol{f} : [0, 2\pi] \to \mathbb{R}^2$ définie par $\boldsymbol{f}(x) := \begin{pmatrix} \cos(x) \\ \sin(x) \end{pmatrix}$.

L'image du segment $[0, 2\pi]$ est un cercle centré à l'origine. On a $\boldsymbol{f}(2\pi) - \boldsymbol{f}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, mais pour tout $c \in]0, 2\pi[$:

$$\boldsymbol{f}'(c) = \begin{pmatrix} -\sin(c) \\ \cos(c) \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Nous conseillons vivement au lecteur de faire un dessin représentant la courbe $\boldsymbol{f}([0, 2\pi])$ et de regarder comment évolue le vecteur tangent $\boldsymbol{f}'(t)$.

Cependant, il y a une inégalité des accroissements finis.

Théorème 22 (Inégalité des accroissements finis). Si $\boldsymbol{f} : [a, b] \to \mathbb{R}^m$ est continue sur le segment $[a, b]$ et dérivable en tout point de $]a, b[$, alors :

$$\big\|\boldsymbol{f}(b) - \boldsymbol{f}(a)\big\| \leqslant \left(\sup_{c \in]a,b[} \big\|\boldsymbol{f}'(c)\big\| \right) \cdot (b - a).$$

En d'autres termes, un point qui se déplace à vitesse inférieure ou égale à M ne peut pas parcourir un trajet de longueur strictement supérieure à $M(b - a)$ entre les temps a et b.

Démonstration. Considérons les fonctions $g : [a, b] \to \mathbb{R}$ et $h : [a, b] \to \mathbb{R}$ définies par :

$$g(t) := \big\|\boldsymbol{f}(t) - \boldsymbol{f}(a)\big\| \quad \text{et} \quad h(t) := g(t) - \left[g(a) + (t - a)\frac{g(b) - g(a)}{b - a} \right].$$

La fonction g est continue sur $[a, b]$. En effet, si $t \in [a, b]$ et si (t_n) est une suite qui tend vers t, alors :

$$\big|g(t_n) - g(t)\big| = \Big| \big\|\boldsymbol{f}(t_n) - \boldsymbol{f}(a)\big\| - \big\|\boldsymbol{f}(t) - \boldsymbol{f}(a)\big\| \Big| \leqslant \big\|\boldsymbol{f}(t_n) - \boldsymbol{f}(t)\big\| \underset{t_n \to t}{\longrightarrow} 0.$$

La fonction h est donc continue sur $[a,b]$ comme somme de deux fonctions continues. Elle est donc bornée et atteint ses bornes dans $[a,b]$. Comme $h(t)$ est la distance selon la verticale du point d'ordonnée $g(t)$ à la corde du graphe de g pour a et b, on a $h(a) = h(b) = 0$. Comme dans la démonstration du théorème de Rolle, on voit donc qu'il existe un point $c \in]a,b[$ où h atteint un maximum ou un minimum. Plus précisément,

- soit $h(t) = 0$ pour tout $t \in [a,b]$ et dans ce cas h atteint un maximum et un minimum en tout point de l'intervalle $]a,b[$;
- soit il existe un point $t \in]a,b[$ avec $h(t) > 0$; dans ce cas le maximum de h sur $[a,b]$ ne peut être atteint en a ou en b et il est atteint en un point $c \in]a,b[$;
- soit il existe un point $t \in]a,b[$ avec $h(t) < 0$; dans ce cas le minimum de h sur $[a,b]$ ne peut être atteint en a ou en b et il est atteint en un point $c \in]a,b[$.

On peut donc trouver une suite (t_n) qui converge vers c, avec $t_n < c$ si h atteint un maximum en c et $t_n > c$ si h atteint un minimum en c, telle que :

$$0 \leqslant \frac{h(t_n) - h(c)}{t_n - c} = \frac{g(t_n) - g(c)}{t_n - c} - \frac{g(b) - g(a)}{b - a}.$$

On a alors :

$$\begin{aligned}
\frac{\|\boldsymbol{f}(b) - \boldsymbol{f}(a)\|}{b-a} = \frac{g(b) - g(a)}{b-a} &\leqslant \frac{g(t_n) - g(c)}{t_n - c} \\
&= \left| \frac{\|\boldsymbol{f}(t_n) - \boldsymbol{f}(a)\| - \|\boldsymbol{f}(c) - \boldsymbol{f}(a)\|}{t_n - c} \right| \\
&\leqslant \left\| \frac{\boldsymbol{f}(t_n) - \boldsymbol{f}(c)}{t_n - c} \right\| \underset{n \to +\infty}{\longrightarrow} \|\boldsymbol{f}'(c)\| \leqslant \left(\sup_{]a,b[} \|\boldsymbol{f}'\| \right).
\end{aligned}$$

■

2.5 Intégration

Nous allons maintenant étendre la notion d'intégrale d'une fonction. Comme dans les cas précédents, nous pouvons le faire directement, ou bien coordonnée par coordonnée. L'existence de la limite dans la définition qui suit est admise.

Définition 17. Soit $\boldsymbol{f} : [a,b] \to \mathbb{R}^m$ une fonction vectorielle continue sur un intervalle $[a,b] \subset \mathbb{R}$. L'intégrale $\displaystyle\int_a^b \boldsymbol{f}(x)\,\mathrm{d}x$ est définie comme limite des sommes régulières de Riemann :

$$\int_a^b \boldsymbol{f}(x)\,\mathrm{d}x := \lim_{n \to +\infty} \frac{b-a}{n} \sum_{k=0}^{n-1} \boldsymbol{f}(x_{k,n}) \quad \text{avec} \quad x_{k,n} := a + k\frac{b-a}{n}.$$

L'intégrale d'une fonction vectorielle est un vecteur. La $i^{ème}$ coordonnée de ce vecteur s'obtient en intégrant la $i^{ème}$ coordonnée de $\boldsymbol{f}$.

Proposition 23. Soit $\boldsymbol{f} := (f_1, \ldots, f_m) : [a,b] \to \mathbb{R}^m$ une fonction vectorielle continue définie sur un segment $[a,b] \subset \mathbb{R}$. Alors :

$$\int_a^b \boldsymbol{f}(x)\,\mathrm{d}x = \begin{pmatrix} \int_a^b f_1(x)\,\mathrm{d}x \\ \vdots \\ \int_a^b f_m(x)\,\mathrm{d}x \end{pmatrix}.$$

Démonstration. La $i^{ème}$ coordonnée de $\displaystyle\int_a^b \boldsymbol{f}(x)\,\mathrm{d}x$ est la limite de sommes de Riemann :

$$\frac{b-a}{n}\sum_{k=0}^{n-1} f_i(x_{k,n}) \qquad \text{avec} \qquad x_{k,n} := a + k\frac{b-a}{n}\cdot$$

Exercice 9.

Déterminer $\displaystyle\int_0^1 \boldsymbol{f}(x)\,\mathrm{d}x$ pour $\boldsymbol{f}(x) := (x^2, x^3)$.

Solution. On a :

$$\int_0^1 \boldsymbol{f}(x)\,\mathrm{d}x = \begin{pmatrix} \int_0^1 x^2\,\mathrm{d}x \\ \int_0^1 x^3\,\mathrm{d}x \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/4 \end{pmatrix}.$$

Les propriétés concernant l'intégrale d'une fonction réelle s'étendent au cas des fonctions vectorielles. Nous ne démontrerons pas celles qui s'obtiennent directement en travaillant coordonnée par coordonnée.

Proposition 24 (Linéarité). Soient $\boldsymbol{f} : [a,b] \to \mathbb{R}^m$ et $\boldsymbol{g} : [a,b] \to \mathbb{R}^m$ deux fonctions vectorielles continues et $\lambda \in \mathbb{R}$ un réel. Alors :

$$\int_a^b (\boldsymbol{f} + \lambda \boldsymbol{g})(x)\,\mathrm{d}x = \int_a^b \boldsymbol{f}(x)\,\mathrm{d}x + \lambda \int_a^b \boldsymbol{g}(x)\,\mathrm{d}x.$$

Proposition 25 (Relation de Chasles). Soient $\boldsymbol{f} : [a,b] \to \mathbb{R}^m$ une fonction vectorielle continue et $c \in \,]a,b[$. Alors :

$$\int_a^b \boldsymbol{f}(x)\,\mathrm{d}x = \int_a^c \boldsymbol{f}(x)\,\mathrm{d}x + \int_c^b \boldsymbol{f}(x)\,\mathrm{d}x.$$

Il est alors naturel d'introduire la définition suivante afin que la relation de Chasles reste valide même si c n'est pas compris entre a et b.

Définition 18. Soit $\boldsymbol{f} : [a,b] \to \mathbb{R}^m$ une fonction vectorielle continue sur un segment $[a,b] \subset \mathbb{R}$. On définit :

$$\int_b^a \boldsymbol{f}(x)\,\mathrm{d}x := -\int_a^b \boldsymbol{f}(x)\,\mathrm{d}x.$$

Il est possible de majorer la valeur absolue de l'intégrale d'une fonction réelle par l'intégrale de sa valeur absolue. Nous allons étendre ce résultat au cas des fonctions vectorielles. La démonstration ne se fait pas en travaillant coordonnée par coordonnée. C'est un résultat excessivement important.

Théorème 26. Si $\boldsymbol{f} : [a, b] \to \mathbb{R}^m$ est une fonction vectorielle continue, alors :

$$\left\| \int_a^b \boldsymbol{f}(x)\,\mathrm{d}x \right\| \leqslant \int_a^b \|\boldsymbol{f}(x)\|\,\mathrm{d}x.$$

Démonstration. Remarquons d'abord que pour que l'intégrale de droite ait un sens, il faut vérifier que $\|\boldsymbol{f}\|$ est une fonction continue. Pour cela, il suffit d'observer que si (x_n) est une suite qui converge vers a, alors, d'après le corollaire 5 de la page 782 :

$$\Big| \|\boldsymbol{f}(x_n)\| - \|\boldsymbol{f}(a)\| \Big| \leqslant \|\boldsymbol{f}(x_n) - \boldsymbol{f}(a)\| \underset{n\to+\infty}{\longrightarrow} 0.$$

Revenons maintenant à la définition de l'intégrale :

$$\int_a^b \boldsymbol{f}(x)\,\mathrm{d}x = \lim_{n\to+\infty} \frac{b-a}{n} \sum_{k=0}^{n-1} \boldsymbol{f}(x_{k,n}) \quad \text{avec} \quad x_{k,n} := a + k\frac{b-a}{n}\cdot$$

L'inégalité triangulaire établie dans le théorème 4 de la page 781 implique que :

$$\left\| \sum_{k=0}^{n-1} \boldsymbol{f}(x_{k,n}) \right\| \leqslant \sum_{k=0}^{n-1} \|\boldsymbol{f}(x_{k,n})\|.$$

On a donc :

$$\begin{aligned} \left\| \int_a^b \boldsymbol{f}(x)\,\mathrm{d}x \right\| &= \lim_{n\to+\infty} \left\| \frac{b-a}{n} \sum_{k=0}^{n-1} \boldsymbol{f}(x_{k,n}) \right\| \\ &\leqslant \lim_{n\to+\infty} \frac{b-a}{n} \sum_{k=0}^{n-1} \|\boldsymbol{f}(x_{k,n})\| = \int_a^b \|\boldsymbol{f}(x)\|\,\mathrm{d}x. \end{aligned}$$

■

3 Équations différentielles linéaires

Une équation différentielle est une relation entre une fonction inconnue et ses dérivées. Une équation différentielle d'ordre m relie la fonction et ses m premières dérivées. Dans le cas d'une fonction à valeurs réelles, nous dirons qu'il s'agit d'une équation différentielle scalaire. Dans le cas d'une fonction vectorielle, nous dirons qu'il s'agit d'une équation différentielle vectorielle.

Nous avons vu au début du module qu'il existe un lien étroit entre les suites réelles définies par une récurrence portant sur plusieurs termes consécutifs, de la forme $u_{n+2} = au_{n+1} + bu_n$ avec $a, b \in \mathbb{R}$ par exemple, et les suites vectorielles définie par une récurrence de la forme $\boldsymbol{u}_{n+1} := A \cdot \boldsymbol{u}_n$ avec $A \in M_m(\mathbb{R})$.

Nous verrons qu'il existe un lien similaire entre les équations différentielles scalaires d'ordre m et les équations différentielles vectorielle d'ordre 1.

Dans ce module, nous n'étudierons qu'un cas très particulier d'équation différentielle : les équations différentielles linéaires.

3.1 Équations différentielles linéaires scalaires d'ordre 1

Définition 19. Une équation différentielle scalaire, linéaire, du premier ordre, est une équation de la forme :

$$y' = ay + b$$

où $a : I \to \mathbb{R}$ et $b : I \to \mathbb{R}$ sont des fonctions continues sur un intervalle $I \subset \mathbb{R}$. Une fonction $f : J \to \mathbb{R}$ est une solution de l'équation différentielle si f est dérivable sur un intervalle $J \subset I$ et si pour tout $t \in J$, on a :

$$f'(t) = a(t)\, f(t) + b(t).$$

Avec un abus de notation, nous noterons les équations différentielles sous la forme $y' = a(t)y + b(t)$ pour insister sur le fait que les fonctions $a : I \to \mathbb{R}$ et $b : I \to \mathbb{R}$ ne sont pas nécessairement constantes.

Exemples.

1. Si $\alpha \in \mathbb{R}^*$ et $\beta \in \mathbb{R}$, l'équation différentielle $y' = \alpha y + \beta$ est une équation différentielle linéaire d'ordre 1 *à coefficients constants*. La fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(t) := \mathrm{e}^{\alpha t} - \beta/\alpha$ est solution de cette équation différentielle :
$$\forall t \in \mathbb{R}, \quad f'(t) = \alpha \mathrm{e}^{\alpha t} = \alpha f(t) + \beta.$$
2. L'équation différentielle $y' = 2ty$ est une équation différentielle linéaire d'ordre 1. La fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(t) := \mathrm{e}^{t^2}$ est solution de cette équation différentielle :
$$\forall t \in \mathbb{R}, \quad f'(t) = 2t\mathrm{e}^{t^2} = 2tf(t).$$

Proposition 27.
Si $f : J \to \mathbb{R}$ est solution de l'équation différentielle $y' = a(t)y + b(t)$, alors f est de classe $\mathcal{C}^1$ sur J.

Démonstration. Si $f : J \to \mathbb{R}$ est solution de l'équation différentielle, alors f est par définition dérivable sur J. En particulier elle est continue sur J. Sa dérivée f' est égale à $af + b$ qui est une fonction continue sur J comme somme et produit de fonctions continues sur J. La fonction f est donc de classe $\mathcal{C}^1$ sur J. ■

Exercice 10.
Montrer par récurrence que si a et b sont de classe $\mathcal{C}^k$ sur l'intervalle I, alors toute solution $f : J \subset I \to \mathbb{R}^m$ de l'équation $y' = a(t)y + b(t)$ est de classe $\mathcal{C}^{k+1}$ sur J.
Solution. Supposons a et b de classe $\mathcal{C}^k$. Si f est de classe $\mathcal{C}^\ell$, $\ell \leqslant k$, alors $f' = af + b$ est de classe $\mathcal{C}^\ell$, donc f est de classe $\mathcal{C}^{\ell+1}$. Une récurrence élémentaire donne le résultat.

Notons que si $f : J \subset I \to \mathbb{R}$ est solution d'une équation différentielle, alors la restriction de f à n'importe quel intervalle contenu dans J est également solution de l'équation différentielle.

Définition 20. Une solution $f : J \to \mathbb{R}$ d'une équation différentielle est une *solution maximale* si elle n'est pas la restriction d'une solution $f_1 : J_1 \to \mathbb{R}$ avec $J \subset J_1$ et $J \neq J_1$.

Observons maintenant que si $f_1 : J \to \mathbb{R}$ et $f_2 : J \to \mathbb{R}$ sont deux solutions de l'équation différentielle $y' = a(t)y + b(t)$, alors $f_1 - f_2$ est solution de l'équation différentielle $y' = a(t)y$.

Définition 21. L'équation *homogène* associée à $y' = a(t)y + b(t)$ est l'équation différentielle $y' = a(t)y$.

Nous allons, dans un premier temps, étudier les solutions de l'équation homogène. On dit également que c'est une équation sans second membre.

Proposition 28.
Pour tout $t_0 \in \mathbb{R}$ et tout $y_0 \in \mathbb{R}$, la fonction $f : I \to \mathbb{R}$ définie par :

$$f(t) := \mathrm{e}^{A(t)} \cdot y_0 \quad \text{avec} \quad A : \begin{array}{ccc} I & \longrightarrow & \mathbb{R} \\ t & \longmapsto & \displaystyle\int_{t_0}^{t} a(u)\,\mathrm{d}u \end{array}$$

est solution de l'équation différentielle $y' = a(t)y$.

Démonstration. La primitive $A : I \to \mathbb{R}$ est dérivable sur I de dérivée $A' = a$. Par conséquent $f : t \mapsto \mathrm{e}^{A(t)} \cdot y_0$ est dérivable comme composée de fonctions dérivables et :

$$f'(t) = A'(t)\mathrm{e}^{A(t)} \cdot y_0 = a(t) \cdot f(t).$$

■

Théorème 29. Considérons l'équation différentielle homogène $y' = a(t)y$ avec $a : I \to \mathbb{R}$ continue. Toute solution maximale est de la forme :

$$f : \begin{array}{ccc} I & \longrightarrow & \mathbb{R} \\ t & \longmapsto & \mathrm{e}^{A(t)} \cdot f(t_0) \end{array} \quad \text{avec} \quad t_0 \in I \quad \text{et} \quad A : \begin{array}{ccc} I & \longrightarrow & \mathbb{R} \\ t & \longmapsto & \displaystyle\int_{t_0}^{t} a(u)\,\mathrm{d}u. \end{array}$$

Démonstration. Soit $f : J \to \mathbb{R}$ une solution de l'équation différentielle $y' = a(t)y$. Définissons $g : J \to \mathbb{R}$ par :

$$g(t) := \mathrm{e}^{-A(t)} \cdot f(t).$$

La fonction g est alors dérivable sur J et pour tout $t \in J$:

$$\begin{aligned} g'(t) &= -A'(t)\mathrm{e}^{-A(t)} \cdot f(t) + \mathrm{e}^{-A(t)} \cdot f'(t) \\ &= -a(t)\mathrm{e}^{-A(t)} \cdot f(t) + \mathrm{e}^{-A(t)} \cdot a(t)f(t) = 0. \end{aligned}$$

On voit donc que pour tout $t \in J$, $g'(t) = 0$. On en déduit que la fonction g est constante sur l'intervalle J et donc qu'il existe $y_0 \in \mathbb{R}$ tel que pour tout $t \in J$:

$$f(t) = \mathrm{e}^{A(t)} \cdot y_0.$$

En remplaçant t par t_0, on obtient $y_0 = f(t_0)$. ■

Remarque. Nous venons de montrer que l'ensemble des solutions de l'équation différentielle $y' = a(t)y$ est une droite vectorielle. Étant donné $t_0 \in \mathbb{R}$, l'application qui a une solution f associe le réel $f(t_0)$ est un isomorphisme d'espaces vectoriels entre l'ensemble des solutions et $\mathbb{R}$.

Revenons maintenant sur les solutions de l'équation avec second membre. Pour trouver la forme des solutions de l'équation avec second membre, nous allons utiliser la méthode dite de *variation de la constante*. Nous avons vu que les solutions de l'équation homogène sont de la forme $t \mapsto \mathrm{e}^{A(t)} \cdot y_0$ avec $y_0 \in \mathbb{R}$. La méthode de variation de la constante consiste à rechercher les solutions $f : J \to \mathbb{R}$ de l'équation avec second membre sous la forme $f(t) = \mathrm{e}^{A(t)} \cdot g(t)$ avec $g : J \to \mathbb{R}$ une fonction à valeurs réelles.

Exercice 11.

On suppose que $f : J \to \mathbb{R}$ est solution de $y' = a(t)y + b(t)$ avec $a : I \to \mathbb{R}$ et $b : I \to \mathbb{R}$ continues. Étant donné $t_0 \in J$, on définit $A : I \to \mathbb{R}$ par $A(t) := \displaystyle\int_{t_0}^{t} a(u)\,\mathrm{d}u$.

1. On définit $g : J \to \mathbb{R}$ par $g(t) := \mathrm{e}^{-A(t)} \cdot f(t)$.
   Montrer que pour tout $t \in J$, $g'(t) = \mathrm{e}^{-A(t)} \cdot b(t)$.
2. En déduire que si $f : J \to \mathbb{R}$ est une solution, alors pour tout $t \in J$:
$$f(t) = \mathrm{e}^{A(t)} \left(f(t_0) + \int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u)\,\mathrm{d}u \right).$$
3. En déduire que les solutions maximales sont définies sur I et sont exactement les fonctions de la forme :
$$\begin{array}{rccl} f : & I & \longrightarrow & \mathbb{R} \\ & t & \longmapsto & \mathrm{e}^{A(t)} \left(y_0 + \displaystyle\int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u)\,\mathrm{d}u \right) \end{array}$$
   avec $t_0 \in I$ et $y_0 \in \mathbb{R}$ un réel quelconque.

Solution.

1. La fonction g est dérivable comme composée et produit de deux fonctions dérivables. Sa dérivée vérifie pour tout $t \in J$:
$$g'(t) = -A'(t)\mathrm{e}^{-A(t)} \cdot f(t) + \mathrm{e}^{-A(t)} \cdot f'(t) = \mathrm{e}^{-A(t)} \cdot b(t)$$
   car $f'(t) = A'(t) \cdot f(t) + b(t)$.
2. Supposons que $f : J \to \mathbb{R}$ est une solution et que $g(t) := \mathrm{e}^{-A(t)} \cdot f(t)$. D'après la question précédente, pour tout $t \in J$:
$$g(t) = g(t_0) + \int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u)\,\mathrm{d}u = f(t_0) + \int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u)\,\mathrm{d}u.$$
   Par conséquent :
$$f(t) = \mathrm{e}^{A(t)} \left(f(t_0) + \int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u)\,\mathrm{d}u \right).$$

3. Soit $f : J \to \mathbb{R}$ une solution maximale et soit $t_0 \in J$. On définit $f_1 : I \to \mathbb{R}$ par la formule :

$$f_1(t) := \mathrm{e}^{A(t)} \left(f(t_0) + \int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u) \, \mathrm{d}u \right).$$

On vérifie que f_1 est solution de l'équation différentielle :

$$\begin{aligned} f_1'(t) &= A'(t)\mathrm{e}^{A(t)} \left(f(t_0) + \int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u) \, \mathrm{d}u \right) + \mathrm{e}^{A(t)} \cdot \mathrm{e}^{-A(t)} \cdot b(t) \\ &= a(t) \cdot f(t) + b(t). \end{aligned}$$

Comme $f = f_1$ sur J d'après le point précédent, et comme f est une solution maximale, on a $J = I$.

On vient également de montrer que pour tout $t_0 \in I$ et tout $y_0 \in \mathbb{R}$, la fonction :

$$\begin{array}{rccl} f : & I & \longrightarrow & \mathbb{R} \\ & t & \longmapsto & \mathrm{e}^{A(t)} \left(y_0 + \displaystyle\int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u) \, \mathrm{d}u \right) \end{array} \qquad \text{avec} \quad A(t) = \int_{t_0}^{t} a(u) \, \mathrm{d}u$$

est solution de l'équation différentielle.

L'exercice précédent a pour conséquence immédiate le théorème suivant.

Théorème 30. Considérons l'équation différentielle :

$$y' = a(t)y + b(t) \qquad \text{avec} \qquad a : I \to \mathbb{R} \quad \text{et} \quad b : I \to \mathbb{R} \quad \text{continues.}$$

Soit $t_0 \in I$.

- Toute solution maximale de l'équation différentielle est de la forme :

$$\begin{array}{rccl} f : & I & \longrightarrow & \mathbb{R} \\ & t & \longmapsto & \mathrm{e}^{A(t)} \left(f(t_0) + \displaystyle\int_{t_0}^{t} \mathrm{e}^{-A(u)} \cdot b(u) \, \mathrm{d}u \right) \end{array}$$

avec $t_0 \in I$, $A(t) := \displaystyle\int_{t_0}^{t} a(u) \, \mathrm{d}u$ et $f(t_0) \in \mathbb{R}$ un réel quelconque.

- Si $f_{\text{particulière}} : I \to \mathbb{R}$ est une solution particulière de l'équation différentielle, alors les solutions maximales de l'équation avec second membre sont exactement les fonctions $f : I \to \mathbb{R}$ de la forme :

$$f = f_{\text{homogène}} + f_{\text{particulière}},$$

où $f_{\text{homogène}} : I \to \mathbb{R}$ est solution de l'équation homogène.

- L'ensemble des solutions maximales est une droite affine.

3.2 Équations vectorielles d'ordre 1

Définition 22. Une équation différentielle linéaire du premier ordre est une équation de la forme :

$$\boldsymbol{y}' = \boldsymbol{A}(t) \cdot \boldsymbol{y} + \boldsymbol{b}(t)$$

où $\boldsymbol{A} : I \to M_m(\mathbb{R})$ et $\boldsymbol{b} : I \to \mathbb{R}^m$ sont des applications continues sur un intervalle $I \subset \mathbb{R}$. Une fonction vectorielle $\boldsymbol{f} : J \to \mathbb{R}^m$ est une solution de l'équation différentielle si $\boldsymbol{f}$ est dérivable sur un intervalle $J \subset I$ et si, pour tout $t \in J$, on a :

$$\boldsymbol{f}'(t) = \boldsymbol{A}(t) \cdot \boldsymbol{f}(t) + \boldsymbol{b}(t).$$

Proposition 31. Si $\boldsymbol{f} : J \to \mathbb{R}^m$ est solution d'une équation différentielle linéaire d'ordre 1, alors $\boldsymbol{f}$ est de classe $\mathcal{C}^1$ sur J.

Démonstration. La preuve est similaire à celle exposée dans le cas d'une équation scalaire (voir proposition 27 de la page 798). ■

La notion de solution maximale est la même dans le cas vectoriel que dans le cas scalaire.

Définition 23. Une solution $\boldsymbol{f} : J \to \mathbb{R}^m$ d'une équation différentielle est une solution *maximale* si elle n'est pas la restriction d'une solution $\boldsymbol{f}_1 : J_1 \to \mathbb{R}^m$ avec $J \subset J_1$ et $J \neq J_1$.

Comme dans le cas d'une équation scalaire, l'étude d'une équation différentielle vectorielle linéaire passe par l'étude de l'équation homogène associée.

Définition 24. L'équation *homogène* associée à l'équation différentielle linéaire $\boldsymbol{y}' = \boldsymbol{A}(t) \cdot \boldsymbol{y} + \boldsymbol{b}(t)$ est l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A}(t) \cdot \boldsymbol{y}$.

Théorème 32. Soit $J \subset I$ un intervalle. L'ensemble des solutions $\boldsymbol{f} : J \to \mathbb{R}^m$ de l'équation homogène est un sous-espace vectoriel de l'espace vectoriel des fonctions de classe $\mathcal{C}^1$ sur J.

Démonstration. D'après la proposition 31, toute solution $\boldsymbol{f} : J \to \mathbb{R}^m$ est de classe $\mathcal{C}^1$ sur J.

La fonction constante égale au vecteur nul de $\mathbb{R}^m$ est solution de l'équation homogène. L'ensemble des solutions définies sur J n'est donc pas vide.

Si $\boldsymbol{f}_1 : J \to \mathbb{R}^m$ et $\boldsymbol{f}_2 : J \to \mathbb{R}^m$ sont deux solutions de l'équation homogène et si $\lambda \in \mathbb{R}$, alors $\lambda \boldsymbol{f}_1 + \boldsymbol{f}_2 : J \to \mathbb{R}^m$ est une solution de l'équation homogène. En effet, pour tout $t \in J$:

$$\begin{aligned}(\lambda \boldsymbol{f}_1 + \boldsymbol{f}_2)'(t) &= \lambda \boldsymbol{f}_1'(t) + \boldsymbol{f}_2'(t) \\ &= \lambda \boldsymbol{A}(t) \cdot \boldsymbol{f}_1(t) + \boldsymbol{A}(t)\boldsymbol{f}_2(t) \\ &= \boldsymbol{A}(t) \cdot (\lambda \boldsymbol{f}_1 + \boldsymbol{f}_2)(t).\end{aligned}$$

■

Pour le moment, nous ne pouvons malheureusement pas aller beaucoup plus loin dans le cas où $\boldsymbol{A} : I \to \mathbb{R}$ est une application continue quelconque. Nous traiterons le cas général dans le volume de L2 à l'aide du théorème de Cauchy-Lipschitz.

> ***Remarque***. Ce qui nous empêche de traiter le cas général, c'est que pour $t \neq t'$, les matrices $\boldsymbol{A}(t)$ et $\boldsymbol{A}(t')$ ne commutent généralement pas.

Dans ce module, nous nous contenterons de traiter le cas où $\boldsymbol{A}$ est une application constante. Dans ce cas, les matrices $\boldsymbol{A}(t)$ et $\boldsymbol{A}(t')$ commutent. Nous dirons alors que l'équation $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ est à coefficients constants. À partir de maintenant, nous supposerons donc que $\boldsymbol{A}$ est une matrice de $M_m(\mathbb{R})$.

Définition 25. L'exponentielle d'une matrice carrée $\boldsymbol{A} \in M_m(\mathbb{R})$ est la matrice définie par :

$$\mathrm{e}^{\boldsymbol{A}} := \sum_{n=0}^{+\infty} \frac{\boldsymbol{A}^n}{n!} = \mathrm{I} + \boldsymbol{A} + \frac{\boldsymbol{A}^2}{2} + \frac{\boldsymbol{A}^3}{6} + \cdots .$$

Nous verrons dans le volume de L2 que cette série est toujours convergente et que si $\boldsymbol{A}$ et $\boldsymbol{B}$ sont deux matrices qui commutent, alors $\mathrm{e}^{\boldsymbol{A}+\boldsymbol{B}} = \mathrm{e}^{\boldsymbol{A}} \cdot \mathrm{e}^{\boldsymbol{B}} = \mathrm{e}^{\boldsymbol{B}} \cdot \mathrm{e}^{\boldsymbol{A}}$.
Pour tout $t \in \mathbb{R}$, la matrice $\boldsymbol{A}$ commute avec la matrice $t\boldsymbol{A}$. Elle commute donc avec $\mathrm{e}^{t\boldsymbol{A}}$.
De plus, pour tout $t_1 \in \mathbb{R}$ et $t_2 \in \mathbb{R}$, les matrices $t_1\boldsymbol{A}$ et $t_2\boldsymbol{A}$ commutent.
Par conséquent, $\mathrm{e}^{t_1\boldsymbol{A}} \cdot \mathrm{e}^{t_2\boldsymbol{A}} = \mathrm{e}^{(t_1+t_2)\boldsymbol{A}}$.

Proposition 33. Pour tout $\boldsymbol{y}_0 \in \mathbb{R}^m$ et tout $\boldsymbol{A} \in M_m(\mathbb{R})$, la fonction $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^m$ définie par :

$$\boldsymbol{f}(t) := \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{y}_0$$

est solution de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$.

> ***Remarque***. Puisque $\boldsymbol{A} \in M_m(\mathbb{R})$ est une matrice et $\boldsymbol{y}_0 \in \mathbb{R}^m$ un vecteur, le produit $\boldsymbol{y}_0 \cdot \mathrm{e}^{t\boldsymbol{A}}$ n'a pas de sens. C'est le produit $\mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{y}_0$ qui a un sens.

Démonstration. Montrons que pour tout $t_0 \in \mathbb{R}$, $\boldsymbol{f}$ est dérivable en t_0. On a :

$$\frac{\boldsymbol{f}(t) - \boldsymbol{f}(t_0)}{t - t_0} = \frac{\mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{y}_0 - \mathrm{e}^{t_0\boldsymbol{A}} \cdot \boldsymbol{y}_0}{t - t_0} = \frac{\mathrm{e}^{(t-t_0)\boldsymbol{A}} - \boldsymbol{I}_m}{t - t_0} \cdot \mathrm{e}^{t_0\boldsymbol{A}} \cdot \boldsymbol{y}_0.$$

Or :

$$\frac{\mathrm{e}^{(t-t_0)\boldsymbol{A}} - \boldsymbol{I}_m}{t - t_0} = \sum_{n=1}^{+\infty} \frac{(t-t_0)^{n-1}}{n!} \boldsymbol{A}^n = \boldsymbol{A} + \sum_{n=2}^{+\infty} \frac{(t-t_0)^{n-1}}{n!} \boldsymbol{A}^n$$

et dès que $|t - t_0| < 1$, on a :

$$\left\| \frac{\mathrm{e}^{(t-t_0)\boldsymbol{A}} - \boldsymbol{I}_m}{t - t_0} - \boldsymbol{A} \right\| \leqslant |t - t_0| \sum_{n=2}^{+\infty} \frac{\|\boldsymbol{A}\|^n}{n!} \leqslant |t - t_0| \mathrm{e}^{\|\boldsymbol{A}\|} \underset{t \to t_0}{\longrightarrow} 0.$$

Par conséquent, $\boldsymbol{f}$ est dérivable en t_0 et :

$$\boldsymbol{f}'(t_0) = \lim_{t \to t_0} \frac{\boldsymbol{f}(t) - \boldsymbol{f}(t_0)}{t - t_0} = \boldsymbol{A} \cdot \mathrm{e}^{t_0 \boldsymbol{A}} \cdot \boldsymbol{y}_0 = \boldsymbol{A} \cdot \boldsymbol{f}(t_0).$$

■

Remarque. On a utilisé de manière cruciale le fait que pour tout $t_1 \in \mathbb{R}$ et $t_2 \in \mathbb{R}$, les matrices $t_1\boldsymbol{A}$ et $t_2\boldsymbol{A}$ commutent. Si $\boldsymbol{A} : \mathbb{R} \to M_m(\mathbb{R})$ est une fonction dérivable quelconque, la dérivée de $\mathrm{e}^{\boldsymbol{A}(t)}$ n'est en général pas égale à $\boldsymbol{A}'(t) \cdot \mathrm{e}^{\boldsymbol{A}(t)}$.

Théorème 34. Considérons l'équation différentielle homogène $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ avec $\boldsymbol{A} \in M_m(\mathbb{R})$. Toute solution maximale est de la forme :

$$\begin{array}{rccc} \boldsymbol{f} : & \mathbb{R} & \longrightarrow & \mathbb{R}^m \\ & t & \longmapsto & \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{f}(0). \end{array}$$

Démonstration. Soit $\boldsymbol{f} : J \to \mathbb{R}^m$ une solution de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$. Définissons $\boldsymbol{g} : J \to \mathbb{R}^m$ par :

$$\boldsymbol{g}(t) := \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}(t).$$

La fonction $\boldsymbol{g}$ est alors dérivable sur J et :

$$\begin{aligned} \boldsymbol{g}'(t) &= -\boldsymbol{A}\mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}(t) + \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}'(t) \\ &= -\boldsymbol{A}\mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}(t) + \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{A}\boldsymbol{f}(t). \end{aligned}$$

Or, $\boldsymbol{A}$ et $\mathrm{e}^{t\boldsymbol{A}}$ commutent. On voit donc que pour tout $t \in J$, $\boldsymbol{g}'(t) = 0$. On en déduit que la fonction $\boldsymbol{g}$ est constante sur l'intervalle J et donc qu'il existe $\boldsymbol{y}_0 \in \mathbb{R}^m$ tel que :

$$\boldsymbol{f}(t) = \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{y}_0.$$

Cela montre que toute solution maximale est de la forme : $\begin{array}{rccc} \boldsymbol{f} : & \mathbb{R} & \longrightarrow & \mathbb{R}^m \\ & t & \longmapsto & \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{f}(0). \end{array}$ ■

Théorème 35. L'ensemble des solutions maximales est un espace vectoriel de dimension m. Pour tout $t_0 \in \mathbb{R}$, l'application qui à une solution $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^m$ associe $\boldsymbol{f}(t_0) \in \mathbb{R}^m$ est un isomorphisme d'espaces vectoriels.

Démonstration. Toute solution $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^m$ est de la forme :

$$t \mapsto \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{f}(0) = \mathrm{e}^{(t-t_0)\boldsymbol{A}} \cdot \boldsymbol{f}(t_0).$$

On voit donc que l'application linéaire qui a une solution maximale $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^m$ associe $\boldsymbol{f}(t_0) \in \mathbb{R}^m$ est une bijection de l'ensemble des solutions maximales de l'équation homogène dans $\mathbb{R}^m$. ■

Revenons maintenant sur les solutions de l'équation avec second membre. Pour trouver la forme des solutions de l'équation avec second membre, nous allons utiliser la méthode de variation de la constante. Autrement dit, nous allons rechercher les solutions de l'équation avec second membre sous la forme $\boldsymbol{f}(t) := \mathrm{e}^{t\boldsymbol{A}}\boldsymbol{g}(t)$ avec $\boldsymbol{g}$ une fonction vectorielle.

Exercice 12.

On suppose que $\boldsymbol{f} : J \to \mathbb{R}^m$ est solution de $\boldsymbol{y}' = \boldsymbol{A}\cdot\boldsymbol{y} + \boldsymbol{b}(t)$ avec $\boldsymbol{b} : I \to \mathbb{R}^m$ continue. On définit $\boldsymbol{g} : J \to \mathbb{R}^m$ par $\boldsymbol{g}(t) := \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}(t)$.

1. Montrer que $\boldsymbol{g} : J \to \mathbb{R}^m$ est dérivable et que pour tout $t \in J$, $\boldsymbol{g}'(t) = \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{b}(t)$.
2. En déduire que si $\boldsymbol{f} : J \to \mathbb{R}^m$ est une solution maximale, alors $J = I$ et pour $t_0 \in I$:

$$\boldsymbol{f}(t) = \mathrm{e}^{(t-t_0)\boldsymbol{A}} \cdot \boldsymbol{f}(t_0) + \mathrm{e}^{t\boldsymbol{A}} \cdot \int_{t_0}^{t} \mathrm{e}^{-u\boldsymbol{A}} \cdot \boldsymbol{b}(u)\,\mathrm{d}u \quad \text{avec} \quad \boldsymbol{f}(t_0) \in \mathbb{R}^m.$$

Solution.

1. La fonction $\boldsymbol{g} : \begin{array}{ccc} J & \longrightarrow & \mathbb{R}^m \\ t & \longmapsto & \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}(t) \end{array}$ est dérivable comme produit de deux fonctions dérivables. Sa dérivée est donnée par :

$$\boldsymbol{g}'(t) = -\boldsymbol{A}\mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}(t) + \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}'(t) = \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{b}(t)$$

car $\boldsymbol{A}$ et $\mathrm{e}^{-t\boldsymbol{A}}$ commutent et $\boldsymbol{f}'(t) = \boldsymbol{A} \cdot \boldsymbol{f}(t) + \boldsymbol{b}(t)$.

2. Supposons que $\boldsymbol{f} : J \to \mathbb{R}^m$ est une solution maximale. On définit $\boldsymbol{g} : J \to \mathbb{R}$ par $\boldsymbol{g}(t) := \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}(t)$. On choisit $t_0 \in J$ et l'on définit $\boldsymbol{g}_1 : I \to \mathbb{R}^m$ par :

$$\boldsymbol{g}_1(t) := \boldsymbol{g}(t_0) + \int_{t_0}^{t} \mathrm{e}^{-u\boldsymbol{A}} \cdot \boldsymbol{b}(u)\,\mathrm{d}u.$$

Enfin, on définit $\boldsymbol{f}_1 : I \to \mathbb{R}^m$ par $\boldsymbol{f}_1(t) := \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{g}_1(t)$. La fonction $\boldsymbol{f}_1$ est dérivable sur I et :

$$\boldsymbol{f}_1{}'(t) = \boldsymbol{A}\mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{g}_1(t) + \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{g}_1'(t) = \boldsymbol{A} \cdot \boldsymbol{f}_1(t) + \mathrm{e}^{t\boldsymbol{A}} \cdot \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{b}(t) = \boldsymbol{A} \cdot \boldsymbol{f}_1(t) + \boldsymbol{b}(t).$$

La fonction $\boldsymbol{f}_1$ est donc solution de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t)$.

La fonction $\boldsymbol{g}_1$ est dérivable sur I, et sur $J \subset I$, on a $\boldsymbol{g}_1' = \boldsymbol{g}'$ d'après la question précédente. De plus, $\boldsymbol{g}_1(t_0) = \boldsymbol{g}(t_0)$. On en déduit que sur J, $\boldsymbol{g}_1$ et $\boldsymbol{g}$ coïncident. Par conséquent, $\boldsymbol{f}$ et la restriction à J de la solution $\boldsymbol{f}_1$. Comme $\boldsymbol{f}$ est une solution maximale, on a nécessairement $J = I$ et :

$$\begin{aligned} \boldsymbol{f}(t) &= \mathrm{e}^{t\boldsymbol{A}} \cdot \left(\mathrm{e}^{-t_0\boldsymbol{A}} \cdot \boldsymbol{f}(t_0) + \int_{t_0}^{t} \mathrm{e}^{-u\boldsymbol{A}} \cdot \boldsymbol{b}(u)\,\mathrm{d}u \right) \\ &= \mathrm{e}^{(t-t_0)\boldsymbol{A}} \cdot \boldsymbol{f}(t_0) + \mathrm{e}^{t\boldsymbol{A}} \cdot \int_{t_0}^{t} \mathrm{e}^{-u\boldsymbol{A}} \cdot \boldsymbol{b}(u)\,\mathrm{d}u. \end{aligned}$$

L'exercice précédent a pour conséquence immédiate le théorème suivant.

Théorème 36. Considérons l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t)$ avec $\boldsymbol{A} \in M_m(\mathbb{R})$ et $\boldsymbol{b} : I \to \mathbb{R}^m$ continue.

- Toute solution maximale de l'équation différentielle est de la forme :

$$\begin{array}{rccl} \boldsymbol{f} : & I & \longrightarrow & \mathbb{R}^m \\ & t & \longmapsto & \mathrm{e}^{(t-t_0)\boldsymbol{A}} \cdot \boldsymbol{f}(t_0) + \mathrm{e}^{t\boldsymbol{A}} \cdot \displaystyle\int_{t_0}^{t} \mathrm{e}^{-u\boldsymbol{A}} \cdot b(u)\,\mathrm{d}u \end{array}$$

  avec $t_0 \in I$ et $\boldsymbol{f}(t_0) \in \mathbb{R}^m$ un vecteur quelconque.
- Si $\boldsymbol{f}_{\text{particulière}} : I \to \mathbb{R}^m$ est une solution particulière de l'équation différentielle, alors les solutions maximales de l'équation avec second membre sont exactement les fonctions $\boldsymbol{f} : I \to \mathbb{R}^m$ de la forme :

$$\boldsymbol{f} = \boldsymbol{f}_{\text{homogène}} + \boldsymbol{f}_{\text{particulière}},$$

  où $\boldsymbol{f}_{\text{homogène}} : I \to \mathbb{R}^m$ est solution maximale de l'équation homogène.
- L'ensemble des solutions maximales est un espace affine de dimension m.

La méthode de variation de la constante peut être présentée de la manière suivante.
Si $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m)$ est une base de solutions de l'équation différentielle homogène $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$, alors pour tout $t \in \mathbb{R}$,

$$\big(\boldsymbol{f}_1(t), \ldots, \boldsymbol{f}_m(t)\big) = \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{P} \quad \text{avec} \quad \boldsymbol{P} := \big(\boldsymbol{f}_1(0), \ldots, \boldsymbol{f}_m(0)\big)$$

matrice inversible. Par conséquent, toute application dérivable $\boldsymbol{f} : I \to \mathbb{R}^m$ peut s'écrire sous la forme :

$$\boldsymbol{f} = \lambda_1 \boldsymbol{f}_1 + \cdots + \lambda_m \boldsymbol{f}_m$$

avec $\lambda_1 : I \to \mathbb{R}, \ldots, \lambda_m : I \to \mathbb{R}$ dérivables sur I. En effet,

$$\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix} = \boldsymbol{P}^{-1} \cdot \mathrm{e}^{-t\boldsymbol{A}} \cdot \boldsymbol{f}(t).$$

Alors :

$$\begin{aligned} \boldsymbol{f}' &= \big(\lambda_1 \boldsymbol{f}_1' + \cdots + \lambda_m \boldsymbol{f}_m'\big) + \big(\lambda_1' \boldsymbol{f}_1 + \cdots + \lambda_m' \boldsymbol{f}_m\big) \\ &= \boldsymbol{A} \cdot \boldsymbol{f} + \lambda_1' \boldsymbol{f}_1 + \cdots + \lambda_m' \boldsymbol{f}_m. \end{aligned}$$

Pour que $\boldsymbol{f}$ soit solution de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t)$, il faut et il suffit que :

$$\lambda_1' \boldsymbol{f}_1 + \cdots + \lambda_m' \boldsymbol{f}_m = \boldsymbol{b}.$$

Exprimons $\boldsymbol{b}$ dans la base $\big(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m\big)$:

$$\boldsymbol{b} = b_1 \boldsymbol{f}_1 + \cdots + b_m \boldsymbol{f}_m$$

avec $b_1 : I \to \mathbb{R}, \ldots, b_m : I \to \mathbb{R}$ continues. On obtient alors :

$$\lambda_1' = b_1 \quad \ldots \quad \lambda_m' = b_m.$$

Étant donné $t_0 \in I$, les solutions de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t)$ sont donc les fonctions de la forme :

$$\boldsymbol{f} : t \mapsto \left(C_1 + \int_{t_0}^{t} b_1(u)\,\mathrm{d}u \right) \boldsymbol{f}_1(t) + \cdots + \left(C_m + \int_{t_0}^{t} b_m(u)\,\mathrm{d}u \right) \boldsymbol{f}_m(t)$$

avec $C_1, \ldots, C_m \in \mathbb{R}$ les coordonnées de $\boldsymbol{f}(t_0)$ dans la base $\big(\boldsymbol{f}_1(t_0), \ldots, \boldsymbol{f}_m(t_0)\big)$.

3.3 Allure des solutions d'une équation homogène en dimension 2

Dans cette partie, nous allons interpréter graphiquement les résultats de la partie précédente dans le cas d'une équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ avec $\boldsymbol{A} \in M_2(\mathbb{R})$. Les solutions maximales sont des fonctions vectorielles $f : \mathbb{R} \to \mathbb{R}^2$. Ce sont donc des courbes paramétrées du plan. Nous allons expliquer comment tracer ces courbes et comment déterminer le sens de parcours quand la variable t croît. Ce qui n'apparaîtra pas, c'est la vitesse de parcours.

Nous venons de voir que les solutions maximales sont les fonctions $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^2$ de la forme $t \mapsto \mathrm{e}^{t\boldsymbol{A}} \cdot \boldsymbol{f}(0)$. Afin d'étudier ces solutions, il peut être avantageux de changer de base, de travailler dans une base de vecteurs propres quand c'est possible, par exemple.

Observons que comme $\boldsymbol{A}$ est une matrice à coefficients réels, son polynôme caractéristique est un polynôme de degré 2 à coefficients réels. Par conséquent, la matrice $\boldsymbol{A}$ possède :

- soit deux valeurs propres réelles et distinctes
- soit une seule valeur propre réelle
- soit deux valeurs propres non réelles et complexes conjuguées.

Nous allons traiter les trois cas séparément.

3.3.1 Cas de deux valeurs propres réelles distinctes

Supposons d'abord que la matrice $\boldsymbol{A}$ ait deux valeurs propres réelles distinctes λ_1 et λ_2. Soient $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$ deux vecteurs propres associés respectivement à λ_1 et λ_2. Alors, $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$ sont linéairement indépendants et forment une base de $\mathbb{R}^2$. Dans ce cas, si $f_1 : \mathbb{R} \to \mathbb{R}$ et $f_2 : \mathbb{R} \to \mathbb{R}$ sont telles que $\boldsymbol{f} = f_1\boldsymbol{v}_1 + f_2\boldsymbol{v}_2 : \mathbb{R} \to \mathbb{R}^2$ est solution de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$, alors :

$$\begin{pmatrix} f_1'(t) \\ f_2'(t) \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \cdot \begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix}.$$

Exercice 13.

Le vérifier.

Solution.

Comme $f_1'\boldsymbol{v}_1 + f_2'\boldsymbol{v}_2 = \boldsymbol{f}' = \boldsymbol{A} \cdot \boldsymbol{f} = \lambda_1 f_1 \boldsymbol{v}_1 + \lambda_2 f_2 \boldsymbol{v}_2$, on a $f_1' = \lambda_1 f_1$ et $f_2' = \lambda_2 f_2$.

Autrement dit, (f_1, f_2) est solution de l'équation différentielle :

$$\boldsymbol{y}' = \boldsymbol{B} \cdot \boldsymbol{y} \quad \text{avec} \quad \boldsymbol{B} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

Donc :

$$\begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix} = \mathrm{e}^{t\boldsymbol{B}} \cdot \begin{pmatrix} f_1(0) \\ f_2(0) \end{pmatrix}.$$

Or :

$$\mathrm{e}^{t\boldsymbol{B}} = \begin{pmatrix} \mathrm{e}^{t\lambda_1} & 0 \\ 0 & \mathrm{e}^{t\lambda_2} \end{pmatrix}.$$

On en déduit que $\boldsymbol{f}(t) = a_1\mathrm{e}^{t\lambda_1}\boldsymbol{v}_1 + a_2\mathrm{e}^{t\lambda_2}\boldsymbol{v}_2$ où a_1 et a_2 sont les coordonnées de $\boldsymbol{f}(0)$ dans la base $(\boldsymbol{v}_1, \boldsymbol{v}_2)$.
L'allure des trajectoires et le sens de parcours dépendent du signe de λ_1 et de λ_2. Changer simultanément le signe de λ_1 et λ_2 ne change pas les trajectoires mais change le sens de parcours. Nous ne traiterons donc pas tous les cas, laissant au lecteur le soin de compléter les cas manquants en changeant le signe de λ_1 et λ_2, ou en échangeant les rôles de λ_1 et λ_2.
Des allures possibles sont décrites sur la figure IV.6.1 où l'on a choisi $\boldsymbol{v}_1 = (2, 1)$ et $\boldsymbol{v}_2 = (1, 2)$.

$\lambda_2 > \lambda_1 > 0$ $\qquad$ $\lambda_1 > 0 > \lambda_2$

FIGURE IV.6.1 – Deux valeurs propres distinctes

3.3.2 Cas d'une valeur propre double

Supposons maintenant que la matrice $\boldsymbol{A}$ ait une unique valeur propre réelle. Si la matrice $\boldsymbol{A}$ est diagonale, alors c'est la matrice d'une homothétie de rapport λ :

$$\boldsymbol{A} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}, \quad \mathrm{e}^{t\boldsymbol{A}} = \begin{pmatrix} \mathrm{e}^{t\lambda} & 0 \\ 0 & \mathrm{e}^{t\lambda} \end{pmatrix}$$

et les solutions maximales de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ sont les fonctions :

$$\begin{array}{rccl} \boldsymbol{f} : & \mathbb{R} & \longrightarrow & \mathbb{R}^2 \\ & t & \longmapsto & \mathrm{e}^{t\lambda}\boldsymbol{f}(0). \end{array}$$

Les trajectoires sont les demi-droites issues de l'origine si $\lambda > 0$ et aboutissant à l'origine si $\lambda < 0$ (et des constantes si $\lambda = 0$).

Si la matrice $\boldsymbol{A}$ n'est pas diagonale, alors on choisit une base $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ de $\mathbb{R}^2$ telle que $\boldsymbol{v}_1$ est un vecteur propre de $\boldsymbol{A}$ et telle que $\boldsymbol{A} \cdot \boldsymbol{v}_2 = \boldsymbol{v}_1 + \lambda \boldsymbol{v}_2$. On vérifie alors facilement que si $f_1 : \mathbb{R} \to \mathbb{R}$ et $f_2 : \mathbb{R} \to \mathbb{R}$ sont telles que $\boldsymbol{f} := f_1\boldsymbol{v}_1 + f_2\boldsymbol{v}_2 : \mathbb{R} \to \mathbb{R}^2$ est solution de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$, alors :

$$\begin{pmatrix} f_1'(t) \\ f_2'(t) \end{pmatrix} = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \cdot \begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix}.$$

FIGURE IV.6.2 – Une valeur propre double

Exercice 14.

Le vérifier.

Solution. Comme $f_1'\boldsymbol{v}_1 + f_2'\boldsymbol{v}_2 = \boldsymbol{f}' = A\boldsymbol{f} = \lambda f_1 \boldsymbol{v}_1 + f_2 \cdot (\boldsymbol{v}_1 + \lambda \boldsymbol{v}_2)$, on a $f_1' = \lambda f_1 + f_2$ et $f_2' = \lambda f_2$.

Autrement dit, (f_1, f_2) est solution de l'équation différentielle :

$$\boldsymbol{y}' = \boldsymbol{B} \cdot \boldsymbol{y} \quad \text{avec} \quad \boldsymbol{B} = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}.$$

Donc :

$$\begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix} = \mathrm{e}^{t\boldsymbol{B}} \cdot \begin{pmatrix} f_1(0) \\ f_2(0) \end{pmatrix}.$$

Posons $\boldsymbol{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Alors :

$$\boldsymbol{N}^2 = 0, \quad \boldsymbol{B} = \lambda \boldsymbol{I}_2 + \boldsymbol{N} \quad \text{et} \quad \text{pour tout } k \geqslant 1, \quad \boldsymbol{B}^k = \lambda^k \boldsymbol{I}_2 + k\lambda^{k-1}\boldsymbol{N}.$$

On en déduit que :

$$\begin{aligned} \mathrm{e}^{t\boldsymbol{B}} &= \sum_{k=0}^{+\infty} \frac{(t\lambda)^k}{k!} \boldsymbol{I}_2 + \sum_{k=1}^{+\infty} \frac{t^k \lambda^{k-1}}{k!} k\boldsymbol{N}) \\ &= \sum_{k=0}^{+\infty} \frac{(t\lambda)^k}{k!} \boldsymbol{I}_2 + t \sum_{k=1}^{+\infty} \frac{(t\lambda)^{k-1}}{(k-1)!} \boldsymbol{N} = \mathrm{e}^{t\lambda} \boldsymbol{I}_2 + t\mathrm{e}^{t\lambda} \boldsymbol{N}. \end{aligned}$$

Autrement dit : $\mathrm{e}^{t\boldsymbol{B}} = \mathrm{e}^{t\lambda}\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ et $\begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix} = \mathrm{e}^{t\lambda}\begin{pmatrix} f_1(0) + tf_2(0) \\ f_2(0) \end{pmatrix}$.

3.3.3 Cas de deux valeurs propres complexes conjuguées

Supposons maintenant que la matrice $\boldsymbol{A}$ possède deux valeurs propres complexes conjuguées $\lambda := \rho + i\omega$ et $\bar{\lambda} := \rho - i\omega$ avec $\rho \in \mathbb{R}$ et $\omega \in \mathbb{R} \setminus \{0\}$. Soit $\boldsymbol{v} \in \mathbb{C}^2$ un vecteur propre associé à λ. Alors, $\boldsymbol{v} = \boldsymbol{v}_1 + i\boldsymbol{v}_2$ avec $\boldsymbol{v}_1 \in \mathbb{R}^2$ et $\boldsymbol{v}_2 \in \mathbb{R}^2$. Notons que $\boldsymbol{v}_1$ et $\boldsymbol{v}_2$ ne sont pas colinéaires, sinon $\boldsymbol{v}_1$ et $\boldsymbol{v}$ seraient colinéaires et $\boldsymbol{v}_1$ serait un vecteur propre réel, nécessairement associé à une valeur propre réelle.

Il est facile de voir que la fonction $\boldsymbol{f} : \mathbb{R} \to \mathbb{C}$ définie par $\boldsymbol{f}(t) := \mathrm{e}^{\lambda t}\boldsymbol{v}$ est une solution à valeurs complexes de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$.

Exercice 15.

Le vérifier.

Solution. On a $\boldsymbol{f}'(t) = \lambda \mathrm{e}^{\lambda t}\boldsymbol{v}$ et $\boldsymbol{A} \cdot \boldsymbol{f}(t) = \mathrm{e}^{\lambda t} A \cdot \boldsymbol{v} = \lambda \mathrm{e}^{\lambda t}\boldsymbol{v}$.

Notons que $\boldsymbol{f} = \boldsymbol{f}_1 + i\boldsymbol{f}_2$ avec $\boldsymbol{f}_1 : \mathbb{R} \to \mathbb{R}^2$ et $\boldsymbol{f}_2 : \mathbb{R} \to \mathbb{R}^2$ définies par :

$$\boldsymbol{f}_1(t) := \mathrm{e}^{\rho t}\big(\cos(\omega t)\boldsymbol{v}_1 - \sin(\omega t)\boldsymbol{v}_2\big) \quad \text{et} \quad \boldsymbol{f}_2(t) := \mathrm{e}^{\rho t}\big(\sin(\omega t)\boldsymbol{v}_1 + \cos(\omega t)\boldsymbol{v}_2\big).$$

Ces deux fonctions sont linéairement indépendantes puisque, par exemple, $\boldsymbol{f}_1(0) = \boldsymbol{v}_1$ et $\boldsymbol{f}_2(0) = \boldsymbol{v}_2$ qui ne sont pas colinéaires. De plus :

$$\boldsymbol{f}'(t) = \boldsymbol{f}_1'(t) + i\boldsymbol{f}_2'(t) = \boldsymbol{A} \cdot \boldsymbol{f}(t) = \boldsymbol{A} \cdot \boldsymbol{f}_1(t) + \boldsymbol{A} \cdot \boldsymbol{f}_2(t).$$

En identifiant la partie réelle et la partie imaginaire, on voit que $\boldsymbol{f}_1$ et $\boldsymbol{f}_2$ sont solutions de l'équation différentielle. Comme l'ensemble des solutions maximales de l'équation différentielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ est un espace vectoriel de dimension 2, $(\boldsymbol{f}_1, \boldsymbol{f}_2)$ en est une base et les solutions maximales sont les combinaisons linéaires de $\boldsymbol{f}_1$ et $\boldsymbol{f}_2$.

Si $\rho > 0$ les trajectoires sont des spirales issues de $(0,0)$ (quand $t \to -\infty$). Si $\rho < 0$ les trajectoires sont des spirales aboutissant en $(0,0)$ (quand $t \to +\infty$). Si $\rho = 0$, les trajectoires sont des ellipses.

FIGURE IV.6.3 – Deux valeurs propres complexes conjuguées non réelles

3.4 Équations différentielles linéaires d'ordre 2 à coefficients constants

Nous allons maintenant nous intéresser au cas particulier d'une équation différentielle scalaire, linéaire, d'ordre 2, à coefficients constants :

$$ay'' + by' + cy = d(t)$$

avec $a, b, c \in \mathbb{R}$, $a \neq 0$ et $d : I \to \mathbb{R}$ continue. Une fonction $f : J \to \mathbb{R}$ est solution de cette équation différentielle sur un intervalle $J \subset I$ si f est deux fois dérivable sur J et si pour tout $t \in J$, on a $af''(t) + bf'(t) + cf(t) = d(t)$.

3.4.1 Équation vectorielle associée

Pour déterminer les solutions d'une telle équation, on commence par lui associer une équation différentielle vectorielle. Soit $f : J \to \mathbb{R}$ une solution. Considérons la fonction :

$$\boldsymbol{f} : \begin{array}{ccc} J & \longrightarrow & \mathbb{R}^2 \\ t & \longmapsto & \big(f(t), f'(t)\big) \end{array}.$$

Alors $\boldsymbol{f}$ est dérivable sur J et pour tout $t \in J$:

$$\boldsymbol{f}'(t) = \begin{pmatrix} f'(t) \\ f''(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -c/a & -b/a \end{pmatrix} \cdot \begin{pmatrix} f(t) \\ f'(t) \end{pmatrix} + \begin{pmatrix} 0 \\ d(t)/a \end{pmatrix}.$$

Autrement dit, $\boldsymbol{f}$ est solution de l'équation différentielle vectorielle :

$$\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}' + \boldsymbol{b}(t) \quad \text{avec} \quad \boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ -c/a & -b/a \end{pmatrix} \quad \text{et} \quad \boldsymbol{b}(t) = \begin{pmatrix} 0 \\ d(t)/a \end{pmatrix}.$$

Réciproquement, si $\boldsymbol{f} = (f_1, f_2) : J \to \mathbb{R}^2$ est solution de l'équation différentielle :

$$\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t),$$

alors f_1 est solution de l'équation différentielle :

$$a\boldsymbol{y}'' + b\boldsymbol{y}' + c\boldsymbol{y} = d(t)$$

et $f_2 = f_1'$. Autrement dit, la fonction $f : J \to \mathbb{R}$ est solution de l'équation différentielle :

$$ay'' + by' + cy = d(t)$$

si, et seulement si, la fonction $\boldsymbol{f} = (f, f') : J \to \mathbb{R}^2$ est solution de l'équation différentielle :

$$\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t).$$

On en déduit immédiatement le résultat suivant.

Théorème 37. Considérons l'équation différentielle $ay'' + by' + cy = d(t)$ avec $a, b, c \in \mathbb{R}$, $a \neq 0$ et $d : I \to \mathbb{R}$ continue.

1. Toute solution maximale est définie sur I.
2. Si $f_{\text{particulière}} : I \to \mathbb{R}$ est une solution particulière, alors toute solution maximale $f : I \to \mathbb{R}$ est de la forme :
$$f = f_{\text{homogène}} + f_{\text{particulière}}$$
où $f_{\text{homogène}}$ est solution de l'équation homogène $ay'' + by' + cy = 0$.
3. L'ensemble des solutions de l'équation homogène est un espace vectoriel de dimension 2 et l'application qui à une solution f associe le vecteur $\big(f(0), f'(0)\big) \in \mathbb{R}^2$ est un isomorphisme d'espaces vectoriels.

Démonstration.

1. On pose :
$$\boldsymbol{A} := \begin{pmatrix} 0 & 1 \\ -c/a & -b/a \end{pmatrix} \quad \text{et} \quad \boldsymbol{b}(t) := \begin{pmatrix} 0 \\ d(t)/a \end{pmatrix}.$$
Comme les solutions maximales de l'équation :
$$\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t),$$
sont définies sur I, les solutions maximales de l'équation :
$$ay'' + by' + cy = d(t)$$
sont définies sur I.
2. La fonction $f_{\text{particulière}} : I \to \mathbb{R}$ est une solution particulière de l'équation différentielle $ay'' + by' + cy = d(t)$ si, et seulement si, $\boldsymbol{f}_{\text{particulière}} = (f_{\text{particulière}}, f'_{\text{particulière}})$ est une solution particulière de l'équation différentielle :
$$\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t).$$
Les solutions maximales de l'équation avec second membre $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t)$ sont exactement les fonctions $\boldsymbol{f} : I \to \mathbb{R}^2$ de la forme :
$$\boldsymbol{f} = \boldsymbol{f}_{\text{homogène}} + \boldsymbol{f}_{\text{particulière}}$$
où $\boldsymbol{f}_{\text{homogène}} : I \to \mathbb{R}^2$ est solution de l'équation différentielle homogène $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$.
La fonction $f_{\text{homogène}}$ est solution de l'équation homogène $ay'' + by' + cy = 0$ si, et seulement si, $\boldsymbol{f}_{\text{homogène}} = (f_{\text{homogène}}, f'_{\text{homogène}})$ est solution de l'équation homogène $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$.
3. L'application qui a une solution f de l'équation homogène $ay'' + by' + cy = 0$ associe la solution (f, f') de l'équation homogène $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ est un isomorphisme d'espaces vectoriels.
L'ensemble des solutions de l'équation homogène $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ est un espace vectoriel de dimension 2. L'application qui a une solution $\boldsymbol{f} = (f, f')$ associe $\big(f(0), f'(0)\big)$ est un isomorphisme d'espaces vectoriels. ■

Il ne nous reste plus qu'à expliquer comment déterminer les solutions de l'équation homogène et comment déterminer une solution particulière via la méthode de variation de la constante.

3.4.2 Solutions de l'équation homogène

Commençons par déterminer les valeurs propres de la matrice $\boldsymbol{A}$. Observons pour cela que :

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}_2) = \det \begin{pmatrix} -\lambda & 1 \\ -c/a & -b/a - \lambda \end{pmatrix} = \frac{1}{a}(a\lambda^2 + b\lambda + c).$$

Les valeurs propres de la matrice $\boldsymbol{A}$ sont donc les racines de l'équation caractéristique :

$$a\lambda^2 + b\lambda + c = 0.$$

Notons que si λ est valeur propre, alors le vecteur $\boldsymbol{v} = (1, \lambda)$ est un vecteur propre associé à λ. En effet :

$$\boldsymbol{A} \cdot \boldsymbol{v} = \begin{pmatrix} 0 & 1 \\ -c/a & -b/a \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \lambda \end{pmatrix} = \begin{pmatrix} \lambda \\ -\frac{c}{a} - \frac{b}{a}\lambda \end{pmatrix} = \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix}.$$

- S'il y a deux valeurs propres réelles distinctes λ_1 et λ_2, les vecteurs $\boldsymbol{v}_1 := (1, \lambda_1)$ et $\boldsymbol{v}_2 := (1, \lambda_2)$ forment une base de $\mathbb{R}^2$. Les solutions de l'équation $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ sont de la forme $t \mapsto a_1 \mathrm{e}^{\lambda_1 t}\boldsymbol{v}_1 + a_2 \mathrm{e}^{\lambda_2 t}\boldsymbol{v}_2$. En gardant la première coordonnée, on voit que les solutions de l'équation homogène $ay'' + by' + cy = 0$ sont les fonctions de la forme :
$$f : t \mapsto C_1 \mathrm{e}^{\lambda_1 t} + C_2 \mathrm{e}^{\lambda_2 t} \quad \text{avec} \quad C_1, C_2 \in \mathbb{R}.$$
- S'il n'y a qu'une seule valeur propre $\lambda \in \mathbb{R}$, alors $b^2 - 4ac = 0$ et $\lambda = -b/2a$. Le vecteur $\boldsymbol{v}_1 = (1, \lambda)$ est un vecteur propre. De plus, si on pose $\boldsymbol{v}_2 := (0, 1)$, alors :
$$\boldsymbol{A} \cdot \boldsymbol{v}_2 = \begin{pmatrix} 0 & 1 \\ -c/a & -b/a \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -b/a \end{pmatrix}.$$
On a donc :
$$\boldsymbol{A} \cdot \boldsymbol{v}_2 = \begin{pmatrix} \lambda \\ 2\lambda \end{pmatrix} = \boldsymbol{v}_1 + \lambda \boldsymbol{v}_2.$$
Les solutions de l'équation $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ sont les fonctions de la forme :
$$\boldsymbol{f} : t \mapsto \mathrm{e}^{t\lambda}\big((a_1 + ta_2)\boldsymbol{v}_1 + a_2\boldsymbol{v}_2\big).$$
En gardant la première coordonnée, on voit que les solutions de l'équation homogène $ay'' + by' + cy = 0$ sont les fonctions de la forme :
$$f : t \mapsto C_1 \mathrm{e}^{\lambda t} + C_2 t \mathrm{e}^{\lambda_2 t} \quad \text{avec} \quad C_1, C_2 \in \mathbb{R}.$$
- S'il y a deux valeurs propres complexes conjuguées non réelles $\lambda := \rho + i\omega$ et $\bar{\lambda} := \rho - i\omega$ avec $\rho \in \mathbb{R}$ et $\omega \in \mathbb{R} \setminus \{0\}$, le vecteur :
$$\boldsymbol{v} := \begin{pmatrix} 1 \\ \rho + i\omega \end{pmatrix} = \boldsymbol{v}_1 + i\boldsymbol{v}_2 \quad \text{avec} \quad \boldsymbol{v}_1 := \begin{pmatrix} 1 \\ \rho \end{pmatrix} \quad \text{et} \quad \boldsymbol{v}_2 := \begin{pmatrix} 0 \\ \omega \end{pmatrix}$$

est un vecteur propre associé à λ. Les solutions de l'équation $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$ sont les combinaisons linéaires de $\boldsymbol{f}_1$ et $\boldsymbol{f}_2$ avec :

$$\boldsymbol{f}_1(t) = \mathrm{e}^{\rho t}\big(\cos(\omega t)\boldsymbol{v}_1 - \sin(\omega t)\boldsymbol{v}_2\big) \quad \text{et} \quad \boldsymbol{f}_2(t) = \mathrm{e}^{\rho t}\big(\sin(\omega t)\boldsymbol{v}_1 + \cos(\omega t)\boldsymbol{v}_2\big).$$

Les solutions de l'équation homogène $ay''+by'+cy = 0$ sont donc les combinaisons linéaires de :

$$f_1 : t \mapsto \mathrm{e}^{\rho t}\cos(\omega t) \quad \text{et} \quad f_2 : t \mapsto \mathrm{e}^{\rho t}\sin(\omega t),$$

c'est-à-dire les applications de la forme :

$$f : t \mapsto \mathrm{e}^{\rho t}\big(C_1\cos(\omega t) + C_2\sin(\omega t)\big) \quad \text{avec} \quad C_1, C_2 \in \mathbb{R}.$$

On peut résumer l'étude qui vient d'être faite de la manière suivante.

Proposition 38. Pour trouver les solutions de l'équation différentielle homogène à coefficients constants $ay'' + by' + cy = 0$, on commence par chercher les racines de l'équation caractéristique :

$$a\lambda^2 + b\lambda + c = 0.$$

- S'il y a deux racines réelles distinctes λ_1 et λ_2, les solutions sont de la forme :

$$f : t \mapsto C_1\mathrm{e}^{\lambda_1 t} + C_2\mathrm{e}^{\lambda_2 t} \quad \text{avec} \quad C_1, C_2 \in \mathbb{R}.$$

- S'il y a une seule racine λ, les solutions sont de la forme :

$$f : t \mapsto C_1\mathrm{e}^{\lambda t} + C_2 t\mathrm{e}^{\lambda t} \quad \text{avec} \quad C_1, C_2 \in \mathbb{R}.$$

- S'il y a deux racines complexes conjuguées distinctes $\rho + i\omega$ et $\rho - i\omega$, les solutions sont de la forme :

$$f : t \mapsto \mathrm{e}^{\rho t}\big(C_1\cos(\omega t) + C_2\sin(\omega t)\big) \quad \text{avec} \quad C_1, C_2 \in \mathbb{R}.$$

Remarque. S'il y a deux racines complexes conjuguées distinctes $\rho + i\omega$ et $\rho - i\omega$, on peut également exprimer les solutions sous la forme :

$$f : t \mapsto C\mathrm{e}^{\rho t}\cos(\omega t + \varphi) \quad \text{avec} \quad C, \varphi \in \mathbb{R}.$$

3.4.3 Méthode de variation de la constante

Pour trouver une solution particulière de l'équation $ay'' + by' + cy = d(t)$, on cherche une solution particulière de l'équation $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t)$ à l'aide de la méthode de variation de la constante.

Notons que si (f_1, f_2) est une base de solutions de l'équation homogène $ay''+by'+cy = 0$, alors les fonctions $\boldsymbol{f}_1 = (f_1, f_1')$ et $\boldsymbol{f}_2 = (f_2, f_2')$ forment une base de l'ensemble des solutions de l'équation homogène $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y}$.

On cherche les solutions de l'équation différentielle vectorielle $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t)$ sous la forme $\lambda_1\boldsymbol{f}_1 + \lambda_2\boldsymbol{f}_2$ avec $\lambda_1 : I \to \mathbb{R}$ et $\lambda_2 : I \to \mathbb{R}$ dérivables. Si l'on écrit :

$$\boldsymbol{b} = b_1\boldsymbol{f}_1 + b_2\boldsymbol{f}_2$$

avec $b_1 : I \to \mathbb{R}$ et $b_2 : I \to \mathbb{R}$ continues, alors $\boldsymbol{f}$ est solution de $\boldsymbol{y}' = \boldsymbol{A} \cdot \boldsymbol{y} + \boldsymbol{b}(t)$ si, et seulement si, $\lambda_1' = b_1$ et $\lambda_2' = b_2$.

On peut mettre en oeuvre ce procédé de la manière suivante. Étant donnée une base (f_1, f_2) de solutions de l'équation différentielle homogène $ay'' + by' + cy = 0$, on cherche les solutions de l'équation avec second membre $ay'' + by' + cy = d(t)$ sous la forme :

$$f = \lambda_1 f_1 + \lambda_2 f_2$$

avec $\lambda_1 : I \to \mathbb{R}$ et $\lambda_2 : I \to \mathbb{R}$ dérivables en demandant en plus que :

$$f' = \lambda_1 f_1' + \lambda_2 f_2'.$$

On écrit alors une première équation :

$$\lambda_1' f_1 + \lambda_2' f_2 = 0.$$

On a alors :

$$f'' = (\lambda_1 f_1'' + \lambda_2 f_2'') + (\lambda_1' f_1' + \lambda_2' f_2').$$

En écrivant que f est solution de l'équation différentielle avec second membre, on obtient une seconde équation :

$$\lambda_1' f_1' + \lambda_2' f_2' = \frac{d(t)}{a}\cdot$$

On a donc un système de deux équations linéaires à deux inconnues λ_1' et λ_2' que l'on résout.
L'équivalence établie précédemment montre qu'en prenant des primitives quelconques de λ_1' et λ_2', on trouve ainsi une solution particulière de l'équation avec second membre.

Exercice 16.

Résoudre l'équation différentielle :

$$y'' + 3y' + 2y = \frac{t-1}{t^2}\mathrm{e}^{-t} \quad \text{avec} \quad t \in \left]0, +\infty\right[.$$

Solution. L'équation caractéristique est $\lambda^2 + 3\lambda + 2 = 0$ qui admet pour racines -1 et -2. Une base de l'espace vectoriel des solutions de l'équation homogène est donnée par :

$$f_1 : t \mapsto \mathrm{e}^{-t} \quad \text{et} \quad f_2 : t \mapsto \mathrm{e}^{-2t}.$$

On cherche une solution particulière sous la forme :

$$f(t) = \lambda_1(t)\mathrm{e}^{-t} + \lambda_2(t)\mathrm{e}^{-2t}$$

avec :

$$f'(t) = -\lambda_1(t)\mathrm{e}^{-t} - 2\lambda_2(t)\mathrm{e}^{-2t}.$$

Pour cela, il faut que

$$\lambda_1'(t)\mathrm{e}^{-t} + \lambda_2'(t)\mathrm{e}^{-2t} = 0. \tag{1}$$

De plus, pour que f soit solution de l'équation avec second membre, il faut que

$$-\lambda_1'(t)\mathrm{e}^{-t} - 2\lambda_2'(t)\mathrm{e}^{-2t} = \frac{t-1}{t^2}\mathrm{e}^{-t}. \tag{2}$$

Les équations (1) et (2) sont équivalentes à :

$$\lambda_1'(t) = \frac{t-1}{t^2} \quad \text{et} \quad \lambda_2'(t) = \frac{1-t}{t^2}\mathrm{e}^{t}.$$

On peut donc prendre :

$$\lambda_1(t) = \ln(t) + \frac{1}{t} \quad \text{et} \quad \lambda_2(t) = -\frac{\mathrm{e}^t}{t}\cdot$$

La fonction :

$$f_0 : t \mapsto \ln(t)\mathrm{e}^{-t}$$

est donc solution particulière. Les solutions maximales de l'équation avec second membre sont donc les fonctions $f :]0, +\infty[\to \mathbb{R}$ de la forme :

$$f : t \mapsto \ln(t)\mathrm{e}^{-t} + C_1\mathrm{e}^{-t} + C_2\mathrm{e}^{-2t} \quad \text{avec} \quad C_1, C_2 \in \mathbb{R}.$$

Exercice type corrigé

IV.6.0 On considère la suite vectorielle définie par :

$$\boldsymbol{u}_0 := \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{et} \quad \boldsymbol{u}_{n+1} := \begin{pmatrix} 1 & 5 \\ 5 & 1 \end{pmatrix} \cdot \boldsymbol{u}_n.$$

Déterminer l'expression de $\boldsymbol{u}_n$ en fonction de n.

Solution. Posons $A := \begin{pmatrix} 1 & 5 \\ 5 & 1 \end{pmatrix}$. Déterminons d'abord les valeurs propres de A :

$$\det(A - \lambda \mathrm{I}) = \det \begin{pmatrix} 1-\lambda & 5 \\ 5 & 1-\lambda \end{pmatrix} = (1-\lambda)^2 - 5^2.$$

Par conséquent, λ est valeur propre si, et seulement si, $(1-\lambda) = \pm 5$, donc si, et seulement si, $\lambda = 6$ ou $\lambda = -4$.

Déterminons ensuite une base de vecteurs propres de A. Le vecteur $\boldsymbol{v}$ est un vecteur propre associé à la valeur propre $\lambda = 6$ si, et seulement si, $\boldsymbol{v} \neq (0,0)$ et $A\boldsymbol{v} = 6\boldsymbol{v}$. Or :

$$A \cdot \begin{pmatrix} x \\ y \end{pmatrix} = 6 \cdot \begin{pmatrix} x \\ y \end{pmatrix} \iff \begin{cases} x + 5y = 6x \\ 5x + y = 6y \end{cases} \iff x = y.$$

Le vecteur $\boldsymbol{v} := (1,1)$ est donc un vecteur propre associé à la valeur propre 6. Le vecteur $\boldsymbol{w}$ est est un vecteur propre associé à la valeur propre $\lambda = -4$ si, et seulement si, $\boldsymbol{w} \neq (0,0)$ et $A\boldsymbol{w} = -4\boldsymbol{w}$. Or :

$$A \cdot \begin{pmatrix} x \\ y \end{pmatrix} = -4 \cdot \begin{pmatrix} x \\ y \end{pmatrix} \iff \begin{cases} x + 5y = -4x \\ 5x + y = -4y \end{cases} \iff x = -y.$$

Le vecteur $\boldsymbol{w} := (1,-1)$ est donc un vecteur propre associé à la valeur propre -4.

Décomposons maintenant le vecteur $\boldsymbol{u}_0 := (1,2)$ dans la base $(\boldsymbol{v}, \boldsymbol{w})$. On a :

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} = a \begin{pmatrix} 1 \\ 1 \end{pmatrix} + b \begin{pmatrix} 1 \\ -1 \end{pmatrix} \iff \begin{cases} a + b = 1 \\ a - b = 2 \end{cases} \iff \begin{cases} a = 3/2 \\ b = -1/2. \end{cases}$$

Par conséquent :

$$\boldsymbol{u}_n = 6^n \frac{3}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} - (-4)^n \frac{1}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 3 \cdot 6^n - (-4)^n \\ 3 \cdot 6^n + (-4)^n \end{pmatrix}.$$

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Exercices sur la section 1

IV.6.1 Soient $A \in M_m(\mathbb{R})$ et $(A_n)_{n\geqslant 1}$ la suite de matrices définies par $A_n := A - \frac{1}{n}\mathrm{I}_m$.

1. Montrer que la suite (A_n) converge vers A quand $n \to +\infty$.
2. Montrer que pour n assez grand, A_n est une matrice inversible.
3. En déduire que toute matrice $A \in M_m(\mathbb{R})$ peut être approchée par des matrices inversibles.

IV.6.2 Soit $A \in M_2(\mathbb{R})$ une matrice admettant deux valeurs propres réelles $\lambda_1 > \lambda_2 > 0$. On considère la suite $(\boldsymbol{u}_n)_{n\in\mathbb{N}}$ de vecteurs de $\mathbb{R}^2$ définie par :

$$\boldsymbol{u}_0 \in \mathbb{R}^2 \quad \text{et} \quad \boldsymbol{u}_{n+1} := A \cdot \boldsymbol{u}_n.$$

Discuter suivant les valeurs de $\boldsymbol{u}_0$ la limite éventuelle de la suite $\left(\frac{1}{\lambda_1^n}\boldsymbol{u}_n\right)_{n\in\mathbb{N}}$.

IV.6.3 On considère la suite réelle $(u_n)_{n\in\mathbb{N}}$ définie par la récurrence :

$$u_0 := 2, \quad u_1 := -1 \quad \text{et} \quad u_{n+2} := u_{n+1} + 2u_n.$$

1. Déterminer l'expression de u_n en fonction de n.
2. La suite admet-elle une limite quand n tend vers $+\infty$?

Exercices sur la section 2

IV.6.4 Étant donnée une matrice $A \in M_m(\mathbb{R})$, montrer que l'application $f : [0,1] \to M_m(\mathbb{R})$ définie par $f(t) := tA + (1-t)\mathrm{I}_m$ est une application continue et dérivable. Déterminer sa dérivée.

IV.6.5 Soit $\boldsymbol{f} : I \to \mathbb{R}^m$ une application dérivable en x_0. Si $\boldsymbol{f}(x_0) \neq (0,\ldots,0)$, montrer que la fonction $\|\boldsymbol{f}\|$ est dérivable en x_0 et que l'on a :

$$\|\boldsymbol{f}\|'(x_0) = \frac{\big(\boldsymbol{f}'(x_0) \mid \boldsymbol{f}(x_0)\big)}{\|\boldsymbol{f}(x_0)\|}.$$

IV.6.6 Soient $\boldsymbol{f} : I \to \mathbb{R}^3$ et $\boldsymbol{g} : I \to \mathbb{R}^3$ deux applications dérivables en x_0. Montrer que l'application $\boldsymbol{f} \wedge \boldsymbol{g} : I \to \mathbb{R}^3$ est dérivable en x_0 et que :

$$\big(\boldsymbol{f} \wedge \boldsymbol{g}\big)'(x_0) = \boldsymbol{f}'(x_0) \wedge \boldsymbol{g}(x_0) + \boldsymbol{f}(x_0) \wedge \boldsymbol{g}'(x_0).$$

IV.6.7 Soient $I \subset \mathbb{R}$ un intervalle, $\boldsymbol{f} := (f_1, f_2) : I \to \mathbb{R}^2$ et $\boldsymbol{g} := (g_1, g_2) : I \to \mathbb{R}^2$ deux applications dérivables sur I. On considère l'application réelle $h : I \to \mathbb{R}$ définie par :

$$h(x) := \det\begin{pmatrix} f_1(x) & g_1(x) \\ f_2(x) & g_2(x) \end{pmatrix}.$$

Montrer que h est dérivable sur I et que :

$$h'(x) = \det \begin{pmatrix} f_1'(x) & g_1(x) \\ f_2'(x) & g_2(x) \end{pmatrix} + \det \begin{pmatrix} f_1(x) & g_1'(x) \\ f_2(x) & g_2'(x) \end{pmatrix}.$$

IV.6.8 Soit $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^m$ une fonction vectorielle continue et soit $\boldsymbol{F} : \mathbb{R} \to \mathbb{R}^m$ la fonction vectorielle définie par :

$$\boldsymbol{F}(x) := \int_0^x \boldsymbol{f}(t)\, dt.$$

1. Montrer que $\boldsymbol{F}$ est continue sur $\mathbb{R}$.
2. Montrer que $\boldsymbol{F}$ est dérivable sur $\mathbb{R}$ et déterminer sa dérivée.

Exercices sur la section 3

IV.6.9 Tracer l'allure des solutions de l'équation différentielle :

$$\boldsymbol{y}' = \begin{pmatrix} 0 & 1 \\ -6 & 5 \end{pmatrix} \cdot \boldsymbol{y}.$$

IV.6.10 Tracer l'allure des solutions de l'équation différentielle :

$$\boldsymbol{y}' = \begin{pmatrix} 0 & 1 \\ -4 & 4 \end{pmatrix} \cdot \boldsymbol{y}.$$

IV.6.11 Soit $a \in \mathbb{R}$. Montrer que les solutions de l'équation différentielle :

$$\boldsymbol{y}' = \begin{pmatrix} 0 & -a \\ a & 0 \end{pmatrix} \cdot \boldsymbol{y}$$

paramètrent des arcs de cercle dont on précisera le centre.

IV.6.12 Résoudre les équations différentielles suivantes :

1. $y'' - 2y' - 3y = \dfrac{\mathrm{e}^{3t}}{\cosh^2 t}$;
2. $y'' - 2y' + y = t\mathrm{e}^t$;
3. $y'' - 4y' + 3y = 3t + \mathrm{e}^t$;
4. $y'' + 4y = t \sin^2 t$.

Première initiation aux fonctions de plusieurs variables

IV.7

Dans ce module nous allons étudier des fonctions vectorielles de plusieurs variables $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$, où n et m sont des entiers naturels non nuls. Nous avons déjà étudié de telles fonctions dans des cas particuliers :

- $n = m = 1$, ce sont les fonctions réelles d'une variable réelle (voir le module IV.3) ;
- $n = 1$, $m > 1$, ce sont les fonctions vectorielles d'une variable réelle (voir le module IV.6) ; il y a également un traitement géométrique des cas $m = 2$ (courbes du plan) et $m = 3$ (courbes de l'espace) dans le module III.2.

Dans le cours d'algèbre linéaire, nous avons déjà rencontré des fonctions de plusieurs variables : les applications linéaires de $\mathbb{R}^n$ dans $\mathbb{R}^m$ (voir modules II.3, II.4, et II.7). Une connaissance des notions abordées dans ces modules, en particulier en ce qui concerne le traitement des applications linéaires, est une nécessité incontournable pour ce que nous ferons ici. Nous renvoyons également le lecteur au module III.1 où l'on étudie les fonctions affines.
Le but de ce module est d'étendre les notions de continuité et dérivabilité d'une fonction d'une variable réelle au cas des fonctions de plusieurs variables. Nous verrons que sous certaines hypothèses de *différentiabilité*, l'étude du comportement local d'une application non linéaire peut se ramener à l'étude plus simple du comportement d'une application linéaire.
Mentionnons quelques problèmes qui seront traités, pour certains ultérieurement, à l'aide des outils développés dans ce module.

Graphe d'une fonction. Pour comprendre le comportement d'une fonction $f : \mathbb{R} \to \mathbb{R}$, on peut tracer son graphe. Dans ce but, on est généralement amené à étudier le sens de variation de la fonction, les maxima et les minima locaux. Pour cela, on étudie le signe de la dérivée. Si l'on veut étudier une fonction $f : \mathbb{R}^2 \to \mathbb{R}$, on peut tracer l'ensemble des points $(x, y, z) \in \mathbb{R}^3$ tels que $z = f(x, y)$. On obtiendra un relief, avec des collines, des vallées, des sommets, des cuvettes, etc. Comment peut-on trouver les sommets, les lignes de crêtes, le fond des lacs ? Le problème majeur consiste d'abord à déterminer comment étendre la notion de dérivée. C'est ce que nous ferons dans ce module.

Courbes du plan données par une équation. Une droite du plan peut être décrite par une équation affine $ax+by=c$. Un cercle du plan de rayon r centré à l'origine peut être décrit par une équation $x^2+y^2=r^2$. Si l'on se donne une fonction $f:\mathbb{R}^2\to\mathbb{R}$, il est donc naturel de chercher à comprendre à quoi ressemble l'ensemble des points $(x,y)\in\mathbb{R}^2$ tels que $f(x,y)=0$. Si l'on se représente le graphe de la fonction f comme un relief dans $\mathbb{R}^3$, pour chaque $t\in\mathbb{R}$, l'ensemble :

$$X_t=\big\{(x,y)\in\mathbb{R}^2\;\big|\;f(x,y)=t\big\}$$

est une *courbe de niveau*. Les habitués de randonnée en montagne savent bien à quel point comprendre les lignes de niveau peut être important pour comprendre le relief. Par exemple, sur une carte topographique, on repère une région très pentue au fait que les lignes de niveau sont très serrées, alors que dans une région plate ou très peu pentue, les lignes de niveau sont très espacées.
Nous introduisons dans ce module les premiers outils qui nous permettrons par la suite de décrire de tels ensembles.

Surfaces données par une équation. Si l'on considère la fonction $f:\mathbb{R}^3\to\mathbb{R}$ définie par $f(x,y,z):=2x+3y-z$, alors pour $t\in\mathbb{R}$, l'ensemble $P_t:=\big\{(x,y,z)\;\big|\;f(x,y,z)=t\big\}$ est un plan et pour des valeurs de t distinctes, on obtient des plans parallèles.
Si l'on considère la fonction $f:\mathbb{R}^3\to\mathbb{R}$ définie par $f(x,y,z):=x^2+y^2+z^2$, alors pour $t>0$, l'ensemble $P_t:=\big\{(x,y,z)\;\big|\;f(x,y,z)=t\big\}$ est une sphère centrée en $(0,0,0)$.
Plus généralement, étant donnée une fonction $f:\mathbb{R}^3\to\mathbb{R}$, on peut se demander à quoi ressemble l'ensemble des points $(x,y,z)\in\mathbb{R}^3$ tels que $f(x,y,z)=t$, $t\in\mathbb{R}$.

Surfaces paramétrées. Dans le module III.2, nous étudions les courbes paramétrées du plan (ou de l'espace), c'est-à-dire, l'image d'un intervalle $I\subset\mathbb{R}$ par une application continue $\gamma:I\to\mathbb{R}^2$ (ou $\gamma:I\to\mathbb{R}^3$).
Il est possible de mener une étude similaire (bien qu'autrement plus complexe) des surfaces paramétrées. Autrement dit, étant donnée une application continue $\boldsymbol{f}:U\subset\mathbb{R}^2\to\mathbb{R}^3$, le problème consiste à décrire l'image $\boldsymbol{f}(U)$.
Par exemple, l'image de $\mathbb{R}^2$ par l'application :

$$\boldsymbol{f}:\begin{pmatrix}x\\y\end{pmatrix}\mapsto\begin{pmatrix}\dfrac{2x}{1+x^2+y^2}\\[2ex]\dfrac{2y}{1+x^2+y^2}\\[2ex]\dfrac{1-x^2-y^2}{1+x^2+y^2}\end{pmatrix}$$

est la sphère unité de $\mathbb{R}^3$ privée du point $A:=(0,0,-1)$. L'application $\boldsymbol{f}$ est injective. Si $M:=(x,y,0)\in\mathbb{R}^3$, le point $\boldsymbol{f}(x,y)$ est le point d'intersection de la demi-droite $]AM)$ avec la sphère unité.

1 Continuité

Avant de nous lancer dans la définition de la continuité d'une fonction de plusieurs variables, nous allons introduire quelques notions qui nous seront fort utiles. Il s'agit par exemple de généraliser à $\mathbb{R}^n$ les notions d'intervalles fermés et d'intervalles ouverts.

1.1 Ouverts, fermés et compacts

Rappelons que $\|\boldsymbol{x}-\boldsymbol{x}_0\|$ est la distance entre $\boldsymbol{x}\in\mathbb{R}^n$ et $\boldsymbol{x}_0\in\mathbb{R}^n$. On considère que $\boldsymbol{x}$ est « proche » de $\boldsymbol{x}_0$ lorsque $\|\boldsymbol{x}-\boldsymbol{x}_0\|$ est petit. Dans le cas de deux réels x et x_0, nous avons traduit cela par la propriété $|x-x_0|<\varepsilon$, c'est-à-dire $x\in\,]x_0-\varepsilon,x_0+\varepsilon[$. Dans $\mathbb{R}^n$, la notion d'intervalle centré en x_0 est remplacé par la notion de boule centrée en $\boldsymbol{x}_0$.

Définition 1 (Boule ouverte). Pour tout $\boldsymbol{x}_0\in\mathbb{R}^n$ et tout $r>0$, la boule ouverte de centre $\boldsymbol{x}_0$ et de rayon r est l'ensemble :

$$B(\boldsymbol{x}_0,r):=\left\{\boldsymbol{x}\in\mathbb{R}^n \;\middle|\; \|\boldsymbol{x}-\boldsymbol{x}_0\|<r\right\}.$$

Exemple. Dans $\mathbb{R}$, la boule $B(x_0,r)$ est l'intervalle $]x_0-r,x_0+r[$; dans $\mathbb{R}^2$, c'est le disque centré en x_0 de rayon r ; dans $\mathbb{R}^3$, c'est une vraie boule centrée en x_0 de rayon r.

Un ouvert de $\mathbb{R}^n$ est un ensemble qui, s'il contient un point $\boldsymbol{x}_0$, contient tous les points de $\mathbb{R}^n$ « suffisamment proches » de $\boldsymbol{x}_0$.

Définition 2. Un ensemble $U\subset\mathbb{R}^n$ est un ouvert de $\mathbb{R}^n$ si pour tout point $\boldsymbol{x}_0\in U$, il existe $r>0$ tel que $B(\boldsymbol{x}_0,r)\subset U$.

Exemple. Si $a<b$, l'intervalle $]a,b[$ est ouvert dans $\mathbb{R}$, mais l'intervalle $[a,b]$ n'est pas ouvert. En effet, il y a des points proches de a qui ne sont pas dans l'intervalle $[a,b]$. L'intervalle $]a,+\infty[$ est ouvert dans $\mathbb{R}$.

Exercice 1.
Montrer que le rectangle $]0,1[\times]0,2[$ est ouvert dans $\mathbb{R}^2$.
Solution. Supposons $\boldsymbol{x}_0:=(x,y)\in\,]0,1[\times]0,2[$.
Posons $r=\min\{|x|,|x-1|,|y|,|y-2|\}$. Alors $B(\boldsymbol{x}_0,r)\subset\,]0,1[\times]0,2[$.

Intuitivement, un ensemble F de $\mathbb{R}^n$ est fermé s'il contient tous les points qui sont à la frontière de F et de son complémentaire. Le problème est de définir ce que l'on appelle la frontière de F. On utilise donc la définition alternative suivante.

Définition 3. Un ensemble $F\subset\mathbb{R}^n$ est un fermé de $\mathbb{R}^n$ si son complémentaire $U:=\mathbb{R}^n\setminus F$ est un ouvert de $\mathbb{R}^n$.

Exemples.

1. Les intervalles $[a,b]$ et $[a,+\infty[$ sont fermés dans $\mathbb{R}$.
2. Le carré $[0,1]\times[0,1]$ est fermé dans $\mathbb{R}^2$.
3. Le segment $[0,i]$ est fermé dans $\mathbb{C}$ (ou de manière équivalente, le segment $\{0\}\times[0,1]$ est fermé dans $\mathbb{R}^2$).

Théorème 1. Un ensemble $F \subset \mathbb{R}^n$ est fermé dans $\mathbb{R}^n$ si, et seulement si, pour toute suite $(\boldsymbol{x}_k)$ dans F qui converge vers $\boldsymbol{a} \in \mathbb{R}^n$, on a $\boldsymbol{a} \in F$.

Démonstration. Montrons d'abord par l'absurde que si F est fermé, alors toute suite d'éléments de F qui converge, a sa limite dans F. Supposons que ce ne soit pas le cas. Il existe une suite $(\boldsymbol{x}_k)$ dans F qui converge vers $\boldsymbol{a} \in \mathbb{R}^n$ avec $\boldsymbol{a} \notin F$. Alors, $\boldsymbol{a}$ est dans le complémentaire de F qui est ouvert. Donc, il existe $r > 0$ tel que $B(\boldsymbol{a}, r)$ est contenu dans le complémentaire de F. Comme la suite $(\boldsymbol{x}_k)$ converge vers $\boldsymbol{a}$, on a pour k assez grand $\boldsymbol{x}_k \in B(\boldsymbol{a}, r) \subset \mathbb{R}^n \setminus F$, ce qui contredit $\boldsymbol{x}_k \in F$.
Montrons que si toute suite d'éléments de F qui converge a sa limite dans F, alors F est fermé. Sinon, le complémentaire de F n'est pas ouvert, ce qui signifie qu'il existe un point $\boldsymbol{a} \in \mathbb{R}^n \setminus F$ telle que pour tout $r > 0$, la boule $B(\boldsymbol{a}, r)$ rencontre F. Pour chaque $k \geqslant 1$, on peut donc choisir un point $\boldsymbol{x}_k \in B(\boldsymbol{a}, 1/k) \cap F$. La suite $(\boldsymbol{x}_k)$ est contenue dans F et converge vers $\boldsymbol{a}$, ce qui contredit le fait que $\boldsymbol{a}$ n'appartient pas à F. ■

Le théorème 31 de la page 638 affirme que si une fonction $f : [a, b] \to \mathbb{R}$ est continue, elle est bornée et atteint ses bornes. Pour étendre ce théorème, nous introduisons maintenant deux notions qui généralisent celle d'intervalle borné et celle d'intervalle fermé borné.

Définition 4. Un ensemble $B \subset \mathbb{R}^n$ est *borné* s'il existe $M \in \mathbb{R}$ tel que pour tout $\boldsymbol{x} \in B$, on ait $\|\boldsymbol{x}\| \leqslant M$.

Exercice 2.
Montrer qu'un ensemble $B \subset \mathbb{R}^n$ est borné si, et seulement s'il existe $\boldsymbol{x}_0 \in \mathbb{R}^n$ et $r \in \mathbb{R}$ tels que B est contenu dans la boule $B(\boldsymbol{x}_0, r)$.
Solution. Si $\|\boldsymbol{x}\| \leqslant M$ alors $\boldsymbol{x} \in B(0, M)$. Si $\boldsymbol{x} \in B(\boldsymbol{x}_0, r)$ alors $\|\boldsymbol{x}\| \leqslant \|\boldsymbol{x}_0\| + r$.

Définition 5. Un ensemble $K \subset \mathbb{R}^n$ est un *compact* de $\mathbb{R}^n$ s'il est fermé et borné.

Une caractérisation fondamentale des compacts de $\mathbb{R}^n$ est la suivante.

Théorème 2 (Bolzano-Weierstraß). Un ensemble $K \subset \mathbb{R}^n$ est compact si, et seulement si, de toute suite $(\boldsymbol{x}_k)$ dans K, on peut extraire une sous-suite qui converge dans K.

Démonstration. Supposons que K est compact. Comme K est borné, la suite $(\boldsymbol{x}_k)$ est bornée. Le théorème de Bolzano-Weierstraß (voir module IV.6) garantit que l'on peut en extraire une sous-suite convergente. Comme K est fermé, la limite de la suite extraite appartient à K.
Réciproquement, supposons que de toute suite $(\boldsymbol{x}_k)$ contenue dans K, on peut extraire une sous-suite qui converge dans K. Alors, toute suite $(\boldsymbol{x}_k)$ dans K qui converge converge dans $\mathbb{R}^n$ a sa limite dans K. Cela montre que K est fermé. Si K n'était pas borné, il y aurait une suite $(\boldsymbol{x}_k)$ dans K telle que $\|\boldsymbol{x}_k\| \underset{k \to +\infty}{\longrightarrow} +\infty$. De cette suite, on ne pourrait pas extraire de sous-suite qui converge dans $\mathbb{R}^n$ (et *a fortiori* dans K). ■

Exercice 3.
Montrer que pour tout $R \geqslant 0$, la boule fermée :

$$B := B(0,R) := \left\{(x,y) \in \mathbb{R}^2 \mid \left\|(x,y)\right\| \leqslant R\right\}$$

est un compact de $\mathbb{R}^2$.
Solution. La boule B est bornée puisque $\boldsymbol{x} \in B \Rightarrow \|\boldsymbol{x}\| \leqslant R$.
Elle est fermée car si $\boldsymbol{x}_k \in B \to \boldsymbol{x} \in \mathbb{R}^2$, alors :

$$\|\boldsymbol{x}\| \leqslant \|\boldsymbol{x} - \boldsymbol{x}_k\| + \|\boldsymbol{x}_k\| \leqslant \|\boldsymbol{x} - \boldsymbol{x}_k\| + R \to R.$$

Donc $\|\boldsymbol{x}\| \leqslant R$ et $\boldsymbol{x} \in B$.

1.2 Fonctions continues

Dans tout la suite de ce module, U désigne est un ouvert.

Définition 6. On dit que $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est continue en $\boldsymbol{a}$ si $\boldsymbol{a} \in U$ et si, pour toute suite $(\boldsymbol{x}_k)$ dans U qui converge vers $\boldsymbol{a}$, on a $\boldsymbol{f}(\boldsymbol{x}_k) \underset{k\to+\infty}{\longrightarrow} \boldsymbol{f}(\boldsymbol{a})$.

On peut donner la définition équivalente suivante.

Définition 7. La fonction $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est continue en $\boldsymbol{a}$ si $\boldsymbol{a} \in U$ et si :

$$\forall \varepsilon > 0,\ \exists \eta > 0, \quad \|\boldsymbol{x} - \boldsymbol{a}\| \leqslant \eta \quad \Longrightarrow \quad \left\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\right\| \leqslant \varepsilon.$$

Définition 8. On dit que $\boldsymbol{f}$ est continue sur un ensemble $U \subset \mathbb{R}^n$ si $\boldsymbol{f}$ est continue en tout point de U.

La proposition suivante est une conséquence immédiate du corollaire 5 de la page 782.

Proposition 3. La fonction $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ est continue sur $\mathbb{R}^n$.

Comme dans le cas des fonctions vectorielles d'une variable réelle on a :

Théorème 4. Une fonction $\boldsymbol{f} := (f_1, \ldots, f_m) : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est continue en $\boldsymbol{a} \in U$ si, et seulement si, chaque fonction $f_i : U \to \mathbb{R}$ est continue en $\boldsymbol{a}$.

Démonstration. Soit $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ une suite qui converge vers $\boldsymbol{a}$. D'après le théorème 6 de la page 782, la suite $\left(\boldsymbol{f}(\boldsymbol{x}_k)\right)_{k\in\mathbb{N}}$ converge vers $\boldsymbol{f}(\boldsymbol{a})$ si, et seulement si, pour tout entier $i \in [1,m]$, la suite $\left(f_i(\boldsymbol{x}_k)\right)_{k\in\mathbb{N}}$ converge vers $f_i(\boldsymbol{a})$, la $i^{ème}$ coordonnée de $\boldsymbol{f}(\boldsymbol{a})$. ■

Corollaire 5. Toute application linéaire $L : \mathbb{R}^n \to \mathbb{R}^m$ est continue sur $\mathbb{R}^n$.

Démonstration. La $i^{ème}$ coordonnée de $L(x_1, \ldots, x_n)$ est une combinaison linéaire des x_i :

$$L_i(x_1, \ldots, x_n) = A_1 x_1 + \ldots + A_n x_n.$$

On a donc :

$$\big\|L_i(\boldsymbol{x})-L_i(\boldsymbol{y})\big\| \leqslant |A_1(x_1-y_1)|+\ldots+|A_n(x_n-y_n)| \leqslant \big(|A_1|+\ldots+|A_n|\big)\|\boldsymbol{x}-\boldsymbol{y}\|. \quad \blacksquare$$

Théorème 6 (Opérations élémentaires). Soit U un sous-ensemble de $\mathbb{R}^n$. Soient $\boldsymbol{f} : U \to \mathbb{R}^m$, $\boldsymbol{g} : U \to \mathbb{R}^m$ et $\lambda : U \to \mathbb{R}$ des fonctions continues en $\boldsymbol{a} \in U$. Alors,

1. $\boldsymbol{f}+\boldsymbol{g}$ est continue en $\boldsymbol{a}$;
2. $\lambda\boldsymbol{f}$ est continue en $\boldsymbol{a}$;
3. la fonction produit scalaire $(\,\boldsymbol{f} \mid \boldsymbol{g}\,) : \boldsymbol{x} \mapsto (\,\boldsymbol{f}(\boldsymbol{x}) \mid \boldsymbol{g}(\boldsymbol{x})\,)$ est continue en $\boldsymbol{a}$.

Démonstration. Soit $(\boldsymbol{x}_k)$ une suite qui converge vers $\boldsymbol{a}$. On a $\boldsymbol{f}(\boldsymbol{x}_k) \underset{k\to+\infty}{\longrightarrow} \boldsymbol{f}(\boldsymbol{a})$, $\boldsymbol{g}(\boldsymbol{x}_k) \underset{k\to+\infty}{\longrightarrow} \boldsymbol{g}(\boldsymbol{a})$ et $\lambda(\boldsymbol{x}_k) \underset{k\to+\infty}{\longrightarrow} \lambda(\boldsymbol{a})$ par continuité de $\boldsymbol{f}$, $\boldsymbol{g}$ et λ.

1. La suite $\big(\boldsymbol{f}(\boldsymbol{x}_k)+\boldsymbol{g}(\boldsymbol{x}_k)\big)$ converge donc vers $\boldsymbol{f}(\boldsymbol{a})+\boldsymbol{g}(\boldsymbol{a})$.
2. La suite $\big(\lambda(\boldsymbol{x}_k)\boldsymbol{f}(\boldsymbol{x}_k)\big)$ converge donc vers $\lambda(\boldsymbol{a})\boldsymbol{f}(\boldsymbol{a})$.
3. La suite $\big(\big(\,\boldsymbol{f}(\boldsymbol{x}_k) \,\big|\, \boldsymbol{g}(\boldsymbol{x}_k)\,\big)\big)$ converge donc vers $\big(\,\boldsymbol{f}(\boldsymbol{a}) \,\big|\, \boldsymbol{g}(\boldsymbol{a})\,\big)$. $\blacksquare$

Théorème 7 (Fonctions composées). Soient $U \subset \mathbb{R}^n$ et $V \subset \mathbb{R}^m$.
Si $\boldsymbol{f} : U \to V$ est continue en $\boldsymbol{a}$ et si $\boldsymbol{g} : V \to \mathbb{R}^p$ est continue en $\boldsymbol{f}(\boldsymbol{a})$, alors $\boldsymbol{g}\circ\boldsymbol{f} : U \to \mathbb{R}^p$ est continue en $\boldsymbol{a}$.

Démonstration. Soit $(\boldsymbol{x}_k)$ une suite qui converge vers $\boldsymbol{a}$. Alors, la suite $\big(\boldsymbol{y}_k := \boldsymbol{f}(\boldsymbol{x}_k)\big)$ converge vers $\boldsymbol{f}(\boldsymbol{a})$ par continuité de $\boldsymbol{f}$. La suite $\big(\boldsymbol{g}(\boldsymbol{y}_k) = \boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{x}_k)\big)$ converge donc vers $\boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{a})$. $\blacksquare$

1.3 Théorème des bornes

Nous allons maintenant généraliser au cas des fonctions réelles de plusieurs variables réelles le théorème des bornes des fonctions réelles d'une variable réelle.
Rappelons qu'un compact $K \subset \mathbb{R}^n$ est un ensemble fermé et borné. De toute suite $(\boldsymbol{x}_k)$ dans K, on peut extraire une sous-suite $(\boldsymbol{x}_{k_j})$ qui converge vers une limite $\boldsymbol{a} \in K$.

Théorème 8. Soient $K \subset \mathbb{R}^n$ un compact et $f : K \to \mathbb{R}$ une fonction continue. Alors, f est bornée et atteint ses bornes :

- il existe un point $\boldsymbol{a} \in K$ tel que $f(\boldsymbol{a}) \geqslant f(\boldsymbol{x})$ pour tout $\boldsymbol{x} \in K$,
- il existe un point $\boldsymbol{b} \in K$ tel que $f(\boldsymbol{b}) \leqslant f(\boldsymbol{x})$ pour tout $\boldsymbol{x} \in K$.

Démonstration. Posons

$$M := \sup_{\boldsymbol{x}\in K} f(\boldsymbol{x}).$$

A priori, $M \in \mathbb{R} \cup \{+\infty\}$. Par propriété de la borne supérieure, pour tout $\boldsymbol{x} \in K$, on a $f(\boldsymbol{x}) \leqslant M$ et il existe une suite de points $(\boldsymbol{x}_k \in K)$ tels que $f(\boldsymbol{x}_k) \to M$. Comme K est compact, on peut en extraire une sous-suite $(\boldsymbol{x}_{k_j})$ qui converge vers une limite $\boldsymbol{a} \in K$.

Par continuité de f en $\boldsymbol{a}$, on a $f(\boldsymbol{x}_{k_j}) \underset{j \to +\infty}{\longrightarrow} f(\boldsymbol{a})$.

Donc $M = f(\boldsymbol{a}) \in \mathbb{R}$ et pour tout $x \in K$, $f(\boldsymbol{x}) \leqslant f(\boldsymbol{a})$.

On montre de manière similaire que f est minorée sur K et qu'il existe un point $\boldsymbol{b} \in K$ tel que $f(\boldsymbol{b}) \leqslant f(\boldsymbol{x})$ pour tout $\boldsymbol{x} \in K$. ■

Définition 9. Étant donnée une fonction $f : K \to \mathbb{R}$, s'il existe un point $\boldsymbol{a} \in K$ tel que $f(\boldsymbol{a}) \geqslant f(\boldsymbol{x})$ pour tout $\boldsymbol{x} \in K$, on définit $\max\limits_{\boldsymbol{x} \in K} f(\boldsymbol{x}) := f(\boldsymbol{a})$. S'il existe un point $\boldsymbol{b} \in K$ tel que $f(\boldsymbol{b}) \leqslant f(\boldsymbol{x})$ pour tout $\boldsymbol{x} \in K$, on définit $\min\limits_{\boldsymbol{x} \in K} f(\boldsymbol{x}) := f(\boldsymbol{b})$.

1.4 Norme d'une application linéaire

Corollaire 9. Si $L : \mathbb{R}^n \to \mathbb{R}^m$ est une application linéaire, il existe $\boldsymbol{x}_0 \in \mathbb{R}^n$ tel que $\|\boldsymbol{x}_0\| \leqslant 1$ et :

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \quad \|\boldsymbol{x}\| \leqslant 1 \quad \Longrightarrow \quad \big\|L(\boldsymbol{x})\big\| \leqslant \big\|L(\boldsymbol{x}_0)\big\|.$$

Démonstration. On a vu précédemment que la boule unité fermée de $\mathbb{R}^n$, c'est-à-dire $\big\{\boldsymbol{x} \in \mathbb{R}^n \;\big|\; \|\boldsymbol{x}\| \leqslant 1\big\}$ est compacte. La fonction $L : \mathbb{R}^n \to \mathbb{R}^m$ est continue. La fonction $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|$ est continue. Par composition, la fonction $\boldsymbol{x} \mapsto \big\|L(\boldsymbol{x})\big\|$ est continue. Elle est donc bornée et atteint ses bornes sur la boule unité fermée de $\mathbb{R}^n$. ■

Définition 10. Étant donnée une application linéaire $L : \mathbb{R}^n \to \mathbb{R}^m$, on note $|||L|||$ le réel positif :

$$|||L||| := \max_{\boldsymbol{x} \in \mathbb{R}^n,\ \|\boldsymbol{x}\| \leqslant 1} \big\|L(\boldsymbol{x})\big\|.$$

Proposition 10.
Si $L : \mathbb{R}^n \to \mathbb{R}^m$ est une application linéaire, alors pour tout $\boldsymbol{x} \in \mathbb{R}^n$, on a :

$$\big\|L(\boldsymbol{x})\big\| \leqslant |||L||| \cdot \|\boldsymbol{x}\|.$$

Démonstration. Si $\|\boldsymbol{x}\| = 0$, alors $\big\|L(\boldsymbol{x})\big\| = 0 = |||L||| \cdot \|\boldsymbol{x}\|$.

Si $\|\boldsymbol{x}\| \neq 0$, posons $\boldsymbol{y} := \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\cdot$ Alors, $\|\boldsymbol{y}\| = 1$, $L(\boldsymbol{x}) = \|\boldsymbol{x}\| \cdot L(\boldsymbol{y})$ et :

$$\big\|L(\boldsymbol{x})\big\| = \|\boldsymbol{x}\| \cdot \big\|L(\boldsymbol{y})\big\| \leqslant \|\boldsymbol{x}\| \cdot |||L|||.$$ ■

2 Différentiabilité

Nous avons défini la dérivée d'une fonction, réelle ou vectorielle, d'une variable réelle en un point a comme la limite, si elle existe :

$$\boldsymbol{f}'(a) := \lim_{\substack{x \to a \\ x \neq a}} \frac{\boldsymbol{f}(x) - \boldsymbol{f}(a)}{x - a}.$$

Si l'on considère une fonction de plusieurs variables, il n'est pas possible d'utiliser cette définition. En effet, il n'est pas possible de diviser le réel ou le vecteur $\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})$ par le vecteur $\boldsymbol{x} - \boldsymbol{a}$: on ne divise pas par un vecteur. C'est pourquoi il va falloir adapter la définition de dérivée.

2.1 Dérivées partielles

Commençons par reformuler la définition de la dérivée d'une fonction $f : U \to \mathbb{R}$ définie sur un ouvert $U \subset \mathbb{R}$.

Définition 11. Soient $f : U \subset \mathbb{R} \to \mathbb{R}$ une fonction réelle d'une variable réelle et $a \in U$. La dérivée de f en a est la limite, si elle existe :

$$\frac{\partial f}{\partial x}(a) = f'(a) := \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{1}{h}\big(f(a+h) - f(a)\big).$$

Considérons maintenant le cas d'une fonction réelle de plusieurs variables $f : U \subset \mathbb{R}^n \to \mathbb{R}$. Étant donné un point $\boldsymbol{a} := (a_1, \ldots, a_n) \in U$, la fonction f est dérivable en $\boldsymbol{a}$ par rapport à la première variable si la fonction $g : x \mapsto f(x, a_2, \ldots, a_n)$ est dérivable en a_1. Le réel $g'(a_1)$ est la dérivée partielle de f par rapport à x_1 en $\boldsymbol{a}$.

Définition 12. Soient $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction réelle de plusieurs variables et $\boldsymbol{a} \in U$. La dérivée partielle de f en $\boldsymbol{a}$ par rapport à la $i^{ème}$ variable est la limite, si elle existe :

$$\frac{\partial f}{\partial x_i}(\boldsymbol{a}) := \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{1}{h} \left(f \begin{pmatrix} a_1 \\ \vdots \\ a_i + h \\ \vdots \\ a_n \end{pmatrix} - f \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{pmatrix} \right).$$

Si $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ est la base canonique de $\mathbb{R}^n$, cela s'écrit également :

$$\frac{\partial f}{\partial x_i}(\boldsymbol{a}) = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f(\boldsymbol{a} + h\boldsymbol{e}_i) - f(\boldsymbol{a})}{h}.$$

Exemples.

1. Si $f : \mathbb{R}^2 \to \mathbb{R}$ est définie par $f(x, y) := x^2 + 2xy$, alors :

$$\frac{\partial f}{\partial x}\begin{pmatrix} x \\ y \end{pmatrix} = 2x + 2y \qquad \text{et} \qquad \frac{\partial f}{\partial y}\begin{pmatrix} x \\ y \end{pmatrix} = 2x.$$

2. Si $f : \mathbb{R}^3 \to \mathbb{R}$ est définie par $f(x, y, z) := x^2 + y^2 - z^2$, alors :

$$\frac{\partial f}{\partial x}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = 2x, \qquad \frac{\partial f}{\partial y}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = 2y \qquad \text{et} \qquad \frac{\partial f}{\partial z}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = -2z.$$

Remarque. La notation de la dérivée partielle *oublie* l'information la plus importante : quelles variables sont maintenues constantes. Considérons par exemple la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x, y) := x + (x - y)^2$. Alors, la dérivée partielle de f par rapport à x est :

$$\frac{\partial f}{\partial x}(x, y) = 1 + 2(x - y).$$

Si l'on fait un changement de variables qui préserve x, par exemple :

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x \\ z \end{pmatrix} = \begin{pmatrix} x \\ x - y \end{pmatrix}.$$

Alors, dans le système de coordonnées (x, z), l'expression de la fonction f devient $g : (x, z) \mapsto x + z^2$. La dérivée partielle de g par rapport à x n'est pas égale à la dérivée partielle de f par rapport à x :

$$\frac{\partial g}{\partial x}(x, y) = 1 \neq 1 + 2(x - y) = \frac{\partial f}{\partial x}(x, y).$$

Venons-en maintenant au cas général d'une fonction vectorielle de plusieurs variables réelles. La définition est similaire.

Définition 13. Soient $\boldsymbol{f} := (f_1, \ldots, f_m) : U \subset \mathbb{R}^n \to \mathbb{R}^m$ une fonction vectorielle de plusieurs variables et $\boldsymbol{a} \in U$. La dérivée partielle de $\boldsymbol{f}$ en $\boldsymbol{a}$ par rapport à la $i^{ème}$ variable est la limite, si elle existe :

$$\frac{\partial \boldsymbol{f}}{\partial x_i}(\boldsymbol{a}) := \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{1}{h} \left(\boldsymbol{f} \begin{pmatrix} a_1 \\ \vdots \\ a_i + h \\ \vdots \\ a_n \end{pmatrix} - \boldsymbol{f} \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{pmatrix} \right) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_i}(\boldsymbol{a}) \\ \vdots \\ \dfrac{\partial f_m}{\partial x_i}(\boldsymbol{a}) \end{pmatrix}.$$

Remarque. Une dérivée partielle d'une fonction vectorielle est un vecteur.

Si $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ est la base canonique de $\mathbb{R}^n$, cela s'écrit également :

$$\frac{\partial \boldsymbol{f}}{\partial x_i}(\boldsymbol{a}) = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{\boldsymbol{f}(\boldsymbol{a} + h\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{a})}{h}.$$

Exercice 4.

Déterminer les dérivées partielles de la fonction $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ définie par :

$$\boldsymbol{f} \begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} x^2 - y + 1 \\ x \\ xy \end{pmatrix}.$$

Solution.

$$\frac{\partial \boldsymbol{f}}{\partial x}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x \\ 1 \\ y \end{pmatrix} \quad \text{et} \quad \frac{\partial \boldsymbol{f}}{\partial y}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ x \end{pmatrix}.$$

2.2 Dérivée suivant un vecteur

La dérivée partielle :

$$\frac{\partial \boldsymbol{f}}{\partial x_i}(\boldsymbol{a}) := \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{\boldsymbol{f}(\boldsymbol{a} + h\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{a})}{h}$$

mesure la rapidité à laquelle $\boldsymbol{f}$ varie lorsque $\boldsymbol{x}$ s'éloigne de $\boldsymbol{a}$ dans la direction du vecteur $\boldsymbol{e}_i$. On peut mesurer cette variation lorsque $\boldsymbol{x}$ s'éloigne de $\boldsymbol{a}$ dans une direction quelconque.

Définition 14. Soient $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ une fonction vectorielle de plusieurs variables et $\boldsymbol{a} \in U$. La dérivée partielle de $\boldsymbol{f}$ en $\boldsymbol{a}$ suivant le vecteur $\boldsymbol{v} \in \mathbb{R}^n$ est la limite, si elle existe :

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{\boldsymbol{f}(\boldsymbol{a} + h\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{a})}{h}.$$

Exercice 5.

Soit $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ la fonction définie par :

$$\boldsymbol{f}\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} x^2 - y + 1 \\ x \\ xy \end{pmatrix}.$$

Calculer la dérivée partielle de $\boldsymbol{f}$ suivant $\boldsymbol{v} := (1, 2)$ au point $\boldsymbol{a} := (2, 3)$.

Solution. On a :

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{a} + h\boldsymbol{v}) = \boldsymbol{f}\begin{pmatrix} 2 + h \\ 3 + 2h \end{pmatrix} &= \begin{pmatrix} 4 + 4h + h^2 - 3 - 2h + 1 \\ 2 + h \\ 6 + 7h + 2h^2 \end{pmatrix} \\ &= \begin{pmatrix} 2 \\ 2 \\ 6 \end{pmatrix} + h\begin{pmatrix} 2 \\ 1 \\ 7 \end{pmatrix} + h^2\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}. \end{aligned}$$

Comme $\boldsymbol{f}(\boldsymbol{a}) = (2, 2, 6)$, on a :

$$\frac{\boldsymbol{f}(\boldsymbol{a} + h\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{a})}{h} = \begin{pmatrix} 2 \\ 1 \\ 7 \end{pmatrix} + h\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \underset{h \to 0}{\longrightarrow} \begin{pmatrix} 2 \\ 1 \\ 7 \end{pmatrix}.$$

La dérivée de $\boldsymbol{f}$ suivant $\boldsymbol{v} = (1, 2)$ au point $\boldsymbol{a} = (2, 3)$ est le vecteur $(2, 1, 7)$.

2.3 Différentielle

Comme mentionné dans le module IV.3, une autre façon d'écrire la dérivabilité en a d'une fonction réelle f d'une variable réelle est de poser :

$$\delta(x,a) := \frac{f(x)-f(a)}{x-a} - f'(a),$$

de sorte que :

$$f(x) = f(a) + (x-a)f'(a) + (x-a)\delta(x,a) \quad \text{et} \quad \lim_{x\to a} \delta(x,a) = 0. \tag{1}$$

L'application $x \mapsto f(a) + (x-a)f'(a)$ est l'application affine qui approche le mieux f au point a. Le graphe de l'application f est une courbe qui passe par le point $\big(a, f(a)\big)$ et plus l'on zoome sur ce point, plus la forme de la courbe épouse celle d'une droite : la tangente d'équation $y = f(a) + (x-a)f'(a)$.
On peut également écrire l'égalité (1) sous la forme :

$$f(a+h) = f(a) + hf'(a) + h\varepsilon(h) \quad \text{avec} \quad \varepsilon(h) \underset{h\to 0}{\longrightarrow} 0.$$

Nous allons maintenant généraliser cette notion au cas des fonctions vectorielles de plusieurs variables réelles.

Définition 15. Soient $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ une fonction vectorielle de plusieurs variables et $\boldsymbol{a} \in U$. On dit que $\boldsymbol{f}$ est *différentiable* en $\boldsymbol{a}$ s'il existe une application linéaire $L : \mathbb{R}^n \to \mathbb{R}^m$ telle que :

$$\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{a}) + L(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon(\boldsymbol{h}) \quad \text{avec} \quad \varepsilon(\boldsymbol{h}) \underset{h\to 0}{\longrightarrow} \boldsymbol{0}.$$

L'application linéaire L est alors appelée la *différentielle* de $\boldsymbol{f}$ en $\boldsymbol{a}$. On la note $D_{\boldsymbol{a}}\boldsymbol{f}$.

Remarque. Dans la définition précédente, $\varepsilon : \mathbb{R}^n \to \mathbb{R}^m$ est une fonction vectorielle de plusieurs variables.

Si $\boldsymbol{f}$ est différentiable en $\boldsymbol{a}$, l'application affine qui « approche le mieux » $\boldsymbol{f}$ est l'application $\boldsymbol{x} \mapsto \boldsymbol{f}(\boldsymbol{a}) + L(\boldsymbol{x}-\boldsymbol{a})$. Le comportement de cette application affine est relativement simple à étudier. Il nous fournira des informations sur le comportement local de $\boldsymbol{f}$ prés du point $\boldsymbol{a}$.

Proposition 11. Soient $f : U \subset \mathbb{R} \to \mathbb{R}$ une fonction réelle d'une variable réelle et $a \in U$. Alors, f est différentiable en a si, et seulement si, f est dérivable en a. On a alors :

$$D_a f : h \mapsto f'(a) \cdot h.$$

Démonstration. Les applications linéaires $L : \mathbb{R} \to \mathbb{R}$ sont les applications de la forme $h \mapsto \lambda h$ avec $\lambda \in \mathbb{R}$. ■

Proposition 12. Si $\boldsymbol{f}$ est différentiable en $\boldsymbol{a}$, alors $\boldsymbol{f}$ est continue en $\boldsymbol{a}$.

Démonstration. $\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a}) = D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{x}-\boldsymbol{a}) + \|\boldsymbol{x}-\boldsymbol{a}\|\varepsilon(\boldsymbol{x}-\boldsymbol{a}) \underset{\boldsymbol{x}\to\boldsymbol{a}}{\longrightarrow} 0.$ ■

Définition 16. L'application $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est *différentiable sur* U si $\boldsymbol{f}$ est différentiable en tout point $\boldsymbol{a} \in U$.

2.4 Matrice jacobienne

Il existe un lien étroit entre la notion de différentiabilité et celle de dérivée suivant un vecteur.

Proposition 13. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est différentiable en $\boldsymbol{a} \in U$, alors la dérivée partielle de $\boldsymbol{f}$ en $\boldsymbol{a}$ suivant un vecteur quelconque $\boldsymbol{v} \in \mathbb{R}^n$ existe et vaut $D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{v})$.

Démonstration. Si $\boldsymbol{f}$ est différentiable en $\boldsymbol{a}$, alors :

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{a}+h\boldsymbol{v}) &= \boldsymbol{f}(\boldsymbol{a}) + D_{\boldsymbol{a}}\boldsymbol{f}(h\boldsymbol{v}) + \|h\boldsymbol{v}\|\varepsilon(h\boldsymbol{v}) \\ &= \boldsymbol{f}(\boldsymbol{a}) + hD_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{v}) + |h|\varepsilon_1(h) \quad \text{avec} \quad \varepsilon_1(h) \underset{h\to 0}{\longrightarrow} \mathbf{0}.\end{aligned}$$

On a donc :

$$\frac{\boldsymbol{f}(\boldsymbol{a}+h\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{a})}{h} = D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{v}) + \frac{|h|}{h}\varepsilon_1(h) \underset{h\to 0}{\longrightarrow} D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{v}).$$

Attention, la réciproque est fausse. Une application peut avoir des dérivées partielles suivant n'importe quel vecteur sans pour autant être différentiable (ni même continue).

Exercice 6.
Étudier la différentiabilité en $(0,0)$ de la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par :

$$f\begin{pmatrix} x \\ y \end{pmatrix} := \begin{cases} \dfrac{3x^2y - y^3}{x^2+y^2} & \text{si } (x,y) \neq (0,0) \\ 0 & \text{si } (x,y) = (0,0). \end{cases}$$

Solution. Pour tout $h \in \mathbb{R}$, $f(hu, hv) = hf(u,v)$. Par conséquent, f admet une dérivée partielle suivant n'importe quel vecteur $(u,v) \in \mathbb{R}^2$; cette dérivée partielle vaut $f(u,v)$. Cependant, f n'est pas différentiable à l'origine. Sinon, d'après la proposition 13, la différentielle serait l'application $(u,v) \mapsto f(u,v)$, c'est-à-dire qu'on aurait $D_{(0,0)}f = f$. Cependant, f n'est pas une application linéaire.

Exercice 7.

Étudier la continuité en $(0,0)$ de la fonction définie par :

$$f\begin{pmatrix} x \\ y \end{pmatrix} := \begin{cases} \dfrac{|y|}{x^2}e^{-|y|/x^2} & \text{si } x \neq 0 \\ 0 & \text{si } x = 0. \end{cases}$$

Solution. Pour tout $(u,v) \in \mathbb{R}^2 \setminus \{(0,0)\}$, on a $f(hu,hv) \underset{h\to 0}{\longrightarrow} 0$. L'application f admet donc des dérivées partielles en $(0,0)$ suivant n'importe quel vecteur.
Cependant, $f(t,t^2) = e^{-1}$ qui ne tend pas vers $0 = f(0,0)$ quand t tend vers 0. L'application f n'est donc pas continue.

Définition 17. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ admet des dérivées partielles en $\boldsymbol{a}$ par rapport à toutes les variables, on appelle *matrice jacobienne* de $\boldsymbol{f}$ en $\boldsymbol{a}$ la matrice $J_{\boldsymbol{a}}\boldsymbol{f}$ dont la $i^{ème}$ colonne est le vecteur $\dfrac{\partial \boldsymbol{f}}{\partial x_i}(\boldsymbol{a})$.

Exemples. Dans le cas d'une application $\boldsymbol{f} := (f_1, f_2, f_3)$ de $\mathbb{R}^2$ dans $\mathbb{R}^3$, la matrice jacobienne de $\boldsymbol{f}$ au point $\boldsymbol{a} := (a_1, a_2)$ est :

$$J_{\boldsymbol{a}}\boldsymbol{f} = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(a_1,a_2) & \dfrac{\partial f_1}{\partial x_2}(a_1,a_2) \\ \dfrac{\partial f_2}{\partial x_1}(a_1,a_2) & \dfrac{\partial f_2}{\partial x_2}(a_1,a_2) \\ \dfrac{\partial f_3}{\partial x_1}(a_1,a_2) & \dfrac{\partial f_3}{\partial x_2}(a_1,a_2) \end{pmatrix}.$$

Dans le cas d'une application $\boldsymbol{f} := (f_1, f_2)$ de $\mathbb{R}^3$ dans $\mathbb{R}^2$, la matrice jacobienne de $\boldsymbol{f}$ au point $\boldsymbol{a} := (a_1, a_2, a_3)$ est :

$$J_{\boldsymbol{a}}\boldsymbol{f} = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(a_1,a_2,a_3) & \dfrac{\partial f_1}{\partial x_2}(a_1,a_2,a_3) & \dfrac{\partial f_1}{\partial x_3}(a_1,a_2,a_3) \\ \dfrac{\partial f_2}{\partial x_1}(a_1,a_2,a_3) & \dfrac{\partial f_2}{\partial x_2}(a_1,a_2,a_3) & \dfrac{\partial f_2}{\partial x_3}(a_1,a_2,a_3) \end{pmatrix}.$$

Exercice 8.

Déterminer la matrice jacobienne au point (r,θ) de l'application :

$$\begin{array}{rccc} \boldsymbol{f} : & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ & (r,\theta) & \longmapsto & (r\cos\theta, r\sin\theta). \end{array}$$

Solution.

$$J_{(r,\theta)}\boldsymbol{f} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}.$$

Proposition 14. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est différentiable en $\boldsymbol{a}$, alors $\boldsymbol{f}$ admet des dérivées partielles en $\boldsymbol{a}$ par rapport à toutes les variables. La matrice de $D_{\boldsymbol{a}}\boldsymbol{f}$ dans la base canonique est la matrice jacobienne $J_{\boldsymbol{a}}\boldsymbol{f}$.

Démonstration. D'après la proposition 13 de la page 830, on a pour tout $i \in [1, n]$:

$$D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{e}_i) = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{\boldsymbol{f}(\boldsymbol{a} + h\boldsymbol{e}_i) - \boldsymbol{f}(\boldsymbol{a})}{h} = \frac{\partial \boldsymbol{f}}{\partial x_i}(\boldsymbol{a}),$$

où les $\boldsymbol{e}_i$ sont les vecteurs de la base canonique. ■

Comme signalé précédemment, ce n'est pas parce qu'une application admet une matrice jacobienne en tout point que cette application est différentiable (ni même continue).

3 Propriétés fondamentales

3.1 Opérations élémentaires

Théorème 15. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est constante, alors $\boldsymbol{f}$ est différentiable en tout point $\boldsymbol{a} \in U$ et la différentielle de $\boldsymbol{f}$ en $\boldsymbol{a}$ est l'application nulle.

Démonstration. Pour tout $\boldsymbol{a} \in U$, on a $\boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{a}) + \boldsymbol{0}$. ■

Théorème 16. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est une application linéaire, alors $\boldsymbol{f}$ est différentiable en tout point $\boldsymbol{a} \in U$ et la différentielle de $\boldsymbol{f}$ en $\boldsymbol{a}$ est $D_{\boldsymbol{a}}\boldsymbol{f} = \boldsymbol{f}$.

Démonstration. Pour tout $\boldsymbol{a} \in U$, on a $\boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{a}) + \boldsymbol{f}(\boldsymbol{h})$. ■

Théorème 17. Si $\boldsymbol{f} := (f_1, \ldots, f_m)$ et si les m fonctions $f_i : U \subset \mathbb{R}^n \to \mathbb{R}$ sont différentiables en $\boldsymbol{a} \in U$, alors $\boldsymbol{f}$ est différentiable en $\boldsymbol{a}$ est donnée par :

$$D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h}) = \begin{pmatrix} D_{\boldsymbol{a}}f_1(\boldsymbol{h}) \\ \vdots \\ D_{\boldsymbol{a}}f_m(\boldsymbol{h}) \end{pmatrix}.$$

Démonstration. Pour tout $\boldsymbol{a} \in U$, on a :

$$\begin{aligned}
\boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) = \begin{pmatrix} f_1(\boldsymbol{a} + \boldsymbol{h}) \\ \vdots \\ f_m(\boldsymbol{a} + \boldsymbol{h}) \end{pmatrix} &= \begin{pmatrix} f_1(\boldsymbol{a}) + D_{\boldsymbol{a}}f_1(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon_1(\boldsymbol{h}) \\ \vdots \\ f_m(\boldsymbol{a}) + D_{\boldsymbol{a}}f_m(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon_m(\boldsymbol{h}) \end{pmatrix} \\
&= \begin{pmatrix} f_1(\boldsymbol{a}) \\ \vdots \\ f_m(\boldsymbol{a}) \end{pmatrix} + \begin{pmatrix} D_{\boldsymbol{a}}f_1(\boldsymbol{h}) \\ \vdots \\ +D_{\boldsymbol{a}}f_m(\boldsymbol{h}) \end{pmatrix} + \|\boldsymbol{h}\|\varepsilon(\boldsymbol{h})
\end{aligned}$$

avec :

$$\varepsilon(\boldsymbol{h}) = \begin{pmatrix} \varepsilon_1(\boldsymbol{h}) \\ \vdots \\ \varepsilon_m(\boldsymbol{h}) \end{pmatrix} \underset{\boldsymbol{h}\to 0}{\longrightarrow} 0.$$

Théorème 18. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ et $\boldsymbol{g} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ sont différentiables en $\boldsymbol{a} \in U$, alors $\boldsymbol{f} + \boldsymbol{g}$ est différentiable en $\boldsymbol{a}$ et :

$$D_{\boldsymbol{a}}(\boldsymbol{f} + \boldsymbol{g}) = D_{\boldsymbol{a}}\boldsymbol{f} + D_{\boldsymbol{a}}\boldsymbol{g}.$$

Démonstration. Pour tout $\boldsymbol{a} \in U$, on a :

$$\begin{aligned}(\boldsymbol{f} + \boldsymbol{g})(\boldsymbol{a} + \boldsymbol{h}) &= \boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) + \boldsymbol{g}(\boldsymbol{a} + \boldsymbol{h}) \\ &= \Big(\boldsymbol{f}(\boldsymbol{a}) + D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon(\boldsymbol{h})\Big) + \Big(\boldsymbol{g}(\boldsymbol{a}) + D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon(\boldsymbol{h})\Big) \\ &= (\boldsymbol{f} + \boldsymbol{g})(\boldsymbol{a}) + (D_{\boldsymbol{a}}\boldsymbol{f} + D_{\boldsymbol{a}}\boldsymbol{g})(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon(\boldsymbol{h}).\end{aligned}$$

Théorème 19. Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ et $\boldsymbol{g} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ sont différentiables en $\boldsymbol{a} \in U$, alors $f\boldsymbol{g}$ est différentiable en $\boldsymbol{a}$ et :

$$D_{\boldsymbol{a}}(f\boldsymbol{g})(\boldsymbol{h}) = \underbrace{f(\boldsymbol{a})}_{\in\mathbb{R}} \cdot \underbrace{D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h})}_{\in\mathbb{R}^m} + \underbrace{\boldsymbol{g}(\boldsymbol{a})}_{\in\mathbb{R}^m} \cdot \underbrace{D_{\boldsymbol{a}}f(\boldsymbol{h})}_{\in\mathbb{R}}.$$

Remarque. Le lecteur remarquera que dans le deuxième terme, nous multiplions un vecteur (à gauche) par un scalaire (à droite). Cela nous permet d'écrire $D_{\boldsymbol{a}}(f\boldsymbol{g}) = f(\boldsymbol{a}) \cdot D_{\boldsymbol{a}}\boldsymbol{g} + \boldsymbol{g}(\boldsymbol{a}) \cdot D_{\boldsymbol{a}}f$.

Démonstration. On a, pour tout $\boldsymbol{a} \in U$:

$$\begin{aligned}(f\boldsymbol{g})(\boldsymbol{a} + \boldsymbol{h}) &= f(\boldsymbol{a} + \boldsymbol{h})\boldsymbol{g}(\boldsymbol{a} + \boldsymbol{h}) \\ &= \Big(f(\boldsymbol{a}) + D_{\boldsymbol{a}}f(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon_1(\boldsymbol{h})\Big)\Big(\boldsymbol{g}(\boldsymbol{a}) + D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon_2(\boldsymbol{h})\Big) \\ &= (f\boldsymbol{g})(\boldsymbol{a}) + \Big(f(\boldsymbol{a})D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h}) + D_{\boldsymbol{a}}f(\boldsymbol{h})\boldsymbol{g}(\boldsymbol{a})\Big) + \|\boldsymbol{h}\|\varepsilon_3(\boldsymbol{h}).\end{aligned}$$

En effet :

$$|D_{\boldsymbol{a}}f(\boldsymbol{h})| \leqslant |||D_{\boldsymbol{a}}f||| \cdot \|\boldsymbol{h}\| \quad \text{et} \quad \big\|D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h})\big\| \leqslant |||D_{\boldsymbol{a}}\boldsymbol{g}||| \cdot \|\boldsymbol{h}\|,$$

de sorte que :

$$D_{\boldsymbol{a}}f(\boldsymbol{h})D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h}) = \|\boldsymbol{h}\|\varepsilon(\boldsymbol{h}) \quad \text{avec} \quad \varepsilon(\boldsymbol{h}) \underset{\boldsymbol{h}\to 0}{\longrightarrow} \boldsymbol{0}.$$

Corollaire 20.
Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ et $\boldsymbol{g} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ sont différentiables en $\boldsymbol{a} \in U$, alors :

$$J_{\boldsymbol{a}}(f\boldsymbol{g}) = f(\boldsymbol{a})J_{\boldsymbol{a}}\boldsymbol{g} + \boldsymbol{g}(\boldsymbol{a})J_{\boldsymbol{a}}f.$$

Remarque. Attention à l'ordre dans lequel on multiplie le vecteur $\boldsymbol{g}(\boldsymbol{a})$ par la matrice ligne $J_{\boldsymbol{a}}f$, ce qui donne une matrice à m lignes et n colonnes.

Exercice 9.
Calculer la matrice jacobienne de $f\boldsymbol{g}$ pour $f(x,y) := xy$ et $\boldsymbol{g}(x,y) := (x+2y, 3x-y)$.
Solution. L'application :

$$f\boldsymbol{g} : \begin{array}{ccc} \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ (x,y) & \longmapsto & \big(xy(x+2y), xy(3x-y)\big) \end{array}$$

est différentiable et sa matrice jacobienne est :

$$\begin{aligned} J_{(x,y)}(f\boldsymbol{g}) &= xy \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix} + \begin{pmatrix} x+2y \\ 3x-y \end{pmatrix} \cdot \begin{pmatrix} y & x \end{pmatrix} \\ &= \begin{pmatrix} xy+(x+2y)y & 2xy+(x+2y)x \\ 3xy+(3x-y)y & -xy+(3x-y)x \end{pmatrix} \\ &= \begin{pmatrix} 2xy+2y^2 & x^2+4xy \\ 6xy-y^2 & 3x^2-2xy \end{pmatrix}. \end{aligned}$$

Théorème 21. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ et $\boldsymbol{g} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ sont différentiables en $\boldsymbol{a} \in U$, alors le produit scalaire $(\boldsymbol{f} \mid \boldsymbol{g})$ est différentiable en $\boldsymbol{a}$ et :

$$D_{\boldsymbol{a}}(\boldsymbol{f} \mid \boldsymbol{g})(\boldsymbol{h}) = \big(\boldsymbol{f}(\boldsymbol{a}) \,\big|\, D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h})\big) + \big(\boldsymbol{g}(\boldsymbol{a}) \,\big|\, D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})\big).$$

Démonstration. On a, pour tout $\boldsymbol{a} \in U$:

$$\begin{aligned} (\boldsymbol{f} \mid \boldsymbol{g})(\boldsymbol{a}+\boldsymbol{h}) &= \big(\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) \,\big|\, \boldsymbol{g}(\boldsymbol{a}+\boldsymbol{h})\big) \\ &= \big(\boldsymbol{f}(\boldsymbol{a}) + D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon_1(\boldsymbol{h}) \,\big|\, \boldsymbol{g}(\boldsymbol{a}) + D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h}) + \|\boldsymbol{h}\|\varepsilon_2(\boldsymbol{h})\big) \\ &= (\boldsymbol{f} \mid \boldsymbol{g})(\boldsymbol{a}) + \big(\boldsymbol{f}(\boldsymbol{a}) \,\big|\, D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h})\big) + \big(D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h}) \,\big|\, \boldsymbol{g}(\boldsymbol{a})\big) + \|\boldsymbol{h}\|\varepsilon_3(\boldsymbol{h}). \end{aligned}$$

En effet :

$$\big\|D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})\big\| \leqslant |\!|\!| D_{\boldsymbol{a}}\boldsymbol{f} |\!|\!| \cdot \|\boldsymbol{h}\| \quad \text{et} \quad \big\|D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h})\big\| \leqslant |\!|\!| D_{\boldsymbol{a}}\boldsymbol{g} |\!|\!| \cdot \|\boldsymbol{h}\|,$$

de sorte que :

$$\big|\big(D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h}) \,\big|\, D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h})\big)\big| \leqslant \big\|D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})\big\| \cdot \big\|D_{\boldsymbol{a}}\boldsymbol{g}(\boldsymbol{h})\big\| = \|\boldsymbol{h}\|\varepsilon(\boldsymbol{h}) \quad \text{avec} \quad \varepsilon(\boldsymbol{h}) \underset{\|\boldsymbol{h}\|\to 0}{\longrightarrow} \boldsymbol{0}.$$

■

Corollaire 22. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ et $\boldsymbol{g} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ sont différentiables en $\boldsymbol{a} \in U$, alors :

$$J_{\boldsymbol{a}}(\boldsymbol{f} \mid \boldsymbol{g}) = {}^t\boldsymbol{f}(\boldsymbol{a}) \cdot J_{\boldsymbol{a}}\boldsymbol{g} + {}^t\boldsymbol{g}(\boldsymbol{a}) \cdot J_{\boldsymbol{a}}\boldsymbol{f}.$$

Exercice 10.
Calculer la matrice jacobienne de $(\boldsymbol{f} \mid \boldsymbol{g})$ pour

$$\boldsymbol{f}\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} xy \\ \cos(x) \\ \sin(y) \end{pmatrix} \quad \text{et} \quad \boldsymbol{g}\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} 1 \\ \cos(x) \\ \sin(y) \end{pmatrix}.$$

Solution. L'application :

$$(\boldsymbol{f} \mid \boldsymbol{g}) : (x, y) \mapsto xy + \cos^2(x) + \sin^2(y)$$

a pour matrice jacobienne :

$$\begin{aligned} J_{(x,y)}(\boldsymbol{f} \mid \boldsymbol{g}) &= \begin{pmatrix} xy & \cos(x) & \sin(y) \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ -\sin(x) & 0 \\ 0 & \cos(y) \end{pmatrix} \\ &\quad + \begin{pmatrix} 1 & \cos(x) & \sin(y) \end{pmatrix} \cdot \begin{pmatrix} y & x \\ -\sin(x) & 0 \\ 0 & \cos(y) \end{pmatrix} \\ &= \begin{pmatrix} 0 - \cos(x)\sin(x) + 0 & 0 + 0 + \sin(y)\cos(y) \end{pmatrix} \\ &\quad + \begin{pmatrix} y - \cos(x)\sin(x) + 0 & x + 0 + \sin(y)\cos(y) \end{pmatrix} \\ &= \begin{pmatrix} y - 2\cos(x)\sin(x) & x + 2\sin(y)\cos(y) \end{pmatrix}. \end{aligned}$$

3.2 Différentielle d'une application composée

Rappelons que si $f : U \subset \mathbb{R} \to V \subset \mathbb{R}$ est dérivable en a et si $g : V \to \mathbb{R}$ est dérivable en $f(a)$, alors $g \circ f : U \to \mathbb{R}$ est dérivable en a et :

$$(g \circ f)'(a) = g'(f(a)) \cdot f'(a).$$

Ce résultat s'étend aux fonctions vectorielles de plusieurs variables de la manière suivante.

Théorème 23. Soient $U \subset \mathbb{R}^n$ et $V \subset \mathbb{R}^m$. Si $\boldsymbol{f} : U \to V$ est différentiable en $\boldsymbol{a}$ et si $\boldsymbol{g} : V \to \mathbb{R}^p$ est différentiable en $\boldsymbol{f}(\boldsymbol{a})$, alors $\boldsymbol{g} \circ \boldsymbol{f} : U \to \mathbb{R}^p$ est différentiable en $\boldsymbol{a}$ et :

$$D_{\boldsymbol{a}}(\boldsymbol{g} \circ \boldsymbol{f}) = D_{\boldsymbol{f}(\boldsymbol{a})}\boldsymbol{g} \circ D_{\boldsymbol{a}}\boldsymbol{f}.$$

Démonstration. Rappelons que :

$$\boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{a}) + D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h}) + \delta_1(\boldsymbol{h}) \quad \text{avec} \quad \delta_1(\boldsymbol{h}) = \|\boldsymbol{h}\|\varepsilon_1(\boldsymbol{h}) \quad \text{et} \quad \varepsilon_1(\boldsymbol{h}) \underset{\|\boldsymbol{h}\| \to 0}{\longrightarrow} \boldsymbol{0}$$

et que, en notant $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$:

$$\boldsymbol{g}(\boldsymbol{b} + \boldsymbol{k}) = \boldsymbol{g}(\boldsymbol{b}) + D_{\boldsymbol{b}}\boldsymbol{g}(\boldsymbol{k}) + \delta_2(\boldsymbol{k}) \quad \text{avec} \quad \delta_2(\boldsymbol{k}) = \|\boldsymbol{k}\|\varepsilon_2(\boldsymbol{k}) \quad \text{et} \quad \varepsilon_2(\boldsymbol{k}) \underset{\|\boldsymbol{k}\| \to 0}{\longrightarrow} \boldsymbol{0}.$$

On écrit alors :

$$\begin{aligned}\boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) &= \boldsymbol{g}\big(\boldsymbol{f}(\boldsymbol{a})+D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})+\delta_1(\boldsymbol{h})\big)\\ &= \boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{a})+D_{\boldsymbol{f}(\boldsymbol{a})}\boldsymbol{g}\circ D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})+D_{\boldsymbol{f}(\boldsymbol{a})}\boldsymbol{g}\big(\delta_1(\boldsymbol{h})\big)+\delta_2\big(D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})+\delta_1(\boldsymbol{h})\big).\end{aligned}$$

Notons que :

$$\big\|D_{\boldsymbol{f}(\boldsymbol{a})}\boldsymbol{g}\big(\delta_1(\boldsymbol{h})\big)\big\| \leqslant |\!|\!|D_{\boldsymbol{f}(\boldsymbol{a})}\boldsymbol{g}|\!|\!|\cdot\|\delta_1(\boldsymbol{h})\| = \|\boldsymbol{h}\|\varepsilon_3(\boldsymbol{h}) \quad \text{avec} \quad \varepsilon_3(\boldsymbol{h}) \underset{\|\boldsymbol{h}\|\to 0}{\longrightarrow} \boldsymbol{0}.$$

De plus :

$$\big\|D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})\big\| \leqslant |\!|\!|D_{\boldsymbol{a}}\boldsymbol{f}|\!|\!|\cdot\|\boldsymbol{h}\|,$$

ce qui implique que :

$$\big\|\delta_2\big(D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})+\delta_1(\boldsymbol{h})\big)\big\| = \|\boldsymbol{h}\|\varepsilon_4(\boldsymbol{h}) \quad \text{avec} \quad \varepsilon_4(\boldsymbol{h}) \underset{\|\boldsymbol{h}\|\to 0}{\longrightarrow} \boldsymbol{0}.$$

Donc :

$$\boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) = \boldsymbol{g}\circ\boldsymbol{f}(\boldsymbol{a})+D_{\boldsymbol{f}(\boldsymbol{a})}\boldsymbol{g}\circ D_{\boldsymbol{a}}\boldsymbol{f}(\boldsymbol{h})+\|\boldsymbol{h}\|\varepsilon_5(\boldsymbol{h}) \quad \text{avec} \quad \varepsilon_5(\boldsymbol{h}) \underset{\|\boldsymbol{h}\|\to 0}{\longrightarrow} \boldsymbol{0}. \quad \blacksquare$$

Corollaire 24. Si $\boldsymbol{f} : U \to V$ est différentiable en $\boldsymbol{a}$ et si $\boldsymbol{g} : V \to \mathbb{R}^p$ est différentiable en $\boldsymbol{f}(\boldsymbol{a})$, alors les matrices jacobiennes vérifient :

$$J_{\boldsymbol{a}}(\boldsymbol{g}\circ\boldsymbol{f}) = J_{\boldsymbol{f}(\boldsymbol{a})}\boldsymbol{g}\cdot J_{\boldsymbol{a}}\boldsymbol{f}.$$

Exercice 11.

Déterminer la matrice jacobienne de $\boldsymbol{g}\circ\boldsymbol{f}$ pour :

$$\boldsymbol{f}\begin{pmatrix} x\\ y\\ z\end{pmatrix} := \begin{pmatrix} x^2-y^2\\ xz\end{pmatrix} \quad \text{et} \quad \boldsymbol{g}\begin{pmatrix} u\\ v\end{pmatrix} := \begin{pmatrix} \sin(u)\\ \cos(v)\\ uv\end{pmatrix}.$$

Solution. On a :

$$J_{(x,y,z)}\boldsymbol{f} = \begin{pmatrix} 2x & -2y & 0\\ z & 0 & x\end{pmatrix} \quad \text{et} \quad J_{(u,v)}\boldsymbol{g} = \begin{pmatrix} \cos(u) & 0\\ 0 & -\sin(v)\\ v & u\end{pmatrix}.$$

Par conséquent, la matrice jacobienne de $\boldsymbol{g}\circ\boldsymbol{f}$ est :

$$\begin{aligned}J_{(x,y,z)}(\boldsymbol{g}\circ\boldsymbol{f}) &= J_{(x^2-y^2,xz)}\boldsymbol{g}\cdot J_{(x,y,z)}\boldsymbol{f}\\ &= \begin{pmatrix} \cos(x^2-y^2) & 0\\ 0 & -\sin(xz)\\ xz & x^2-y^2\end{pmatrix}\cdot\begin{pmatrix} 2x & -2y & 0\\ z & 0 & x\end{pmatrix}\\ &= \begin{pmatrix} 2x\cos(x^2-y^2) & -2y\cos(x^2-y^2) & 0\\ -z\sin(xz) & 0 & -x\sin(xz)\\ 2x^2z+z(x^2-y^2) & -2xyz & x(x^2-y^2)\end{pmatrix}.\end{aligned}$$

Nous avons vu que si I est un intervalle et si $f : I \subset \mathbb{R} \to \mathbb{R}$ est dérivable sur I et si $f' > 0$ sur I (ou bien $f' < 0$ sur I), alors sa réciproque f^{-1} est dérivable sur $f(I)$, et vérifie :

$$\left(f^{-1}\right)'(y) = \frac{1}{f'\left(f^{(-1)}(y)\right)}.$$

Dans le volume de L2, nous verrons qu'il y a une version locale de ce théorème dans le cas de fonctions vectorielles de plusieurs variables : le théorème d'inversion locale. Ici, nous en donnons une version « faible ».

Théorème 25. Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to V \subset \mathbb{R}^m$ est une bijection différentiable sur U et si la fonction réciproque $\boldsymbol{f}^{-1} : V \to U$ est différentiable sur V, alors $m = n$ et pour tout $\boldsymbol{a} \in U$:

$$D_{\boldsymbol{f}(\boldsymbol{a})}(\boldsymbol{f}^{-1}) = (D_{\boldsymbol{a}}\boldsymbol{f})^{-1}.$$

Démonstration. Comme :

$$\boldsymbol{f}^{-1} \circ \boldsymbol{f} = \mathrm{Id}_U \quad \text{et} \quad \boldsymbol{f} \circ \boldsymbol{f}^{-1} = \mathrm{Id}_V,$$

le théorème 23 de la page 835 implique que :

$$D_{\boldsymbol{f}(\boldsymbol{a})}(\boldsymbol{f}^{-1}) \circ D_{\boldsymbol{a}}\boldsymbol{f} = \mathrm{Id}_{\mathbb{R}^n} \quad \text{et} \quad D_{\boldsymbol{a}}\boldsymbol{f} \circ D_{\boldsymbol{f}(\boldsymbol{a})}(\boldsymbol{f}^{-1}) = \mathrm{Id}_{\mathbb{R}^m}.$$

L'application linéaire $D_{\boldsymbol{a}}\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ est donc inversible (ce qui implique que $m = n$) et d'inverse $D_{\boldsymbol{f}(\boldsymbol{a})}(\boldsymbol{f}^{-1})$. ■

Corollaire 26. Sous les mêmes hypothèse que dans le théorème précédent, on a l'égalité suivante entre matrices jacobiennes :

$$J_{\boldsymbol{f}(\boldsymbol{a})}(\boldsymbol{f}^{-1}) = (J_{\boldsymbol{a}}\boldsymbol{f})^{-1}.$$

Exercice 12.

Montrer que l'application $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^2$ donnée par $\boldsymbol{f}\begin{pmatrix} x \\ y \end{pmatrix} := \begin{pmatrix} x^2 - y + 3 \\ x \end{pmatrix}$ est une bijection différentiable d'inverse différentiable et déterminer la matrice jacobienne de $\boldsymbol{f}^{-1}$.

Solution. Observons que :

$$\begin{cases} X = x^2 - y + 3 \\ Y = x \end{cases} \iff \begin{cases} x = Y \\ y = -X + Y^2 + 3 \end{cases}.$$

L'application $\boldsymbol{f}$ est donc inversible d'inverse :

$$\boldsymbol{f}^{-1}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} Y \\ -X + Y^2 + 3 \end{pmatrix}.$$

On a :

$$J_{(x,y)}\boldsymbol{f} = \begin{pmatrix} 2x & -1 \\ 1 & 0 \end{pmatrix}$$

et :

$$J_{(X,Y)}\boldsymbol{f}^{-1} = \begin{pmatrix} 0 & 1 \\ -1 & 2Y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2x \end{pmatrix} = \begin{pmatrix} 2x & -1 \\ 1 & 0 \end{pmatrix}^{-1}.$$

3.3 Applications continûment différentiables

Comme mentionné précédemment, une application $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ peut avoir des dérivées partielles en tout point $\boldsymbol{a} \in U$ sans pour autant être différentiable. Nous allons voir maintenant que cette pathologie ne se produit pas si les dérivées partielles sont des fonctions continues sur U.

Définition 18. On dit que $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est *continûment différentiable* (ou de classe $\mathcal{C}^1$) sur U si toutes les dérivées partielles de $\boldsymbol{f}$ existent et sont des fonctions continues sur U.

Théorème 27. Si $\boldsymbol{f}$ est continûment différentiable sur U, alors $\boldsymbol{f}$ est différentiable sur U.

Démonstration. Une fonction vectorielle de plusieurs variables est différentiable si, et seulement si, chaque coordonnée est différentiable (théorème 17 de la page 832). Il suffit donc de montrer le théorème pour les fonctions réelles de plusieurs variables.

On écrit :

$$f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a}) = f\begin{pmatrix} a_1+h_1 \\ a_2+h_2 \\ \vdots \\ a_n+h_n \end{pmatrix} - f\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \sum_{i=1}^{n} \Delta_i$$

avec :

$$\Delta_i = f\begin{pmatrix} a_1 \\ \vdots \\ a_{i-1} \\ a_i+h_i \\ a_{i+1}+h_{i+1} \\ \vdots \\ a_n+h_n \end{pmatrix} - f\begin{pmatrix} a_1 \\ \vdots \\ a_{i-1} \\ a_i \\ a_{i+1}+h_{i+1} \\ \vdots \\ a_n+h_n \end{pmatrix}.$$

Pour chaque i, considérons la fonction affine $\boldsymbol{x}_i : [0, h_i] \to \mathbb{R}^n$ définie par :

$$\boldsymbol{x}_i(t) := \begin{pmatrix} a_1 \\ \vdots \\ a_{i-1} \\ a_i+t \\ a_{i+1}+h_{i+1} \\ \vdots \\ a_n+h_n \end{pmatrix}$$

et la fonction réelle d'une variable réelle $g_i = f \circ \boldsymbol{x}_i : [0, h_i] \to \mathbb{R}$. Cette fonction est bien définie dès que les h_i sont suffisamment petits pour que $\boldsymbol{x}_i\big([0, h_i]\big)$ soit contenu dans U.

La fonction g_i est alors continue sur $[0, h_i]$, dérivable sur $]0, h_i[$ et la dérivée vaut :

$$g_i'(t) = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f\big(\boldsymbol{x}_i(t+h)\big) - f\big(\boldsymbol{x}_i(t)\big)}{h} = \frac{\partial f}{\partial x_i}\big(\boldsymbol{x}_i(t)\big).$$

En appliquant le théorème des accroissements finis a $g_i : [0, h_i] \to \mathbb{R}$, on déduit qu'il existe $t_i \in]0, h_i[$ tel que :

$$g_i(h_i) - g_i(0) = h_i g_i'(t_i) = h_i \frac{\partial f}{\partial x_i}\big(\boldsymbol{x}_i(t_i)\big).$$

On a donc :

$$f(\boldsymbol{a} + \boldsymbol{h}) - f(\boldsymbol{a}) = \sum_{i=1}^{n} h_i \frac{\partial f}{\partial x_i}(\boldsymbol{c}_i) \quad \text{avec} \quad \boldsymbol{c}_i = \boldsymbol{x}_i(t_i).$$

Nous pouvons maintenant utiliser l'hypothèse de continuité sur les dérivées partielles de f pour conclure que :

$$f(\boldsymbol{a} + \boldsymbol{h}) - f(\boldsymbol{a}) - J_{\boldsymbol{a}} f \cdot \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} = \sum_{i=1}^{n} h_i \left(\frac{\partial f}{\partial x_i}(\boldsymbol{c}_i) - \frac{\partial f}{\partial x_i}(\boldsymbol{a}) \right) = \|\boldsymbol{h}\| \varepsilon(\boldsymbol{h}).$$

En effet, lorsque $\|\boldsymbol{h}\|$ tend vers 0, on a $\boldsymbol{c}_i \to \boldsymbol{a}$ et :

$$\begin{aligned} \frac{1}{\|\boldsymbol{h}\|} \left| \sum_{i=1}^{n} h_i \left(\frac{\partial f}{\partial x_i}(\boldsymbol{c}_i) - \frac{\partial f}{\partial x_i}(\boldsymbol{a}) \right) \right| &\leqslant \sum_{i=1}^{n} \frac{|h_i|}{\|\boldsymbol{h}\|} \left| \frac{\partial f}{\partial x_i}(\boldsymbol{c}_i) - \frac{\partial f}{\partial x_i}(\boldsymbol{a}) \right| \\ &\leqslant \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(\boldsymbol{c}_i) - \frac{\partial f}{\partial x_i}(\boldsymbol{a}) \right| \underset{\|\boldsymbol{h}\| \to 0}{\longrightarrow} 0. \end{aligned}$$

■

3.4 Théorème des accroissements finis

Nous venons d'utiliser de manière cruciale le théorème des accroissements finis pour les fonctions réelles d'une variable réelle. Nous allons maintenant voir que l'on peut généraliser ce théorème des accroissements finis au cas des fonctions réelles de plusieurs variables. Commençons par définir la notion de segment dans $\mathbb{R}^n$.

Définition 19. Soient $\boldsymbol{a}$ et $\boldsymbol{b}$ deux éléments de $\mathbb{R}^n$. Le *segment* reliant $\boldsymbol{a}$ et $\boldsymbol{b}$, noté $[\boldsymbol{a}, \boldsymbol{b}]$ est l'ensemble des barycentres à coefficients positifs de $\boldsymbol{a}$ et $\boldsymbol{b}$.
Autrement dit :

$$[\boldsymbol{a}, \boldsymbol{b}] := \big\{ (1-t)\,\boldsymbol{a} + t\,\boldsymbol{b} \;\big|\; t \in [0,1] \big\}.$$

On note aussi $]\boldsymbol{a}, \boldsymbol{b}[:= \big\{ (1-t)\,\boldsymbol{a} + t\,\boldsymbol{b} \;\big|\; t \in]0,1[\big\}$.

Attention à cette dernière notation : contrairement à ce qui se passe dans $\mathbb{R}$, $]\boldsymbol{a}, \boldsymbol{b}[$ n'est pas vide lorsque $\boldsymbol{a} = \boldsymbol{b}$ mais est égal à $\{\boldsymbol{a}\}$.

Théorème 28 (des accroissements finis). Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est continue sur le segment $[\boldsymbol{a}, \boldsymbol{b}] \subset U$ et différentiable en tout point de $]\boldsymbol{a}, \boldsymbol{b}[$, alors, il existe $\boldsymbol{c} \in]\boldsymbol{a}, \boldsymbol{b}[$ tel que :

$$f(\boldsymbol{b}) - f(\boldsymbol{a}) = D_{\boldsymbol{c}} f(\boldsymbol{b} - \boldsymbol{a}).$$

Démonstration. Lorsque t parcourt le segment $[0, 1]$, le point $\boldsymbol{x}(t) = (1 - t)\boldsymbol{a} + t\boldsymbol{b}$ parcourt le segment $[\boldsymbol{a}, \boldsymbol{b}]$. Considérons la fonction $g : [0, 1] \to \mathbb{R}$ définie par :

$$g(t) := f\big(\boldsymbol{x}(t)\big).$$

La fonction g est continue sur $[0, 1]$ et différentiable (c'est-à-dire dérivable) sur $]0, 1[$ comme composée de la fonction affine $\boldsymbol{x}$ et de la fonction différentiable f. Pour tout $t \in]0, 1[$, on a :

$$g'(t) = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{g(t + h) - g(t)}{h} = \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f\big(\boldsymbol{x}(t) + h(\boldsymbol{b} - \boldsymbol{a})\big) - f\big(\boldsymbol{x}(t)\big)}{h} = D_{\boldsymbol{x}(t)} f(\boldsymbol{b} - \boldsymbol{a}).$$

Le théorème des accroissements finis d'une variable réelle appliqué à g dit qu'il existe un point $t_0 \in]0, 1[$ tel que :

$$g'(t_0) = \frac{g(1) - g(0)}{1 - 0} = f(\boldsymbol{b}) - f(\boldsymbol{a}).$$

Posons $\boldsymbol{c} = \boldsymbol{x}(t_0)$. Alors, $\boldsymbol{c} \in]\boldsymbol{a}, \boldsymbol{b}[$ et :

$$f(\boldsymbol{b}) - f(\boldsymbol{a}) = g'(t_0) = D_{\boldsymbol{c}} f(\boldsymbol{b} - \boldsymbol{a}).$$

■

Rappelons que le théorème des accroissements finis est valide pour les fonctions réelles de plusieurs variables réelles, mais pas pour les fonctions vectorielles (même d'une variable réelle). Cependant, comme corollaire du théorème des accroissements finis, on obtient l'*inégalité* des accroissements finis pour les fonctions réelles de plusieurs variables.

Corollaire 29. Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est continue sur le segment $[\boldsymbol{a}, \boldsymbol{b}] \subset U$ et différentiable en tout point de $]\boldsymbol{a}, \boldsymbol{b}[$, alors :

$$\big|f(\boldsymbol{b}) - f(\boldsymbol{a})\big| \leqslant \Big(\sup_{\boldsymbol{c} \in]\boldsymbol{a}, \boldsymbol{b}[} |||D_{\boldsymbol{c}} f|||\Big) \|\boldsymbol{b} - \boldsymbol{a}\|.$$

Ce résultat s'étend, lui, au cas des fonctions vectorielles. Rappelons que pour une application linéaire L, $|||L|||$ a été introduit dans la définition 10 de la page 825.

Théorème 30 (Inégalité des accroissements finis). Si $\boldsymbol{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est continue sur le segment $[\boldsymbol{a}, \boldsymbol{b}] \subset U$ et différentiable en tout point de $]\boldsymbol{a}, \boldsymbol{b}[$, alors :

$$\big\|\boldsymbol{f}(\boldsymbol{b}) - \boldsymbol{f}(\boldsymbol{a})\big\| \leqslant \Big(\sup_{\boldsymbol{c} \in]\boldsymbol{a}, \boldsymbol{b}[} |||D_{\boldsymbol{c}} \boldsymbol{f}|||\Big) \|\boldsymbol{b} - \boldsymbol{a}\|.$$

Démonstration. Considérons la fonction vectorielle d'une variable réelle $\boldsymbol{g} : [0, 1] \to \mathbb{R}$ définie par :

$$\boldsymbol{g}(t) := \boldsymbol{f}\big((1 - t)\boldsymbol{a} + t\boldsymbol{b}\big).$$

Comme dans la démonstration précédente, on voit que g est continue sur $[0,1]$, dérivable sur $]0,1[$ et :

$$\boldsymbol{g}'(t) = \lim_{\substack{h\to 0\\ h\neq 0}} \frac{\boldsymbol{f}\big(\boldsymbol{x}(t)+h(\boldsymbol{b}-\boldsymbol{a})\big)-\boldsymbol{f}\big(\boldsymbol{x}(t)\big)}{h} = D_{\boldsymbol{x}(t)}\boldsymbol{f}(\boldsymbol{b}-\boldsymbol{a}).$$

Posons :

$$M = \sup_{\boldsymbol{c}\in]\boldsymbol{a},\boldsymbol{b}[} |||D_{\boldsymbol{c}}f|||.$$

On a alors :

$$\big\|\boldsymbol{g}'(t)\big\| \leqslant M\|\boldsymbol{b}-\boldsymbol{a}\|.$$

D'après l'inégalité des accroissements finis pour les fonctions vectorielles d'une variable réelle, on a donc :

$$\big\|\boldsymbol{f}(\boldsymbol{b})-\boldsymbol{f}(\boldsymbol{a})\big\| = \big\|\boldsymbol{g}(1)-\boldsymbol{g}(0)\big\| \leqslant M\|\boldsymbol{b}-\boldsymbol{a}\|.$$ ■

4 Applications de la notion de différentiabilité

4.1 Plan tangent au graphe d'une fonction $f : \mathbb{R}^2 \to \mathbb{R}$

Rappelons que la courbe représentative d'une fonction $f : I \subset \mathbb{R} \to \mathbb{R}$, où $I \subset \mathbb{R}$ est un intervalle, est la courbe $\gamma : I \to \mathbb{R}^2$ définie par $\gamma(x) := \big(x, f(x)\big)$. Le graphe de f est le support de cette courbe : c'est l'ensemble $\Gamma_f \subset \mathbb{R}^2$ défini par :

$$\Gamma_f := \big\{(x,y)\in\mathbb{R}^2 \;\big|\; x\in I \text{ et } y = f(x)\big\}.$$

Si f est dérivable en $x_0 \in \overset{\circ}{I}$, la courbe représentative γ admet une tangente en x_0. C'est la droite passant par $\gamma(x_0)$ et dirigée par le vecteur $\big(1, f'(x_0)\big)$ (voir module III.2). Une équation de cette droite est :

$$y = f(x_0) + (x-x_0)f'(x_0).$$

On reconnaît les premiers termes du « développement » définissant la dérivée de f en x_0 :

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + (x-x_0)\varepsilon(x) \quad \text{avec} \quad \varepsilon(x) \underset{x\to x_0}{\longrightarrow} 0.$$

Nous allons maintenant généraliser les notions de graphe et de tangente au graphe pour les fonctions réelles de 2 variables réelles.

Définition 20. Le *graphe* d'une fonction $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ est l'ensemble $\Gamma_f \subset \mathbb{R}^3$ défini par :

$$\Gamma_f := \big\{(x,y,z)\in\mathbb{R}^3 \;\big|\; (x,y)\in U \text{ et } z = f(x,y)\big\}.$$

Le graphe d'une fonction continue est une « surface » qui contient un et un seul point (x,y,z) au dessus de chaque point $(x,y)\in U$.

Exemples.

1. Le graphe de la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par :
$$f(x,y) := x^2 + y^2$$
est un paraboloïde de révolution , c'est-à-dire la « surface » obtenue en faisant tourner une parabole autour de son axe de symétrie (voir figure IV.7.1).

FIGURE IV.7.1 – Le graphe de la fonction $(x,y) \mapsto x^2 + y^2$.

2. Si $U \subset \mathbb{R}^2$ est le disque unité (de centre $(0,0)$ et de rayon 1), le graphe de la fonction $f : U \to \mathbb{R}$ définie par :
$$f(x,y) := \sqrt{1 - x^2 - y^2}$$
est une calotte sphérique (la moitié supérieure de la sphère unité de $\mathbb{R}^3$).

Définition 21. Si $M_0 := (x_0, y_0, z_0) \in \Gamma_f$, le plan P est appelé *plan tangent* au graphe de f en M_0 si, pour tout $M \in \Gamma_f$, il existe $M' \in P$ tel que :
$$\|\overrightarrow{MM'}\| = \|\overrightarrow{M_0M}\|\varepsilon(M) \quad \text{avec} \quad \varepsilon(M) \underset{M \to M_0}{\longrightarrow} 0.$$

Proposition 31.
Si $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ est différentiable sur U, alors pour tout $(x_0, y_0) \in U$, le plan P d'équation :
$$z = f(x_0, y_0) + D_{(x_0,y_0)}f \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}$$
est tangent au graphe de f en $\big(x_0, y_0, f(x_0, y_0)\big)$.

Démonstration. Un point $M \in \Gamma_f$ est de la forme $\big(x, y, f(x,y)\big)$ avec $(x,y) \in U$. On lui associe le point $M' := \big(x, y, g(x,y)\big)$ avec :
$$g(x,y) = f(x_0, y_0) + D_{(x_0,y_0)}f \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}.$$
Le point M' appartient au plan P. On a :
$$\begin{aligned} \|\overrightarrow{MM'}\| &= \left| f(x,y) - f(x_0,y_0) + D_{(x_0,y_0)}f \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \right| \\ &= \left\| \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \right\| \varepsilon(x,y) \quad \text{avec} \quad \varepsilon(x,y) \underset{(x,y) \to (x_0,y_0)}{\longrightarrow} 0. \end{aligned}$$

De plus :

$$\|\overrightarrow{M_0M}\| = \left\| \begin{pmatrix} x - x_0 \\ y - y_0 \\ f(x,y) - f(x_0,y_0) \end{pmatrix} \right\| \geqslant \left\| \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \right\|.$$

Par conséquent, on a bien :

$$\|\overrightarrow{MM'}\| = \|\overrightarrow{M_0M}\|\varepsilon(M) \quad \text{avec} \quad \varepsilon(M) \underset{M \to M_0}{\longrightarrow} 0.$$

■

Exercice 13.

Déterminer une équation du plan tangent au graphe de la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x,y) := xy$ au point $(x_0, y_0) \in \mathbb{R}^2$.

Solution. Les dérivées partielles :

$$\frac{\partial f}{\partial x} = y \quad \text{et} \quad \frac{\partial f}{\partial y} = x$$

sont continues sur $\mathbb{R}^2$. La fonction est donc différentiable sur $\mathbb{R}^2$ et la matrice jacobienne est $J_{(x,y)}f = \begin{pmatrix} y & x \end{pmatrix}$. Pour tout $(x_0, y_0) \in \mathbb{R}^2$, le plan tangent au graphe de f au point (x_0, y_0, x_0y_0) est le plan d'équation :

$$z = x_0y_0 + \begin{pmatrix} y_0 & x_0 \end{pmatrix} \cdot \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} = x_0y_0 + y_0(x - x_0) + x_0(y - y_0).$$

Par exemple, à l'origine $(x_0, y_0) = (0,0)$, c'est le plan d'équation $z = 0$.

Recherchons maintenant les points $(x,y) \in U$ où f atteint un maximum ou un minimum (il se peut qu'il n'y ait pas de tel point). Il est plus ou moins évident qu'en un tel point le plan tangent au graphe doit être un plan horizontal, c'est-à-dire un plan d'équation $z =$ constante. Ceci ne se produit que si la différentielle s'annule. C'est ce que dit le résultat suivant.

Proposition 32. Si $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ est différentiable sur U et admet un maximum local ou un minimum local en un point $(x_0, y_0) \in U$, alors la différentielle de f au point (x_0, y_0) est l'application nulle. Autrement dit :

$$\frac{\partial f}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0) = 0.$$

Démonstration. Supposons que f admette un extremum local en (x_0, y_0). Alors, les applications $g_1 : t \mapsto f(x_0 + t, y_0)$ et $g_2 : t \mapsto f(x_0, y_0 + t)$ admettent un extremum local en $t = 0$. L'application f étant différentiable en (x_0, y_0), les applications g_1 et g_2 sont dérivables en $t = 0$:

$$\frac{\partial f}{\partial x}(x_0, y_0) = g_1'(0) = 0 \quad \text{et} \quad \frac{\partial f}{\partial y}(x_0, y_0) = g_2'(0) = 0.$$

La matrice jacobienne de f en (x_0, y_0) est donc la matrice nulle. Par conséquent, la différentielle de f en (x_0, y_0) est l'application nulle. ■

Exercice 14.

Considérons la fonction $f : U \to \mathbb{R}^2$ définie par $f(x,y) := \sqrt{1-x^2-y^2}$ où U est le disque unité de $\mathbb{R}^2$. Étudier le comportement de f en $(0,0)$.

Solution. La fonction f admet un maximum pour $(x,y) = (0,0)$. On a :

$$\frac{\partial f}{\partial x}(x,y) = \frac{-x}{\sqrt{1-x^2-y^2}} \qquad \text{et} \qquad \frac{\partial f}{\partial y}(x,y) = \frac{-y}{\sqrt{1-x^2-y^2}}.$$

Ces dérivées partielles s'annulent en $(x,y) = (0,0)$.

Comme dans le cas d'une fonction réelle d'une variable réelle, il ne suffit pas que la différentielle s'annule pour avoir un extremum local. Dans le cas d'une fonction réelle d'une variable réelle, on peut par exemple avoir un point d'inflexion.

Exercice 15.

Étudier le comportement de la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x,y) := x^2 - y^2$ au voisinage de $(0,0)$.

Solution. La matrice jacobienne de f est :

$$J_{(x,y)}f = \begin{pmatrix} 2x & -2y \end{pmatrix}.$$

Cette matrice est égale à la matrice nulle quand $x = y = 0$. Cependant :

$$f(0,0) = 0, \quad f(0,y) = -y^2 \leqslant 0 \quad \text{et} \quad f(x,0) = x^2 \geqslant 0.$$

La fonction n'admet donc ni maximum ni minimum en $(x,y) = (0,0)$.

L'exercice suivant est un exemple typique d'application des techniques que nous avons développées dans ce module.

Exercice 16.

Soit $f : \mathbb{R}^2 \to \mathbb{R}$ la fonction de deux variables définie par :

$$f(x,y) := (x-y)e^{-(x^2+y^2)}.$$

1. Montrer que $f(x,y) \to 0$ lorsque $\|(x,y)\| \to +\infty$.
2. En déduire que f est bornée et atteint ses bornes dans $\mathbb{R}^2$.
3. Montrer que f est différentiable sur $\mathbb{R}^2$.
4. Déterminer en quels points le graphe de f admet un plan tangent horizontal.
5. Déterminer le minimum et le maximum de f.

Solution.

1. Posons $r := \|(x,y)\|$. Comme $|x-y| \leqslant |x| + |y| \leqslant 2r$, on a :

$$\|f(x,y)\| \leqslant 2re^{-r^2} \underset{r\to+\infty}{\longrightarrow} 0.$$

2. D'après la question précédente, pour tout $\varepsilon > 0$, il existe $R > 0$ tel que si $\|(x,y)\| > R$, alors $\|f(x,y)\| < \varepsilon$. Notons que $f(1,0) = e^{-1} > 0$ et que $f(0,1) = -e^{-1} < 0$. On peut choisir $\varepsilon := e^{-1}$. On voit alors qu'il existe $R > 0$ tel que si $\|(x,y)\| > R$, alors $\|f(x,y)\| < e^{-1}$. La fonction f est continue sur $\mathbb{R}^2$ comme somme, produit et composée de fonctions continues. D'après le théorème des bornes, elle est bornée et atteint ses bornes sur le compact :
$$K := \{(x,y) \in \mathbb{R}^2 \mid \|(x,y)\| \leqslant R\}.$$
Comme $(1,0)$ et $(0,1)$ appartiennent à K, le maximum est supérieur à e^{-1} et le minimum est inférieur $-e^{-1}$. Comme en dehors de K on a $\|f(x,y)\| \leqslant e^{-1}$, on voit que f est bornée sur $\mathbb{R}^2$ et atteint ses bornes dans K.
3. Notons que :
$$\frac{\partial f}{\partial x}(x,y) = (1 - 2x(x-y))e^{-(x^2+y^2)} = (-2x^2 + 2xy + 1)e^{-(x^2+y^2)}$$
et :
$$\frac{\partial f}{\partial y}(x,y) = (-1 - 2y(x-y))e^{-(x^2+y^2)} = (2y^2 - 2xy - 1)e^{-(x^2+y^2)}.$$
Les dérivées partielles de f par rapport à x et à y sont donc continues. La fonction f est donc différentiable sur $\mathbb{R}^2$.
4. Pour que le plan tangent au graphe de f soit horizontal, il faut que la différentielle de f s'annule, c'est-à-dire que :
$$(-2x^2 + 2xy + 1)e^{-(x^2+y^2)} = 0 \quad \text{et} \quad (2y^2 - 2xy - 1)e^{-(x^2+y^2)} = 0.$$
Cela ne peut se produire que lorsque :
$$-2x^2 + 2xy + 1 = 0 \quad \text{et} \quad 2y^2 - 2xy - 1 = 0.$$
Nous sommes donc amenés à résoudre un système de deux équations (non linéaires) à deux inconnues. En ajoutant terme à terme, on voit que $-x^2+y^2 = 0$. Donc $x = \pm y$. Si $x = y$, alors en remplaçant dans la première équation, on obtient $1 = 0$, ce qui n'est pas possible. Donc, $x = -y$. En remplaçant dans la première équation, on obtient $4y^2 = 1$. Donc $y = \pm 1/2$. En résumé, les seuls points où les deux dérivées partielles peuvent s'annuler simultanément sont les points :
$$(-1/2, 1/2) \quad \text{et} \quad (1/2, -1/2).$$
On vérifie facilement qu'en ces points, les dérivées partielles s'annulent effectivement simultanément. Donc, les points où le graphe de f admet un plan tangent horizontal sont les points :
$$(-1/2, 1/2, -e^{-1/2}) \quad \text{et} \quad (1/2, -1/2, e^{-1/2}).$$
5. Aux points où f atteint son maximum, le graphe de f admet un plan horizontal, de même qu'au point où f admet un minimum. On a donc :
$$\min_{(x,y)\in\mathbb{R}^2} f(x,y) = -e^{-1/2} \quad \text{et} \quad \max_{(x,y)\in\mathbb{R}^2} f(x,y) = e^{-1/2}.$$

4.2 Dérivation sur $\mathbb{C}$

Comme mentionné plus haut, il n'est généralement pas possible de définir la dérivée d'une fonction de plusieurs variables réelles comme la limite d'un taux d'accroissement car il n'est pas possible de diviser par un vecteur. On peut cependant le faire dans certains cas, par exemple si l'on identifie le plan réel $\mathbb{R}^2$ à $\mathbb{C}$ via $(x, y) \leftrightarrow x + iy$. Dans toute cette partie, lorsque nous utilisons la notation $z = x + iy$, nous considérons que x et y sont des réels.

Une fonction complexe d'une variable complexe est une fonction $f : U \subset \mathbb{C} \to \mathbb{C}$ qui, à un nombre complexe $z \in U$, associe un nombre complexe $f(z) \in \mathbb{C}$.

Définition 22. La fonction $f : U \subset \mathbb{C} \to \mathbb{C}$ est *continue* en $z_0 \in U$ si $\lim\limits_{z \to z_0} f(z) = f(z_0)$.

La fonction f est $\mathbb{C}$-*dérivable* en z_0 si la limite :

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

existe. On la note alors $f'(z_0)$.

Définition 23. Si $U \subset \mathbb{C}$ est un ouvert et si $f : U \subset \mathbb{C} \to \mathbb{C}$ est $\mathbb{C}$-dérivable en tout point de U, on dit que f est une fonction *holomorphe* sur U.

Exercice 17.

Montrer que la fonction $f : z \mapsto z^2$ est holomorphe sur $\mathbb{C}$ et que pour tout $z \in \mathbb{C}$, on a $f'(z) = 2z$.

Solution. On a pour tout $z \in \mathbb{C} \setminus \{z_0\}$:

$$\frac{z^2 - z_0^2}{z - z_0} = z + z_0 \underset{z \to z_0}{\longrightarrow} 2z_0.$$

Il est facile de voir que les propriétés élémentaires des fonctions réelles d'une variable réelle qui sont dérivables passent aux fonctions complexes d'une variable complexe qui sont $\mathbb{C}$-dérivables. Voir le module IV.3 pour les démonstrations qui sont en tout point similaires.

Théorème 33. Soient f et g deux fonctions $\mathbb{C}$-dérivables au point z_0. Alors $f + g$, $f - g$, fg sont dérivables en z_0, ainsi que f/g si $g(z_0) \neq 0$. On a de plus les formules :

$$(f + g)'(z_0) = f'(z_0) + g'(z_0) \qquad (f - g)'(z_0) = f'(z_0) - g'(z_0)$$

$$(fg)'(z_0) = f'(z_0)g(z_0) + f(z_0)g'(z_0) \quad (f/g)'(z_0) = \frac{f'(z_0)g(z_0) - f(z_0)g'(z_0)}{\big(g(z_0)\big)^2}.$$

Il est naturel d'identifier la fonction $f : x + iy \mapsto f(x + iy)$ avec la fonction :

$$\boldsymbol{f} : (x, y) \mapsto \big(P(x, y), Q(x, y)\big)$$

avec :

$$P(x,y) := \mathrm{Re}\big(f(x+iy)\big) \quad \text{et} \quad Q(x,y) := \mathrm{Im}\big(f(x+iy)\big).$$

La suite de nombres complexes $(z_k := x_k + iy_k)$ converge vers $z_0 := x_0 + iy_0$ si, et seulement si, la suite $\big((x_k, y_k)\big)$ de points de $\mathbb{R}^2$ converge vers (x_0, y_0). On en déduit facilement que f est continue en z_0 si, et seulement si, $\boldsymbol{f}$ est continue en (x_0, y_0).
Nous allons maintenant étudier comment se traduit la $\mathbb{C}$-dérivabilité de f en terme de $\boldsymbol{f}$.

Théorème 34. La fonction f est $\mathbb{C}$-dérivable en $z_0 := x_0 + iy_0$ si, et seulement si,

1. $\boldsymbol{f}$ est différentiable en (x_0, y_0) et
2. la matrice jacobienne de $\boldsymbol{f}$ en (x_0, y_0) est de la forme :

$$J_{(x_0,y_0)}\boldsymbol{f} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \quad \text{avec} \quad (a,b) \in \mathbb{R}^2.$$

On a alors :

$$f'(z_0) = a + ib.$$

Démonstration. Supposons que f soit $\mathbb{C}$-dérivable en z_0. Cela se traduit par l'existence d'un nombre complexe λ tel que :

$$f(z_0 + h) = f(z_0) + \lambda h + |h|\,\varepsilon(h)$$

avec $\varepsilon(h)$ qui tend vers 0 lorsque $h \in \mathbb{C}$ tend vers 0. Si $\lambda = a + ib$ et $h = h_1 + ih_2$, on peut écrire λh sous la forme :

$$\lambda h = (a+ib)(h_1 + ih_2) = (ah_1 - bh_2) + i(ah_2 + bh_1).$$

La $\mathbb{C}$-dérivabilité de f en z_0 s'écrit en terme de $\boldsymbol{f}$ sous la forme :

$$\boldsymbol{f}(x_0 + h_1, y_0 + h_2) = \boldsymbol{f}(x_0, y_0) + (ah_1 - bh_2, ah_2 + bh_1) + \big\|(h_1, h_2)\big\|\,\boldsymbol{\varepsilon}(h_1, h_2)$$

avec $\lim\limits_{(0,0)} \boldsymbol{\varepsilon} = (0,0)$. Or, l'application :

$$L : (h_1, h_2) \mapsto (ah_1 - bh_2, ah_2 + bh_1)$$

est une application linéaire dont la matrice dans la base canonique de $\mathbb{R}^2$ est :

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

On voit donc que $\boldsymbol{f}$ est différentiable en (x_0, y_0) et que la matrice jacobienne de $\boldsymbol{f}$ en (x_0, y_0) est :

$$J_{(x_0,y_0)}\boldsymbol{f} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

avec $a := \mathrm{Re}\big(f'(z_0)\big)$ et $b := \mathrm{Im}\big(f'(z_0)\big)$.
Réciproquement, supposons que $\boldsymbol{f}$ est différentiable en (x_0, y_0) et que :

$$J_{(x_0,y_0)}\boldsymbol{f} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

Alors :

$$\boldsymbol{f}(x_0+h_1, y_0+h_2) = \boldsymbol{f}(x_0,y_0) + (ah_1 - bh_2, ah_2 + bh_1) + \big\|(h_1,h_2)\big\|\, \boldsymbol{\varepsilon}(h_1,h_2).$$

Donc :

$$f(z_0+h) = f(z_0) + (a+ib)h + |h|\,\varepsilon(h).$$

L'application f est donc $\mathbb{C}$-dérivable en z_0 de dérivée $f'(z_0) = a+ib$. ■

Lorsque nous identifions $\mathbb{C}$ à $\mathbb{R}^2$, la base $(1,i)$ est identifiée à la base orthonormée $(\vec{\imath},\vec{\jmath})$. Dans $\mathbb{R}^2$, une similitude centrée à l'origine $O = (0,0)$ de rapport $t \in \mathbb{R}$ et d'angle $\theta \in [0,2\pi[$ est la composée de l'homothétie H_t de centre O et de rapport t et d'une rotation R_θ de centre O et d'angle $\theta \in [0,2\pi[$.

Exercice 18.

Déterminer les matrices de H_t et R_θ dans la base canonique de $\mathbb{R}^2$.
Vérifier que $H_t \circ R_\theta = R_\theta \circ H_t$ et déterminer la matrice de cette similitude dans la base canonique de $\mathbb{R}^2$.

Solution. Comme $H_t(1) = t$ et $H_t(i) = it$, la matrice de H_t est $A := \begin{pmatrix} t & 0 \\ 0 & t \end{pmatrix}$.
Comme $R_\theta(1) = e^{i\theta} = \cos\theta + i\sin\theta$ et $R_\theta(i) = ie^{i\theta} = -\sin\theta + i\cos\theta$, la matrice de R_θ est $B := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. On a $AB = BA = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ avec $a := t\cos\theta$ et $b := t\sin\theta$.

Ce que nous avons utilisé dans la preuve précédente, c'est que, pour $\lambda \in \mathbb{C}^*$, l'application :

$$\begin{array}{rccc} L: & \mathbb{C} & \longrightarrow & \mathbb{C} \\ & h & \longmapsto & \lambda h \end{array}$$

est une application linéaire réelle particulière : une similitude centrée à l'origine, de rapport $|\lambda|$ et d'angle $\mathrm{Arg}(\lambda)$.

Corollaire 35 (Conditions de Cauchy-Riemann).
Si $f : (x+iy) \mapsto P(x,y) + iQ(x,y)$ est $\mathbb{C}$-dérivable en $z_0 := x_0 + iy_0$, alors :

$$\begin{cases} \dfrac{\partial P}{\partial x}(x_0,y_0) = \dfrac{\partial Q}{\partial y}(x_0,y_0) \\ \dfrac{\partial P}{\partial y}(x_0,y_0) = -\dfrac{\partial Q}{\partial x}(x_0,y_0) \end{cases}$$

et :

$$f'(z_0) = \frac{\partial P}{\partial x}(x_0,y_0) - i\frac{\partial P}{\partial y}(x_0,y_0) = \frac{\partial Q}{\partial y}(x_0,y_0) + i\frac{\partial Q}{\partial x}(x_0,y_0).$$

Démonstration. La matrice jacobienne de l'application $\boldsymbol{f}$ en (x_0,y_0) est :

$$J_{(x_0,y_0)}\boldsymbol{f} = \begin{pmatrix} \dfrac{\partial P}{\partial x} & \dfrac{\partial P}{\partial y} \\ \dfrac{\partial Q}{\partial x} & \dfrac{\partial Q}{\partial y} \end{pmatrix}(x_0,y_0).$$

Par conséquent :

$$J_{(x_0,y_0)}\boldsymbol{f} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \quad \text{avec} \quad f'(z_0) = a + ib$$

si, et seulement si :

$$\begin{cases} \dfrac{\partial P}{\partial x}(x_0, y_0) = \dfrac{\partial Q}{\partial y}(x_0, y_0) \\ \dfrac{\partial P}{\partial y}(x_0, y_0) = -\dfrac{\partial Q}{\partial x}(x_0, y_0) \end{cases}$$

avec :

$$f'(z_0) = \frac{\partial P}{\partial x}(x_0, y_0) - i\frac{\partial P}{\partial y}(x_0, y_0) = \frac{\partial Q}{\partial y}(x_0, y_0) + i\frac{\partial Q}{\partial x}(x_0, y_0).$$

■

Exercice 19.

Montrer que l'application $f : \mathbb{C} \to \mathbb{C}$ définie par $z \mapsto \mathrm{Re}(z)$ n'est holomorphe nulle part.

Solution. On a $f(x + iy) = x$, donc :

$$P(x, y) = x \quad \text{et} \quad Q(x, y) = 0.$$

Par conséquent, pour tout $z_0 := x_0 + iy_0 \in \mathbb{C}$:

$$\frac{\partial P}{\partial x}(x_0, y_0) = 1 \neq 0 = \frac{\partial Q}{\partial y}(x_0, y_0).$$

Ce n'est pas parce que les conditions de Cauchy-Riemann sont vérifiées en un point $(x_0, y_0) \in \mathbb{R}^2$ que la fonction f est $\mathbb{C}$-dérivable en $z_0 := x_0 + iy_0$. En effet, l'existence de dérivées partielles ne suffit pas à conclure à la différentiabilité de la fonction $\boldsymbol{f}$. Cependant, si les conditions de Cauchy-Riemann sont vérifiées sur un ouvert, et si les dérivées partielles de P et Q sont des fonctions continues, on peut conclure, grâce au théorème 27 de la page 838, que $\boldsymbol{f}$ est différentiable sur cet ouvert, et donc que f est holomorphe.

Exercice 20.

Montrer que la fonction $f : \mathbb{C} \to \mathbb{C}$ définie par :

$$f(x + iy) := e^{x+iy} = e^x(\cos y + i \sin y)$$

est holomorphe sur $\mathbb{C}$.

Solution. On a :

$$P(x, y) = e^x \cos y \quad \text{et} \quad Q(x, y) = e^x \sin y.$$

Les dérivées partielles :

$$\frac{\partial P}{\partial x}(x, y) = e^x \cos y = \frac{\partial Q}{\partial y}(x, y) \quad \text{et} \quad \frac{\partial P}{\partial y}(x, y) = -e^x \sin y = -\frac{\partial Q}{\partial x}(x, y)$$

sont des fonctions continues sur $\mathbb{C}$. La dérivée f' vérifie :

$$f'(x + iy) = \frac{\partial P}{\partial x}(x, y) - i\frac{\partial P}{\partial y}(x, y) = f(x + iy).$$

Donc $f'(z) = f(z)$.

Exercice type corrigé

IV.7.0 Étudier la continuité en $(0,0)$ de la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par :

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

Solution. Pour que la fonction f soit continue en $(0,0)$ il faut et il suffit que :

$$f(x,y) \underset{(x,y)\to(0,0)}{\longrightarrow} f(0,0) := 0.$$

Nous allons montrer que la fonction f n'est pas continue en $(0,0)$.
Si x et y tendent vers 0, alors xy tend vers 0 et x^2+y^2 tend vers 0. Nous sommes donc confrontés à une forme indéterminée du type $0/0$. Considérons la suite $(\boldsymbol{x}_k)$ définie par $\boldsymbol{x}_k := (1/k, 1/k)$. Quand $k \to +\infty$, $\boldsymbol{x}_k \to (0,0)$. De plus,

$$f(\boldsymbol{x}_k) = \frac{\sin(1/k^2)}{2/k^2} = \frac{1}{2}\frac{\sin(1/k^2)}{1/k^2}\cdot$$

Or,

$$\lim_{x\to 0} \frac{\sin(x)}{x} = \lim_{x\to 0} \frac{\sin(x)-\sin(0)}{x-0} = \cos(0) = 1$$

(la limite du taux d'accroissement est égale à la dérivée en 0). Par conséquent,

$$\lim_{k\to+\infty} f(\boldsymbol{x}_k) = \frac{1}{2} \neq f(0,0).$$

La fonction f n'est pas continue en $(0,0)$.
Notons que $f(x,0) = f(y,0) = 0$.

Il n'est donc pas possible de prolonger continûment la fonction $(x,y) \mapsto \dfrac{\sin(xy)}{x^2+y^2}$ en $(0,0)$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

IV.7.1 Parmi les ensembles suivants, lesquels sont ouverts ? fermés ? compacts ?

1. $[0,1[$ dans $\mathbb{R}$,
2. $]0,1[$ dans $\mathbb{C}$,
3. $[0,1]$ dans $\mathbb{C}$,
4. $\{(x,y) \in \mathbb{R}^2 \mid x > y\}$ dans $\mathbb{R}^2$,
5. $\{(x,y) \in \mathbb{R}^2 \mid x \geqslant y\}$ dans $\mathbb{R}^2$.

IV.7.2 Étudier la limite en $(0,0)$ des fonctions suivantes :

1. $f(x,y) := \dfrac{\ln(x+y)}{\sqrt{x^2+y^2}}$,
2. $g(x,y) := (|x|+|y|)\ln(x^2+y^4)$,
3. $h(x,y) := \dfrac{x^2-y^2}{x^2+y^2}\cdot$

IV.7.3 Montrer que la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par :

$$f(x,y) := (x^2 - y^3)e^{-x^2+y^3}$$

est continue et tend vers 0 quand $\|(x,y)\| \to +\infty$. En déduire que f est bornée et atteint ses bornes dans $\mathbb{R}^2$.

IV.7.4 Déterminer les dérivées partielles $\dfrac{\partial f}{\partial x}$ et $\dfrac{\partial f}{\partial y}$ pour les fonctions :

1. $f : (x,y) \mapsto e^x \sin(x^2+y)$,
2. $g : (x,y) \mapsto \dfrac{xy^2}{\sqrt{x+y^2}}$,
3. $h : (x,y) \mapsto \sin(x^2-y) + \sin^2(xy)$.

IV.7.5 Déterminer les dérivées partielles $\dfrac{\partial \boldsymbol{f}}{\partial x}$ et $\dfrac{\partial \boldsymbol{f}}{\partial y}$ pour :

$$\boldsymbol{f} : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x^2 y \\ e^x \sin(y) \\ \ln(x^2+y^2) \end{pmatrix}.$$

IV.7.6 Quelle est la différentielle de l'application $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ donnée par $\boldsymbol{f}(\boldsymbol{x}) := \|\boldsymbol{x}\|^2 \boldsymbol{x}$?

IV.7.7 Sur quel ensemble les applications suivantes sont-elles différentiables ?

1. $f : (x,y) \mapsto \ln\|(x,y)\|$,
2. $g : (x,y) \mapsto |x+y|$.

IV.7.8 Déterminer la matrice jacobienne $J_{(x,y)}f$ de l'application $f : \mathbb{R}^2 \to \mathbb{R}^2$ définie par :

$$f(x,y) := (x^2 + 3xy^2, ye^x).$$

IV.7.9 Montrer que l'application $f : \mathbb{R}^2 \to \mathbb{R}^2$ définie par :

$$f(x,y) := \big(\cos(xy), \sin(x+y)\big)$$

est continûment différentiable. Déterminer sa différentielle.

IV.7.10 La fonction :

$$\begin{array}{rccc} f : & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & \dfrac{xy^3}{x^4+y^2} \end{array}$$

prolongée par 0 à l'origine est-elle continue, différentiable, continûment différentiable ?

IV.7.11 On suppose que $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ est une fonction dont toutes les dérivées partielles sont identiquement nulles sur $\mathbb{R}^n$. Montrer que $\boldsymbol{f}$ est constante.

IV.7.12 On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par :

$$f(x,y) := \arctan(x) + \arctan(y) - \arctan\frac{x+y}{1-xy}.$$

Montrer que f est continûment différentiable sur son ensemble de définition et déterminer sa différentielle. En déduire les valeurs de f.

IV.7.13 On dit qu'une application $f : \mathbb{R}^2 \to \mathbb{R}$ est homogène de degré $k \geqslant 1$ si, et seulement si, pour tout $\boldsymbol{x} \in \mathbb{R}^2$ et tout $\lambda \in \mathbb{R}$, on a $f(\lambda \boldsymbol{x}) = \lambda^k f(\boldsymbol{x})$. Montrer qu'alors pour tout $\boldsymbol{x} \in \mathbb{R}^2$:

$$D_{\boldsymbol{x}} f(\boldsymbol{x}) = k f(\boldsymbol{x}).$$

IV.7.14 On considère l'application f de l'ensemble des matrices 2×2 (que vous pouvez identifier à $\mathbb{R}^4$ si vous le souhaitez) à valeurs dans $\mathbb{R}$ qui à une matrice M associe son déterminant :

$$f\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

1. Déterminer $\dfrac{\partial f}{\partial a}$, $\dfrac{\partial f}{\partial b}$, $\dfrac{\partial f}{\partial c}$ et $\dfrac{\partial f}{\partial d}$.
2. Montrer que f est différentiable.
3. Montrer que la différentielle de f en I_2 est l'application $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto a + d$.

IV.7.15 On considère l'application f de l'ensemble des matrices 3×3 (que vous pouvez identifier à $\mathbb{R}^9$ si vous le souhaitez) à valeurs dans $\mathbb{R}$ qui à une matrice M associe son déterminant.
On note tr l'application qui à une matrice carrée associe la somme de ses coefficients diagonaux.
On pourra utiliser la notion de polynôme caractéristique et, en particulier, la proposition 57 de la page 371.

1. Montrer que f est différentiable.
2. Étant donnée une matrice M, montrer que l'application :

$$\begin{array}{rccc} P : & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & h & \longmapsto & f(\mathrm{I}_3 + hM) \end{array}$$

   est un polynôme en h. Déterminer le coefficient constant et le coefficient de h.
3. En déduire que :

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f(\mathrm{I}_3 + hM) - f(\mathrm{I}_3)}{h} = \mathrm{tr}(M).$$

4. Déterminer la différentielle de f en I_3.

IV.7.16 Sur quels ensembles les fonctions suivantes sont-elles holomorphes ?

1. $f : z \mapsto |z|$,
2. $g : z \mapsto \dfrac{z+1}{z^2+2}$,
3. $h : z \mapsto \ln|z| + i\,\mathrm{Arg}(z)$.

IV.7.17 Montrer que l'inverse d'une matrice de la forme $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ avec $a \in \mathbb{R}$ et $b \in \mathbb{R}$ non nuls est de la même forme.

IV.7.18 Soit $f(x,y)$ un polynôme en les deux variables x et y. Montrer que c'est un polynôme en $z := x + iy$ si, et seulement si, f vérifie les conditions de Cauchy-Riemann.
Donner un énoncé analogue pour les polynômes en $\bar{z} := x - iy$.

Approximation IV.8

1 Introduction

Si f est une fonction réelle *continue* au point a, la fonction $f - f(a)$ tend vers zéro lorsque x tend vers a. Si nous posons $\varepsilon_0 = f - f(a)$, alors :

$$f(x) = f(a) + \varepsilon_0(x) \quad \text{où} \quad \lim_{x \to a} \varepsilon_0(x) = 0. \tag{1}$$

Cette relation nous dit que $f(x)$ est « proche » du réel $f(a)$ lorsque x est « proche » de a, la différence étant $\varepsilon_0(x)$ qui est « proche » de 0. On notera que nous n'avons pas assez d'informations sur f pour préciser « combien proche ».
Nous pouvons être plus précis si f est *dérivable* en a. Dans ce cas :

$$f(x) = f(a) + (x-a)f'(a) + (x-a)\varepsilon_1(x) \qquad \text{où} \qquad \lim_{x \to a} \varepsilon_1(x) = 0. \tag{2}$$

Les deux formules (1) et (2) sont liées par la relation $\varepsilon_0 = (x-a)f'(a) + (x-a)\varepsilon_1$, mais la seconde est plus précise, car le deuxième terme est explicite (si l'on connaît la dérivée) et, si $f'(a) \neq 0$, le troisième tend « plus rapidement vers zéro » que lui. On peut interpréter (1) comme l'approximation de f par une fonction constante et (2) comme l'approximation de f par une fonction polynomiale de degré (au plus) 1, c'est-à-dire une fonction affine.
Géométriquement, ces deux formules donnent des informations sur le graphe de f : la première dit que, si x est « proche » de a, les points du graphe sont « proches » du point $\big(a, f(a)\big)$. La seconde indique que le graphe est « proche » de celui de l'application affine $x \mapsto f(a) + (x-a)f'(a)$, qui est la *tangente* au graphe de f au point $\big(a, f(a)\big)$.
Plus généralement, si la fonction f est plusieurs fois dérivable en a, on peut « l'approcher » par un polynôme en utilisant une formule de Taylor (*cf.* à la page 652, la partie 2.6 du module IV.2 *Fonctions réelles*). Par exemple, si $f := \sin$, on peut la comparer aux polynômes de Taylor en 0 :

$$P_1 := x, \quad P_3 := x - \frac{x^3}{6}, \quad P_5 := x - \frac{x^3}{6} + \frac{x^5}{120}$$

$$P_7 := x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040}, \quad P_9 := x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \frac{x^9}{362880},$$

en utilisant les formules de Taylor-Young aux ordres 1, 3, 5, 7, 9.

La figure IV.8.1 montre les graphes de $\sin$, et P_i, pour $i = 1, 3, 5, 7, 9$. Pour $i = 1, 5, 9$ (resp. 3, 7) les graphes sont au-dessus (resp. au-dessous) de celui de $\sin$. On peut deviner que l'utilisation d'un polynôme de Taylor d'ordre plus élevé « améliore l'approximation ».

FIGURE IV.8.1 – Approximations de la fonction $\sin$ par ses polynômes de Taylor

On peut quantifier la différence entre une fonction et l'un de ses polynômes de Taylor en utilisant les formules de Taylor déjà établies dans le module IV.2. Dans ce module, nous les compléterons par une nouvelle formule, la formule de Taylor *avec reste intégral* (théorème 1 de la page ci-contre).

Ces considérations nous amènent à deux questions qui seront traitées dans ce module :

1. Comment comparer plusieurs fonctions différentes à partir de leurs approximations ? Il y a différentes façons de le faire, nous le verrons dans la section 3.
2. Comment faire des calculs d'approximation de fonctions à partir d'autres plus simples (par exemple des polynômes) ? Et comment en déduire divers résultats, en particulier des calculs de limites ? C'est le but de la notion de *développement limité*, que nous introduirons dans la section 4.

Nous terminerons le module par l'étude de méthodes de calcul approché d'intégrales. Ces méthodes sont basées sur l'approximation d'une fonction (suffisamment régulière) par un polynôme d'interpolation (c'est-à-dire un polynôme coïncidant avec la fonction en un certain nombre de points) et l'estimation de l'erreur de calcul est souvent basée sur des formules de Taylor.

Dans ce module tous les intervalles réels sont supposés contenir au moins deux points (ou, ce qui est équivalent, être d'intérieur non vide).

2 Formules de Taylor

Nous avons vu dans la partie 2.6 du module IV.2 que, lorsque f est une fonction réelle plusieurs fois dérivable, on peut l'approximer par un *polynôme* grâce à des formules de Taylor, qui sont toutes des variantes de la même formule initiale dont le *reste* est exprimé différemment. Dans cette partie, nous allons montrer une nouvelle formule de Taylor, dite *formule de Taylor avec reste intégral*, en utilisant la théorie de l'intégration (*cf.* le module IV.5 *Introduction à l'intégration*). Les méthodes très élémentaires que nous avons utilisées dans le

module IV.2 pour obtenir des formules de Taylor ne s'étendent pas aux fonctions à valeurs vectorielles. Nous verrons plus loin que, par contre, la formule de Taylor avec reste intégral reste vraie dans ce cas et qu'elle permet d'établir l'inégalité de Taylor-Lagange et un théorème de Taylor-Young pour les fonctions vectorielles.
On notera la différence entre les formules *globales* (faisant intervenir le comportement de la fonction sur un intervalle) et les formules *locales* (faisant intervenir le comportement de la fonction au voisinage d'un point) : les formules de Taylor avec reste intégral, l'inégalité de Taylor-Lagrange et la formule de Taylor-Lagrange sont des formules globales, tandis que la formule de Taylor-Young est une formule locale.

2.1 Formule de Taylor avec reste intégral

Théorème 1 (Formule de Taylor avec reste intégral). Soient $n \in \mathbb{N}$ et f une fonction $(n+1)$ fois *continûment dérivable* sur un intervalle I. Alors, pour tout point $a \in I$, on a :

$$\forall x \in I\,,\ f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{n!}f^{(n)}(a) + \int_a^x \frac{(x-t)^n}{n!}f^{(n+1)}(t)\,\mathrm{d}t.$$

On dit que l'on a une formule de Taylor *à l'ordre* n avec reste intégral. Son *reste intégral* à l'ordre n est $\displaystyle\int_a^x \frac{(x-t)^n}{n!}f^{(n+1)}(t)\,\mathrm{d}t$.

Démonstration. Nous faisons la démonstration par récurrence sur n. Pour $n := 0$, la formule affirme que :

$$f(x) = f(a) + \int_a^x f'(t)\,\mathrm{d}t$$

ce qui est vrai d'après le théorème fondamental de l'intégration.

Supposons donc la formule vraie à un certain ordre $n-1 \geqslant 0$, et montrons-la à l'ordre n, en supposant la fonction $(n+1)$ fois continûment dérivable. On utilise la formule d'intégration par parties (*cf.* le théorème 24 de la page 761). La dérivée de $f^{(n)}$ est $f^{(n+1)}$, et une primitive de $\frac{(x-t)^{n-1}}{(n-1)!}$ (par rapport à t) est $-\frac{(x-t)^n}{n!}$:

$$\begin{aligned}\int_a^x \frac{(x-t)^{n-1}}{(n-1)!}f^{(n)}(t)\,\mathrm{d}t &= \left[-\frac{(x-t)^n}{n!}f^{(n)}(t)\right]_a^x + \int_a^x \frac{(x-t)^n}{n!}f^{(n+1)}(t)\,\mathrm{d}t \\ &= \frac{(x-a)^n}{n!}f^{(n)}(a) + \int_a^x \frac{(x-t)^n}{n!}f^{(n+1)}(t)\,\mathrm{d}t.\end{aligned}$$

Il suffit de remplacer cette valeur du reste dans la formule à l'ordre $n-1$, pour obtenir exactement la formule à l'ordre n. ∎

Remarque. Le changement de variable affine $t := a + s\,(x-a)$ permet de transformer le reste intégral en une intégrale sur $[0,1]$. On obtient ainsi, pour a et x dans I :

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{n!}f^{(n)}(a) + \frac{(x-a)^{n+1}}{n!}\int_0^1 (1-s)^n f^{(n+1)}\big(a + s\,(x-a)\big)\,\mathrm{d}s. \quad (3)$$

2.2 Formules de Taylor pour les fonctions vectorielles

Nous allons voir que certaines formules de Taylor s'étendent aux fonctions vectorielles : la formule avec reste intégral, l'inégalité de Taylor-Lagrange, la formule de Taylor-Young. Par contre la formule de Taylor-Lagrange ne se généralise pas.

Dans toute cette partie $m \in \mathbb{N}^*$. Nous utiliserons la notion d'intégrale d'une fonction à valeurs dans $\mathbb{R}^m$ et les propriétés de l'intégrale. Nous noterons $\|\mathbf{u}\|$ la *norme* de $\mathbf{u} = (u_1, \ldots, u_m) \in \mathbb{R}^m$.

Théorème 2 (Formule de Taylor avec reste intégral).
Soient $n \in \mathbb{N}$, $I \subset \mathbb{R}$ un intervalle et $\mathbf{f} : I \to \mathbb{R}^m$ une fonction $(n+1)$ fois *continûment dérivable* sur I. Alors, pour tout point $a \in I$, on a :

$$\forall x \in I\,,\ \mathbf{f}(x) = \mathbf{f}(a) + (x-a)\mathbf{f}'(a) + \frac{(x-a)^2}{2!}\mathbf{f}''(a) + \cdots + \frac{(x-a)^n}{n!}\mathbf{f}^{(n)}(a) + \int_a^x \frac{(x-t)^n}{n!}\mathbf{f}^{(n+1)}(t)\,\mathrm{d}t. \quad (4)$$

Démonstration. La preuve est, *mutatis mutandis*, la même que dans le cas $m = 1$. ■

Le second membre de (4) se décompose en deux parties ; la première est :

$$\mathbf{P}_{a,n}(x) := \mathbf{f}(a) + (x-a)\mathbf{f}'(a) + \frac{(x-a)^2}{2!}\mathbf{f}''(a) + \cdots + \frac{(x-a)^n}{n!}\mathbf{f}^{(n)}(a).$$

On l'appelle la *partie régulière* de f au point a à l'ordre n. La seconde partie est appelée *le reste intégral*.

Comme dans le cas $m = 1$, on peut écrire le reste intégral sous une autre forme :

$$\boldsymbol{f}(x) = \boldsymbol{f}(a) + (x-a)\boldsymbol{f}'(a) + \frac{(x-a)^2}{2!}\boldsymbol{f}''(a) + \cdots + \frac{(x-a)^n}{n!}\boldsymbol{f}^{(n)}(a) + \frac{(x-a)^{n+1}}{n!}\int_0^1 (1-s)^n \boldsymbol{f}^{(n+1)}\big(a + s\,(x-a)\big)\,\mathrm{d}s. \quad (5)$$

La preuve de l'inégalité de Taylor-Lagrange est différente de celle donnée pour $m = 1$. On la déduit de la formule intégrale.

Théorème 3 (Inégalité de Taylor-Lagrange). Soient $n \in \mathbb{N}$, $I \subset \mathbb{R}$ un intervalle et $\mathbf{f} : I \to \mathbb{R}^m$ une fonction $(n+1)$ fois continûment dérivable sur I, dont la $(n+1)^{ème}$ dérivée est bornée par $M \geqslant 0$. Alors pour tout point $a \in I$, on a :

$$\forall x \in I\,,\ \left\| \mathbf{f}(x) - \sum_{k=0}^{n} \frac{(x-a)^k}{k!} \mathbf{f}^{(k)}(a) \right\| \leqslant M\, \frac{|x-a|^{n+1}}{(n+1)!}. \tag{6}$$

Démonstration.
Il suffit de majorer la norme du reste intégral. On utilise l'inégalité

$$\left\| \int_a^b \boldsymbol{g}(x)\,\mathrm{d}x \right\| \leqslant \int_a^b \|\boldsymbol{g}(x)\|\,\mathrm{d}x$$

du théorème 26 de la page 797. Pour éviter de discuter suivant le signe de $x-a$, il est plus simple d'utiliser la formule de Taylor sous la forme (5) de la page précédente. On obtient :

$$\begin{aligned}
\frac{(x-a)^{n+1}}{n!} & \left\| \int_0^1 (1-s)^n \boldsymbol{f}^{(n+1)}\big(a + s\,(x-a)\big)\,\mathrm{d}s \right\| \\
& \leqslant \frac{|x-a|^{n+1}}{n!} \int_0^1 (1-s)^n \left\| \boldsymbol{f}^{(n+1)}\big(a + s\,(x-a)\big) \right\| \mathrm{d}s \\
& \leqslant \frac{|x-a|^{n+1}}{n!} \int_0^1 M\,(1-s)^n\,\mathrm{d}s = M\,\frac{|x-a|^{n+1}}{(n+1)!},
\end{aligned}$$

ce qui donne l'inégalité annoncée. ∎

Théorème 4 (Formule de Taylor-Young). Soient $n \in \mathbb{N}$, $I \subset \mathbb{R}$ un intervalle et $\mathbf{f} : I \to \mathbb{R}^m$ une fonction n fois continûment dérivable sur I. Alors :

$$\boldsymbol{f}(x) = \sum_{k=0}^{n} \frac{(x-a)^k}{k!} \boldsymbol{f}^{(k)}(a) + (x-a)^n \varepsilon_n(x) \quad \text{où} \quad \lim_{x \to a} \varepsilon_n(x) = 0. \tag{7}$$

Démonstration. Nous donnons deux preuves. La première consiste à déduire le théorème pour les fonctions vectorielles du théorème pour les fonctions réelles. La seconde utilise l'inégalité de Taylor-Lagrange pour les fonctions vectorielles. Elle peut être omise en première lecture.

Première preuve. On note $\boldsymbol{f} := (f_1, \ldots, f_m)$. On applique la formule de Taylor-Young (*cf.* le théorème 54 de la page 655) aux fonctions réelles $f_1, \ldots, f_m$. Pour tout $j \in [\![1, m]\!]$:

$$f_j(x) = \sum_{k=0}^{n} \frac{(x-a)^k}{k!} f_j^{(k)}(a) + (x-a)^n \varepsilon_{j,n}(x) \qquad \text{où} \qquad \lim_{x \to a} \varepsilon_{j,n}(x) = 0.$$

On pose $\varepsilon_n := (\varepsilon_{1,n}, \ldots, \varepsilon_{m,n})$. On a $\|\varepsilon_n\|^2 = \varepsilon_{1,n}^2 + \cdots + \varepsilon_{m,n}^2$, donc $\|\varepsilon_n\| \leqslant \sqrt{n} \max\limits_{j \in [\![1,m]\!]} |\varepsilon_{j,n}|$. Par suite $\lim\limits_{x \to a} \varepsilon_n(x) = 0$.

Deuxième preuve. Le résultat est immédiat si $n := 0$ (*cf.* la relation (1) de la page 855) ; nous supposerons donc $n \geqslant 1$ dans la suite.

Pour que (7) soit vraie, il suffit de montrer que la limite de $\varepsilon_n(x)$ est nulle lorsque x tend vers a, la fonction $\varepsilon_n(x)$ étant définie, pour $x \neq a$, par :

$$\varepsilon_n(x) = (x-a)^{-n}\left(\boldsymbol{f}(x) - \sum_{k=0}^{n} \frac{(x-a)^k}{k!}\boldsymbol{f}^{(k)}(a)\right) = \frac{\boldsymbol{f}(x) - \boldsymbol{P}_{a,n}(x)}{(x-a)^n}\,.$$

D'après sa définition, $\boldsymbol{P}_{a,n}(x) = \boldsymbol{P}_{a,n-1}(x) + \dfrac{(x-a)^n}{n!}\boldsymbol{f}^{(n)}(a)$.

- Supposons tout d'abord $\boldsymbol{f}^{(n)}(a) = 0$. Pour tout $x \in I$, posons $M_x := \sup\limits_{J_x} \left\|\boldsymbol{f}^{(n)}\right\|$, avec $J_x = [a, x]$ ou $J_x = [x, a]$ selon les positions relatives de x et a. Par continuité de $\boldsymbol{f}^{(n)}$ en a, on a $\lim\limits_{x\to a} M_x = 0$. En effet, pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que :

  $$\forall t \in I\,,\ |t-a| \leqslant \eta \Rightarrow \left\|\boldsymbol{f}^{(n)}(t) - \boldsymbol{f}^{(n)}(a)\right\| \leqslant \varepsilon$$

  et alors, pour $x \in I$ tel que $|x-a| \leqslant \eta$, on a $M_x \leqslant \varepsilon$.

  L'hypothèse $\boldsymbol{f}^{(n)}(a) = 0$ donne d'autre part $\boldsymbol{P}_{a,n} = \boldsymbol{P}_{a,n-1}$ et donc l'inégalité de Taylor-Lagrange, qui s'écrit :

  $$\forall x \in I \setminus \{a\}\,,\ \left\|\boldsymbol{f}(x) - \boldsymbol{P}_{a,n-1}(x)\right\| \leqslant \frac{|x-a|^n}{n!}\,M_x$$

  permet de conclure $\left\|\varepsilon_n(x)\right\| \leqslant M_x/n! \underset{x\to a}{\longrightarrow} 0$.

- Dans le cas général, on considère $\boldsymbol{g} : x \mapsto \boldsymbol{f}(x) - \frac{(x-a)^n}{n!}\,\boldsymbol{f}^{(n)}(a)$.

  C'est une fonction n fois continûment dérivable sur I vérifiant les relations $\boldsymbol{g}^{(n)}(a) = 0$ et $\forall k \in [\![0, n-1]\!]\,,\ \boldsymbol{g}^{(k)}(a) = \boldsymbol{f}^{(k)}(a)$. On a alors, pour $x \neq a$:

  $$\varepsilon_n(x) = (x-a)^{-n}\left(\boldsymbol{g}(x) - \sum_{k=0}^{n} \frac{(x-a)^k}{k!}\boldsymbol{g}^{(k)}(a)\right) \underset{x\to a}{\longrightarrow} 0,$$

  d'après l'étude du premier cas. ■

Comme dans le cas $m = 1$ la formule de Taylor-Young reste valable si, au lieu de supposer que $\boldsymbol{f}$ est n fois dérivable sur I, on suppose seulement qu'elle est n fois dérivable au point a (*cf.* le théorème 55 de la page 656).

La preuve est similaire, *mutatis mutandis*. À la fin de cette preuve, il faut remplacer l'utilisation de l'inégalité des accroissements finis par sa version vectorielle (*cf.* le théorème 22 de la page 794).

3 Équivalents et notations de Landau

Dans cette partie et dans toute la suite de ce module on ne considère que des fonctions réelles.

Comme nous l'avons dit plus haut, les formules de Taylor nous disent qu'une fonction suffisamment dérivable est « proche » d'un polynôme. Pour exploiter cette information et en vue d'autres applications, il est utile d'introduire certaines notations.

Dans les sections suivantes, nous parlons uniquement de l'approximation de fonctions. En réalité les notations introduites s'appliquent aussi aux suites, considérées comme des fonctions de $\mathbb{N}$ vers $\mathbb{R}$. Dans ce cas, le point a est $+\infty$, le vocabulaire étant essentiellement le même.

Notation. Dans les divers énoncés, nous fixons un élément a qui peut être un réel ou $+\infty$ ou $-\infty$ et nous supposons que toutes les fonctions considérées sont définies au voisinage de a sur un même domaine $D \subset \mathbb{R}$.

Rappelons (*cf.* la définition 2 de la page 621) ce que cela signifie : une fonction f est définie au voisinage de a s'il existe une suite à valeurs dans $D \setminus \{a\}$ admettant a pour limite.

3.1 Équivalents

Soit g une fonction définie sur D au voisinage de a.
On dit qu'elle ne s'annule pas au voisinage de a, sauf peut-être en a, si, quitte à restreindre D, on a $g(x) \neq 0$ pour tout $x \in D \setminus \{a\}$.

Définition 1. Soient f et g deux fonctions définies sur D, *la seconde ne s'annulant pas au voisinage de* a, sauf peut-être en a. On dit que f et g sont *équivalentes* en a si :

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 1. \tag{8}$$

On note $f \sim_a g$ (ou $f \sim g$ si le point a est sous-entendu).

Deux fonctions f et g équivalentes en a peuvent être continues en ce point et alors $f(a) = g(a)$; cela s'étend aussitôt à des limites infinies. Mais il est possible que f n'ait pas de limite, auquel cas g n'en a pas non plus. Autrement dit, les équivalents permettent de trouver une limite ou de montrer qu'une limite n'existe pas.
La condition sur g de ne pas s'annuler au voisinage de a, sauf peut-être en a, peut s'exprimer sous la forme : il n'existe aucune suite (x_n) à valeurs dans $D_g \setminus \{a\}$, convergeant vers a, et telle que $g(x_n) = 0$ pour tout n. Cette condition est évidemment nécessaire pour g, sinon la limite dans (8) n'a pas de sens. Le lecteur vérifiera qu'elle est suffisante.
On peut noter qu'aucune fonction n'est équivalente à la fonction nulle, pas même cette dernière.
Notons que si f et g sont équivalentes en a, alors f ne peut pas s'annuler au voisinage de a, sauf éventuellement en a. En effet, le fait que f/g tende vers 1 entraîne que f ne s'annule pas au voisinage de a, sauf éventuellement en a. Nous utiliserons cette remarque pour prouver la symétrie dans la proposition 6 ci-dessous. Dans la suite, quand nous écrirons $f \sim g$, il sera toujours sous-entendu que f et g ne peuvent pas s'annuler au voisinage de a, sauf éventuellement en a

S'il existe une suite (x_n) à valeurs dans $D_f \setminus \{a\}$, convergeant vers a, et telle que $f(x_n) = 0$ pour tout n, alors f n'est pas équivalente à f.
Par exemple, si $f : x \in \mathbb{R}^* \mapsto \sin(1/x)$, f n'est pas équivalente à f en 0.
Pour éviter ce problème on peut introduire une définition de l'équivalence *plus générale* que la nôtre. Soient f et g deux fonctions définies sur D. On dit alors que f et g sont *équivalentes* en a s'il existe h, définie sur D, tendant vers 1 en a et telle que $f = hg$ au voisinage de a. On trouve cette définition dans plusieurs ouvrages. Avec cette définition les fonctions $x \sin(1/x)$ et $\sin x \sin(1/x)$ sont équivalentes en 0 mais elles ne le sont pas avec la nôtre.

Proposition 5. Soient f et g deux fonctions équivalentes en a. On suppose que g admet la limite ℓ (finie ou infinie) au point a, alors f admet également la limite ℓ au point a.

Démonstration. On a $f = hg$, la fonction $h = f/g$ admettant la limite 1 au point a. On en déduit le résultat en utilisant la proposition 11 de la page 626 si $\ell \in \mathbb{R}$ et la proposition 12 de la page 626 si $\ell = \pm\infty$. ■

Proposition 6. La relation « être équivalent(e) au point a » est une relation d'équivalence, c'est-à-dire qu'elle est :

1. réflexive : $f \sim f$;
2. symétrique : $f \sim g$ implique $g \sim f$;
3. transitive : $f \sim g$ et $g \sim h$ impliquent $f \sim h$.

Ces propriétés s'entendent pour les fonctions f, g, h définies sur D et ne s'annulant pas au voisinage de a, sauf peut-être en a.

Démonstration. La réflexivité est évidente (elle sous-entend que la fonction ne s'annule pas sauf peut-être en a). Pour la symétrie, on note que g/f est, au voisinage de a, l'inverse de f/g et donc tend vers $1/1 = 1$ en a.

Pour la transitivité, on écrit simplement que : $\frac{f(x)}{h(x)} = \frac{f(x)}{g(x)}\,\frac{g(x)}{h(x)}$ et l'on passe à la limite. ■

Il résulte immédiatement de la définition 1 que, si f a une limite *non nulle* ℓ au point a, alors f est équivalente à la fonction constante ℓ. Par suite, si deux fonctions f et g ont *chacune* une limite au point a, et si *l'une* des deux limites est non nulle, alors f et g sont équivalentes si, et seulement si, les deux limites sont égales. La notion d'équivalent est donc surtout intéressante pour des fonctions qui s'annulent ou qui tendent vers l'infini en a.
En particulier, on a le théorème suivant :

Théorème 7. Soient f une fonction n fois continûment dérivable sur un intervalle I, et $a \in I$. On note $f^{(0)}(a) := f(a)$. Si l'une des dérivées successives de f est non nulle en a, alors :

$$f(x) \sim_a \frac{(x-a)^p}{p!} f^{(p)}(a) \tag{9}$$

où $p \in [\![0, n]\!]$ est le plus petit entier tel que $f^{(p)}(a) \neq 0$.

Démonstration.
On écrit le développement de Taylor-Young pour f à l'ordre $p \leqslant n$ (théorème 54 de la page 655), en notant que, par hypothèse, $f^{(k)}(a) = 0$ si $k < p$, ce qui élimine les premiers termes dans le polynôme d'approximation. En particulier, on peut factoriser par $(x-a)^p$:

$$f(x) = (x-a)^p \left(\frac{1}{p!} f^{(p)}(a) + \varepsilon_p(x) \right).$$

Posons $g(x) = (x-a)^p f^{(p)}(a)/p!$, qui n'est pas nulle pour $x \neq a$, puisque $f^{(p)}(a) \neq 0$. On peut donc diviser l'égalité précédente par $g(x)$:

$$\frac{f(x)}{g(x)} = 1 + \frac{p!}{f^{(p)}(a)}\varepsilon_p(x).$$

Le second terme du second membre tend vers zéro lorsque x tend vers a. ■

Exemple. En utilisant le théorème 7, on obtient $\mathrm{e}^x - 1 \sim x$ et $\ln(1+x) \sim x$ au voisinage de 0. On en déduit que $\ln u \sim u - 1$ quand u tend vers 1.

Toujours en utilisant le théorème 7, le lecteur vérifiera facilement les autres équivalences figurant dans la table IV.8.1.

$\mathrm{e}^x - 1 \sim x$	$\ln(1+x) \sim x$
$\sin x \sim x$	$\tan x \sim x$
$\arcsin x \sim x$	$\arctan x \sim x$
$\sinh x \sim x$	$\tanh x \sim x$
$\cos x - 1 \sim -\dfrac{x^2}{2}$	$\cosh x - 1 \sim \dfrac{x^2}{2}$
$\sin x - x \sim -\dfrac{x^3}{6}$	$\sinh x - x \sim \dfrac{x^3}{6}$
$(1+x)^\alpha - 1 \sim \alpha x$	$(\alpha \in \mathbb{R})$

TABLE IV.8.1 – Équivalents classiques en 0

On pourrait croire que l'on peut appliquer le théorème 7 à toute fonction indéfiniment dérivable non nulle. Mais ce n'est pas le cas, car une telle fonction peut avoir *toutes ses dérivées nulles en un point*, comme le montre l'exercice ci-après.

Exercice 1.

Soit f la fonction définie sur $\mathbb{R}$ par $f(0) = 0$ et $f(x) = \exp\left(-\frac{1}{x^2}\right)$ pour $x \neq 0$. Montrer que f est indéfiniment dérivable sur $\mathbb{R}$ (même en 0), et que $f^{(n)}(0) = 0$ pour tout $n \in \mathbb{N}$.

Solution. Il est élémentaire de vérifier que f est continue en 0. Il est facile de montrer par récurrence sur $n \in \mathbb{N}^*$ que f est n fois continûment dérivable sur $\mathbb{R}^*$ et qu'il existe un polynôme P_n tel que $f^{(n)}(x) = P_n\left(\frac{1}{x}\right) \exp\left(-\frac{1}{x^2}\right)$.

On démontre ensuite, par récurrence, que f est n fois dérivable en 0 avec $f^{(n)}(0) = 0$. Si $n = 0$, il n'y a rien à démontrer. Supposons le résultat vrai pour un entier n. Alors $\frac{f^{(n)}(x) - f^{(n)}(0)}{x} = \frac{1}{x} P_n\left(\frac{1}{x}\right) \exp\left(-\frac{1}{x^2}\right)$. Il s'agit ainsi d'une combinaison linéaire de termes de la forme $\frac{1}{x^k} \exp\left(-\frac{1}{x^2}\right)$ qui tendent tous vers 0 d'après les résultats sur les croissances comparées (*cf.* la proposition 13 de la page 688). D'où la dérivabilité en 0 de $f^{(n)}$ avec $f^{(n+1)}(0) = 0$.

Les polynômes ont des équivalents très simples.

Exercice 2.
Soit P un polynôme *non nul* à coefficients réels.
On peut écrire $P(x) := a_n x^n + a_{n-1}x^{n-1} + \cdots + a_p x^p$, avec $n \geqslant p$, $a_p \neq 0$ et $a_n \neq 0$.

1. Montrer que $P \sim a_p x^p$ au voisinage de $a = 0$.
2. Montrer que $P \sim a_n x^n$ au voisinage de $a = \pm\infty$.
3. On note $\sigma \in \{+, -\}$ le signe de a_n. Montrer que :
   - si n est pair, P tend vers $\sigma\infty$ quand x tend vers $\pm\infty$;
   - si n est impair, P tend vers $\sigma\infty$ (resp. $-\sigma\infty$) quand x tend vers $+\infty$ (resp. $-\infty$).

On dit qu'au voisinage de zéro un polynôme (non nul) est équivalent à son monôme de plus bas degré et qu'au voisinage de l'infini il est équivalent à son monôme de plus haut degré.
Solution.

1. Il suffit d'écrire que $P(x) = a_p x^p \displaystyle\sum_{k=p}^{n} \frac{a_k}{a_p} x^{k-p}$, où la somme a visiblement 1 pour limite.
2. La démonstration est analogue en écrivant $P(x) = a_n x^n \displaystyle\sum_{k=p}^{n} \frac{a_k}{a_n} \frac{1}{x^{n-k}}$.
3. La preuve est immédiate.

Notons que des fonctions transcendantes « très simples » peuvent avoir un comportement « compliqué » à l'infini.

Exercice 3.
Montrer qu'aucune des fonctions $\exp$, $\ln$, $\sin$, $\cos$ n'est équivalente à un polynôme en $+\infty$.
Solution. Pour les deux premières, cela résulte aussitôt de l'exercice 2 et des croissances comparées des fonctions logarithme, exponentielle et puissances qui donnent :

$$\lim_{x\to+\infty} \mathrm{e}^x/x^p = +\infty \quad \text{ou} \quad \lim_{x\to+\infty} \ln x/x^p = 0 \qquad \text{pour tout } p > 0.$$

Les fonctions $\sin$ et $\cos$ étant *bornées*, elles ne peuvent pas avoir un équivalent en $+\infty$ de la forme ax^m, avec $m \geqslant 1$. Elles n'ont pas non plus d'équivalent de cette forme avec $m = 0$, sinon elles auraient une limite, ce qui n'est pas.

La notion d'équivalent est « compatible » avec les produits et les quotients.

Théorème 8.
Si $f_1 \sim g_1$ et $f_2 \sim g_2$ (au même point a), alors $f_1 f_2 \sim g_1 g_2$ et $f_1/f_2 \sim g_1/g_2$.

La preuve, basée sur les produit et quotient de limites, est laissée au lecteur.
On peut appliquer ce théorème comme ceci : sachant que $\cos x - 1 \sim -x^2/2$ au point 0, ce qui se déduit facilement du théorème 7 de la page 862, on voit que $(\cos x - 1)/x^2 \sim -1/2$ (ce qui donne donc la limite de $(\cos x - 1)/x^2$ en 0), et que $(\cos x - 1)^2 \sim x^4/4$ (ce qui est bien plus simple que de calculer les dérivées successives de $(\cos x - 1)^2$).

Exercice 4.
Donner un équivalent simple d'une fonction fraction rationnelle non nulle au voisinage de $\pm\infty$. En déduire qu'une fraction rationnelle est bornée au voisinage de l'infini si, et seulement si, son degré est positif.
Solution. Une fonction fraction rationnelle f non nulle s'écrit comme le quotient de deux polynômes P et Q non nuls, chacun d'entre eux étant respectivement équivalent, en $\pm\infty$, à un monôme de la forme $\alpha\, x^{\deg P}$ ou $\beta\, x^{\deg Q}$. Par quotient, $f(x) \sim a\, x^d$, où $d = \deg P - \deg Q$ est le degré de f et $a = \alpha/\beta$. Par suite f est bornée au voisinage de l'infini si, et seulement si, $d \geqslant 0$.

Le comportement vis-à-vis de l'addition (ou de la soustraction) n'est pas aussi bon. En fait, $f_1 \sim g_1$ et $f_2 \sim g_2$ n'implique pas nécessairement $f_1 + g_1 \sim f_2 + g_2$. Par exemple $\sin x \sim x$ et $-\tan x \sim -x$ au point 0, mais $\sin x - \tan x$ n'est pas équivalent à $x - x = 0$, puisqu'aucune fonction n'est équivalente à zéro.
On a :

$$\sin x - \tan x = \sin x\,(1 - 1/\cos x) \sim \sin x(\cos x - 1) \sim x(-x^2/2) = -x^3/2.$$

On ne peut pas appliquer les règles d'associativité de la somme pour calculer des équivalents. Soit, par exemple, $f := f_1 + f_2 + f_3$, avec $f_1 := 1 + x^3$, $f_2 := -1$ et $f_3 := x^2$. On a, au voisinage de 0, $f = x^3 + x^2 \sim x^2$, $g := f_2 + f_3 = -1 + x^2 \sim -1$ et $f_1 - 1 = x^3$ n'est pas équivalent à x^2.
La bonne façon d'obtenir des équivalents de sommes de fonctions est d'utiliser des développements limités, comme on le verra dans la suite de ce module.

Proposition 9. Soient deux fonctions f et g définies au voisinage de a.
On a $\mathrm{e}^f \sim \mathrm{e}^g$ si, et seulement si, $\lim\limits_{x\to a}(f - g) = 0$.

Démonstration. Les fonctions e^f et e^g ne s'annulent pas. Donc $\mathrm{e}^f \sim \mathrm{e}^g$ si, et seulement si, $\lim\limits_{x\to a} \mathrm{e}^{f-g} = 1$. Le résultat s'en déduit. ∎

On peut avoir $f \sim g$ sans avoir $\mathrm{e}^f \sim \mathrm{e}^g$. Soient, par exemple $f : x \in \mathbb{R}^* \mapsto \frac{1}{x^2} + \frac{1}{x}$ et $g : x \in \mathbb{R}^* \mapsto \frac{1}{x^2}$, f et g sont équivalentes au voisinage de 0, mais $\mathrm{e}^{f(x)-g(x)} = \mathrm{e}^{1/x}$ n'est pas bornée au voisinage de 0 et ne peut donc pas tendre vers 1.

Les exponentielles de deux fonctions équivalentes ne sont pas nécessairement équivalentes.

Si f et g ont une limite commune ℓ (finie) quand x tend vers a, on a :

$$\mathrm{e}^f - \mathrm{e}^g = \mathrm{e}^g(\mathrm{e}^{f-g} - 1) \sim \mathrm{e}^\ell(f - g) \tag{10}$$

au voisinage de 0. En particulier, si $\ell = 0$, $\mathrm{e}^f - \mathrm{e}^g \sim f - g$.

Proposition 10. Soient f et g à valeurs dans $\mathbb{R}_+^*$ équivalentes au voisinage de a. On suppose que g admet une limite ℓ, finie ou infinie, *différente de* 1, lorsque x tend vers a. Alors $\ln f \sim \ln g$.

Démonstration. Quitte à restreindre le domaine de définition de f et g, on peut supposer que f et g ne prennent pas la valeur 1. Alors $\ln f$ et $\ln g$ ne s'annulent pas et l'on peut écrire :

$$\frac{\ln f}{\ln g} - 1 = \frac{\ln(f/g)}{\ln g}.$$

On convient de $1/\ln \ell = 0$ si $\ell = 0$ ou $\ell = +\infty$. On a $\lim f/g = 1$, $\lim 1/\ln g = 1/\ln \ell$, donc $\lim \frac{\ln(f/g)}{\ln g} = 0$ et $\lim(\ln f/\ln g) = 1$. ■

Exercice 5.

Trouver un équivalent simple en 0 de $\varphi : \begin{array}{rcl}]0,\pi/2[& \longrightarrow & \mathbb{R} \\ x & \longmapsto & x^x - (\sin x)^{\sin x}. \end{array}$

Solution.

On note $f(x) := x\ln x$ et $g(x) := \sin x \ln(\sin x)$. On a $\lim\limits_{x\to 0} f(x) = 0$ $\lim\limits_{x\to 0} g(x) = 0$.

Or, $\varphi(x) = \mathrm{e}^{f(x)} - \mathrm{e}^{g(x)}$, donc, d'après (10), $\varphi \sim (f-g)$, c'est-à-dire :

$$\varphi(x) \sim x\ln x - \sin x \ln(\sin x).$$

On a :

$$\sin x \ln(\sin x) = \sin x \ln \frac{\sin x}{x} + \sin x \ln x,$$

et, en notant $h(x) := \sin x \ln \frac{\sin x}{x}$:

$$\varphi(x) \sim x\ln x - \sin x \ln(\sin x) = (x - \sin x)\ln x - h(x)$$

On a $x - \sin x \sim \frac{x^3}{6}$ et, en utilisant $\frac{\sin x}{x} - 1 = \frac{\sin x - x}{x}$:

$$h(x) = \sin x \ln \frac{\sin x}{x} \sim \sin x \frac{\sin x - x}{x} \sim -\frac{x^3}{6},$$

$$\varphi(x) \sim x^3 \ln x \left(\frac{x - \sin x}{x^3} - \frac{h(x)}{x^3 \ln x} \right) \qquad \text{et} \qquad \lim_{x\to 0} \frac{\sin x - x}{x^3} - \frac{h(x)}{x^3 \ln x} = \frac{1}{6} - 0,$$

donc $\varphi(x) \sim \frac{x^3}{6} \ln x$.

Exercice 6.

On note $I :=]-\pi/2, \pi/2[$. On définit deux fonctions $u,\ v : I \setminus \{0\} \to \mathbb{R}$ par

$$u(x) := \sqrt{3 + \frac{\sinh x}{\sin x}} \qquad \text{et} \qquad v(x) := 1 + \cos x.$$

(i) Montrer que, pour tout $x \in I \setminus \{0\}$, $u(x) > v(x) > 0$.

(ii) Trouver la limite en 0 de la fonction $f := \dfrac{u^v - v^u}{u - v}$.

Solution.

(i) On a, pour tout $x \in]0, \pi/2[$, $\sin x < x$ et $\sinh x > x$, donc, en utilisant la parité, pour tout $x \in I \setminus \{0\}$, $\frac{\sinh x}{\sin x} > 1$ et $u > 2$.
Par ailleurs $0 < v < 2$. D'où $u(x) > v(x) > 0$.

(ii) D'après (i), la fonction f est définie sur $I \setminus \{0\}$. On a :

$$\lim_{x\to 0}(u - v) = \lim_{x\to 0}(u^v - v^u) = 0;$$

on dit que l'on a une forme indéterminée $\frac{0}{0}$.

On a $f = v^u \dfrac{\mathrm{e}^{v\ln u - u\ln v} - 1}{u - v}$ avec $\lim\limits_{x\to 0} v^u = 4$ et $\mathrm{e}^{v\ln u - u\ln v} - 1 \sim v\ln u - u\ln v$, donc, en posant $g := \dfrac{v\ln u - u\ln v}{u - v}$, on obtient $f \sim 4g$.

On a $g + \ln u = \dfrac{\ln\frac{v}{u}}{\frac{v}{u} - 1} = \dfrac{\ln\left(1 + \frac{v}{u} - 1\right)}{\frac{v}{u} - 1}$, donc $\lim\limits_{x\to 0} g + \ln u = 1$ et $\lim\limits_{x\to 0} g = 1 - \ln 2$.
Par suite $\lim\limits_{x\to 0} f = 4(1 - \ln 2)$.

3.2 Notations de Landau

3.2.1 Fonctions négligeables — Notation o de Landau

Définition 2. Soient f et g deux fonctions définies sur D, la seconde ne s'annulant pas au voisinage de a sauf peut-être en a. On dit que f *est négligeable devant* g *au voisinage de* a si :

$$\lim_{x\to a} \frac{f(x)}{g(x)} = 0. \tag{11}$$

On note $f = \mathrm{o}_a(g)$, ou $f = \mathrm{o}(g)$ si le point a est sous-entendu.

On notera qu'aucune fonction n'est négligeable devant la fonction nulle.
La notation, due à Edmund Landau, est doublement inhabituelle. D'une part elle utilise la lettre o (à ne pas confondre avec le chiffre 0) : la relation $f = \mathrm{o}(g)$ est lue « f est un petit-o de g ». D'autre part, il est anormal d'utiliser le signe $=$ pour indiquer ce type de relation : on devrait plutôt écrire $f \in \mathrm{o}(g)$ pour dire que f appartient à la classe des fonctions négligeables devant g au point a.
Cette notation s'est pourtant bien imposée (en mathématiques et en sciences appliquées) pour son caractère commode.
Les deux équivalences suivantes sont très utiles (les vérifications sont faciles et laissées au lecteur) :

$$f = \mathrm{o}_a(1) \iff \lim_{x\to a} f(x) = 0. \tag{12}$$

$$f \sim_a g \iff f - g = \mathrm{o}_a(g). \tag{13}$$

Notation. On utilisera souvent par la suite la notation $f = g + \mathrm{o}_a(h)$. *Par définition*, elle signifie que $f - g = \mathrm{o}_a(h)$.

S'il existe une suite (x_n) à valeurs dans $D \setminus \{a\}$, convergeant vers a, et telle que $g(x_n) = 0$ pour tout n, alors on ne peut pas écrire $f = \mathrm{o}(g)$. On ne peut pas, par exemple, écrire $x\sin(1/x) = \mathrm{o}\left(\sin(1/x)\right)$. Pour éviter ce problème on peut introduire une définition de négligeabilité *plus générale* que la nôtre. Soient f et g deux fonctions définies sur D. On dit alors que f est négligeable devant g s'il existe une fonction ε, définie sur D, tendant vers 0 quand x tend vers a et telle que $f = \varepsilon g$. On trouve cette définition dans plusieurs ouvrages. Avec cette définition $x\sin(1/x) = \mathrm{o}\left(\sin(1/x)\right)$.

Remarque. Soient f et g deux fonctions définies sur D, *la seconde ne s'annulant pas au voisinage de a sauf peut-être en a*. Alors $f = \mathrm{o}(g)$ si, et seulement s'il existe une fonction ε définie au voisinage de a telle que $f = \varepsilon g$ et $\lim\limits_{x \to a} \varepsilon(x) = 0$ (il suffit de poser $\varepsilon = f/g$). Le maniement des o est délicat et le débutant pourra dans certains cas utiliser cette façon d'écrire.

Proposition 11. La relation « être négligeable au voisinage de a » est :

1. antiréflexive, c'est-à-dire que $f = \mathrm{o}(f)$ est toujours faux ;
2. strictement antisymétrique, c'est-à-dire que si $f = \mathrm{o}(g)$, alors $g = \mathrm{o}(f)$ est faux ;
3. transitive, c'est-à-dire que si $f = \mathrm{o}(g)$ et $g = \mathrm{o}(h)$, alors $f = \mathrm{o}(h)$.

On dit alors que la relation « être négligeable au voisinage de a » est une relation d'ordre strict.
La démonstration est laissée au lecteur, qui pourra aussi noter que la fonction nulle est un élément minimal pour cette relation, car elle est négligeable devant toute autre, alors qu'aucune fonction n'est négligeable devant elle (ce n'est pas le seul élément minimal : toute fonction qui s'annule sur un intervalle centré en a est aussi un élément minimal).

Exemple. Les résultats concernant les croissances comparées des fonctions usuelles (proposition 13 de la page 688) peuvent s'écrire, pour tout réel $\alpha > 0$, sous la forme suivante :

$$\ln x = \mathrm{o}(x^\alpha) \quad \text{et} \quad x^\alpha = \mathrm{o}(\mathrm{e}^x) \qquad \text{au voisinage de } +\infty.$$

La notion de négligeabilité se comporte bien pour l'addition :

Proposition 12. Soient f_1, f_2 deux fonctions négligeables devant g.
Alors $f_1 + f_2$, $f_1 - f_2$ sont négligeables devant g.

La preuve est immédiate. Une autre façon d'écrire cette proposition est :

$$\big(f_1 = \mathrm{o}(g) \quad \text{et} \quad f_2 = \mathrm{o}(g)\big) \Longrightarrow f_1 \pm f_2 = \mathrm{o}(g).$$

L'erreur à ne surtout pas faire ici serait de « soustraire les $\mathrm{o}(g)$ » en écrivant, à tort, $f_1 - f_2 = \mathrm{o}(g) - \mathrm{o}(g) = 0$! C'est là le principal piège des notations de Landau.

Proposition 13. Si $f = \mathrm{o}(g)$, alors pour toute constante $\lambda \in \mathbb{R}$, on a $\lambda f = \mathrm{o}(g)$.
De plus, pour toute fonction h qui ne s'annule pas au voisinage de a, sauf éventuellement en a, on a $fh = \mathrm{o}(gh)$.
En particulier, au voisinage de 0, pour tous $p, q \in \mathbb{N}^*$, on a $x^p \mathrm{o}(x^q) = \mathrm{o}(x^{p+q})$.

C'est évident d'après la définition. En conséquence, on peut par exemple, au voisinage de 0, écrire indifféremment $4\,\mathrm{o}(x)$, $\mathrm{o}(3x)$ ou $\mathrm{o}(x)$; ou encore $x^2\mathrm{o}(\ln x)$ ou $\mathrm{o}(x^2 \ln x)$. Dans la pratique, on préfère supprimer les constantes et placer (si possible) toutes les fonctions dans le petit-o.
D'après les deux propositions précédentes, si g est une fonction donnée, ne s'annulant pas au voisinage de a sauf peut-être en a, l'ensemble des fonctions négligeables devant g est un *espace vectoriel*.
Dans la partie 2.6.2 du module IV.2 nous avons vu des formules qui se terminaient par un terme de la forme $(x-a)^n\varepsilon(x)$ avec $\lim\limits_{x\to a} \varepsilon(x) = 0$. Un tel terme est négligeable devant $(x-a)^n$; réciproquement, toute fonction négligeable devant $(x-a)^n$ est de cette forme. Une fonction négligeable devant $(x-a)^n$ est normalement écrite $\mathrm{o}\left((x-a)^n\right)$; on pourra, par abus de notation, l'écrire $\mathrm{o}(x-a)^n$ pour éviter les multiples parenthèses.
On peut réécrire le théorème 54 de la page 655 de la façon suivante :

Théorème 14 (Formule de Taylor-Young avec notations de Landau). Soient $n \in \mathbb{N}^*$ et f une fonction n fois continûment dérivable sur un intervalle I. Alors, pour tout point $a \in I$, on a :

$$f(x) = \sum_{k=0}^{n} \frac{(x-a)^k}{k!} f^{(k)}(a) + \mathrm{o}(x-a)^n.$$

3.2.2 Domination — Notation O de Landau

Définition 3. Soient f et g deux fonctions définies sur D, *la seconde ne s'annulant pas au voisinage de* a, sauf peut-être en a. On dit que f *est dominée par* g *au voisinage de* a si la fonction f/g est bornée au voisinage de a.
On note alors $f = \mathrm{O}_a(g)$, ou $f = \mathrm{O}(g)$ en sous-entendant le point a.

C'est l'autre notation de Landau, lue « f est un grand-O de g ». Notons que $f = \mathrm{o}(g)$ implique $f = \mathrm{O}(g)$ (car une fonction de limite nulle en a est bornée au voisinage de a), mais pas l'inverse.
La relation O possède des propriétés similaires à celles de la relation o. On a, en particulier, pour l'addition et la multiplication le résultat suivant.

Proposition 15. Pour toutes les fonctions f, g, h, on a au voisinage d'un même point a :

1. si $f_1 = \mathrm{O}(g)$ et $f_2 = \mathrm{O}(g)$, alors $f_1 + f_2 = \mathrm{O}(g)$ et $f_1 - f_2 = \mathrm{O}(g)$;
2. si $f = \mathrm{O}(g)$ et $\lambda \in \mathbb{R}$, alors $\lambda f = \mathrm{O}(g)$;
3. si $f = \mathrm{O}(g)$ et si h ne s'annule pas au voisinage de a, alors $fh = \mathrm{O}(gh)$.

On peut combiner des o et O et des équivalents de diverses façons.

Proposition 16. Pour toutes les fonctions f, g, h, on a au voisinage d'un même point a :

1. $f = \mathrm{O}(g)$ et $g = \mathrm{O}(h)$ impliquent $f = \mathrm{O}(h)$;
2. $f = \mathrm{o}(g)$ et $g = \mathrm{O}(h)$ impliquent $f = \mathrm{o}(h)$;
3. $f = \mathrm{O}(g)$ et $g = \mathrm{o}(h)$ impliquent $f = \mathrm{o}(h)$.
4. $f = \mathrm{O}(g)$ et $g \sim h$ impliquent $f = \mathrm{O}(h)$.
5. $f = \mathrm{o}(g)$ et $g \sim h$ impliquent $f = \mathrm{o}(h)$.
6. $f \sim g$ et $g = \mathrm{O}(h)$ impliquent $f = \mathrm{O}(h)$.
7. $f \sim g$ et $g = \mathrm{o}(h)$ impliquent $f = \mathrm{o}(h)$.

La démonstration résulte immédiatement de la définition et du fait que le produit d'une fonction bornée par une fonction tendant vers zéro tend aussi vers zéro.

On notera que O *n'est pas une relation d'ordre entre fonctions*, car on peut très bien avoir $f = \mathrm{O}(g)$ et $g = \mathrm{O}(f)$ (au même point) avec $f \neq g$. Si $f = \mathrm{O}(g)$ et $g = \mathrm{O}(f)$, on dit que f et g sont *du même ordre au point* a.

C'est notamment vrai si les deux fonctions sont équivalentes (comme par exemple $\sin x$ et x au voisinage de 0), ou plus généralement si le quotient $f(x)/g(x)$ tend vers une limite *finie et non nulle*.

Exercice 7.

Montrer que deux fonctions f et g, ne s'annulant pas au voisinage de $+\infty$, sont du même ordre au voisinage de $+\infty$ si, et seulement s'il existe $M \in \mathbb{R}$ et $A \geqslant 1$ tels que :

$$x \geqslant M \Longrightarrow \frac{1}{A} \leqslant \left|\frac{f(x)}{g(x)}\right| \leqslant A.$$

En déduire un exemple de fonctions qui ne sont pas du même ordre au voisinage de $+\infty$, mais dont aucune n'est négligeable devant l'autre.

Solution. Supposons f et g du même ordre. Par définition, on a $|f(x)| \leqslant A_1\,|g(x)|$ et $|g(x)| \leqslant A_2\,|f(x)|$ pour tous les x vérifiant respectivement $x \geqslant M_1$ et $x \geqslant M_2$. Prenant $M := \max(M_1, M_2)$ et $A := \max(A_1, A_2)$ (ou $A := A_1 + A_2$), on obtient les inégalités cherchées. On a bien $A \geqslant 1$ puisqu'en prenant $x \geqslant M$, on a $g(x) \neq 0$ et $\big|g(x)\big| \leqslant A_1\,A_2\,\big|g(x)\big|$, ce qui prouve que $A_1\,A_2 \geqslant 1$ et donc que l'un des deux réels A_1, A_2 est supérieur ou égal à 1.

L'implication réciproque est évidente.

Pour trouver un exemple comme demandé par l'énoncé, il suffit de prendre $g(x) := 1$ et $f(x) := \mathrm{e}^{-x} + \mathrm{e}^{x}\sin^2 x$ sur $\mathbb{R}_+$ puisque le rapport $f(x)/g(x)$ n'est ni majoré (prendre $x := (2n+1)\pi/2$) ni minoré par un réel strictement positif (prendre $x := n\pi$).

---

4 Développements limités

4.1 Définition et premières propriétés — Exemples

Définition 4. Soit $n \in \mathbb{N}$. On dit qu'une fonction f *admet un développement limité à l'ordre* n au point a s'il existe un polynôme P de degré au plus n tel que :

$$f(x) = P(x) + \mathrm{o}(x-a)^n \tag{14}$$

au voisinage de a.

Il est équivalent de dire que $f(x) = P(x) + (x-a)^n \varepsilon(x)$, avec $\lim\limits_{x \to a} \varepsilon(x) = 0$.

Le polynôme P est appelé la *partie régulière* du développement limité.

Les polynômes sont habituellement écrits $P(x) := a_n x^n + \cdots + a_0$. Mais ils peuvent aussi s'écrire comme combinaison de puissances de $(x-a)$ (avec des coefficients différents si $a \neq 0$). En rangeant ces puissances par ordre croissant, on peut réécrire la formule (14) sous la forme :

$$f(x) = \alpha_0 + \alpha_1(x-a) + \cdots + \alpha_n(x-a)^n + \mathrm{o}(x-a)^n = Q(x-a) + \mathrm{o}(x-a)^n, \tag{15}$$

où $Q(t) := \alpha_0 + \alpha_1 t + \cdots + \alpha_n t^n$ et $P(x) = Q(x-a) = \alpha_0 + \alpha_1(x-a) + \cdots + \alpha_n(x-a)^n$.

Les nombres α_k sont souvent appelés les *coefficients du développement limité*.

Notons que les puissances successives de $(x-a)$ forment une suite de fonctions, chacune négligeable devant la précédente au point a. En particulier, une fonction qui a un développement limité à l'ordre n a aussi un développement limité à tout ordre inférieur : il suffit de supprimer les puissances de $(x-a)$ trop grandes dans la formule (15). On prendra garde au fait que ce n'est pas la même chose que de supprimer les puissances de x trop grandes dans P. En pratique, il vaut mieux écrire la partie régulière d'un développement limité en a comme un polynôme en $(x-a)$ plutôt que comme un polynôme en x.

Théorème 17. Si une fonction f admet au voisinage de a un développement limité de la forme :

$$f(x) = \alpha_k(x-a)^k + \cdots + \alpha_n(x-a)^n + \mathrm{o}(x-a)^n$$

avec $\alpha_k \neq 0$, alors elle est *équivalente* à $\alpha_k(x-a)^k$ en a.

Démonstration. Comme pour tout $p > k$, on a $(x-a)^p = \mathrm{o}(x-a)^k$, on en déduit :

$$f(x) = \alpha_k(x-a)^k + \mathrm{o}(x-a)^k \qquad \text{c'est-à-dire} \qquad f(x) \sim \alpha_k(x-a)^k. \qquad \blacksquare$$

Proposition 18. Si une fonction admet un développement limité à l'ordre n au point a, alors celui-ci est unique.

Démonstration. Supposons au contraire que f ait deux développements limités différents, celui donné par (15) et un autre : $f(x) = \beta_0 + \cdots + \beta_n(x-a)^n + \mathrm{o}(x-a)^n$. Comme on les suppose différents, il existe au moins un indice $k \in [\![0, n]\!]$ tel que $\alpha_k \neq \beta_k$. Notons k_0 le

plus petit d'entre eux. En éliminant les premiers termes égaux dans les deux égalités sur f, on trouve, par la proposition précédente :

$$0 = f(x) - f(x) = \sum_{k=k_0}^{n} (\beta_k - \alpha_k)(x-a)^k + \mathrm{o}(x-a)^n \sim (\beta_{k_0} - \alpha_{k_0})(x-a)^{k_0}$$

ce qui est contradictoire. ■

Il résulte en particulier de cette proposition qu'il n'existe qu'un seul polynôme P de degré inférieur ou égal à n tel que $f(x) = P(x) + \mathrm{o}(x-a)^n$. On l'appelle le *polynôme d'approximation* de f à l'ordre n au point a.

Point Méthode

Pour chercher la limite en a d'une fonction f ou, plus généralement, pour chercher un équivalent en a de f de la forme $\alpha(x-a)^p$, on cherche un développement limité de f en a *dont la partie régulière n'est pas nulle*. Si un tel développement existe, alors, d'après le théorème 17, sa partie principale, considérée comme polynôme en $(x-a)$, est équivalente à son monôme non nul de plus bas degré $\alpha_k(x-a)^k$. Si $k=0$, alors $\lim\limits_{x\to a} f(x) = \alpha_0$.

Si f admet, en a, le développement limité :

$$f(x) = \alpha_0 + \alpha_1(x-a) + \cdots + \alpha_n(x-a)^n + \mathrm{o}(x-a)^n,$$

la fonction $g : t \mapsto g(t) := f(a+t)$ admet, en 0, le développement limité :

$$g(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_n t^n + \mathrm{o}(t^n),$$

et inversement. Ainsi, l'on passe de la notion de développement limité en 0 à celle de développement limité en a par le changement de variable $t \mapsto x = t+a$, $x \mapsto t := x-a$.

Point Méthode

Au lieu d'étudier un développement limité au voisinage du point a, il est souvent préférable de se ramener à un développement limité au voisinage de 0 par le changement de variable $t \mapsto x = t+a$, $x \mapsto t := x-a$.

Remarque. Soient $p, q \in \mathbb{N}^*$. On a vu que $x^p\mathrm{o}(x^q) = \mathrm{o}(x^{p+q})$. Par suite, pour avoir un développement limité de $f := x^p g$ à l'ordre $p+q$ en 0, il suffit d'avoir un développement limité de g à l'ordre q en 0 : $g(x) = \alpha_0 + \alpha_1 t + \cdots + \alpha_q x^q + \mathrm{o}(x^q)$:

$$f(t) = \alpha_0 x^p + \alpha_1 x^{p+1} + \cdots + \alpha_q t x^{p+q} + \mathrm{o}(x^{p+q}).$$

Proposition 19. Si des fonctions f et g admettent un développement limité à l'ordre n au point a, alors pour tout réel λ la fonction $\lambda f + g$ admet également un développement limité à l'ordre n au point a obtenu par combinaison linéaire.

Démonstration. Il suffit d'appliquer la définition et les propriétés de $\mathrm{o}(x-a)^n$. ■

Par exemple, au voisinage de 0 :

$$\mathrm{e}^x - 2\cos x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \mathrm{o}(x^3) - 2\left(1 - \frac{x^2}{2} + \mathrm{o}(x^3)\right) = -1 + x + \frac{3x^2}{2} + \frac{x^3}{6} + \mathrm{o}(x^3).$$

Proposition 20. Soit f une fonction admettant un développement limité à l'ordre n en 0 : $f(x) = a_0 + a_1 x + \cdots + a_n x^n + \mathrm{o}(x^n)$.

- Si f est paire, alors le développement limité ne contient pas de termes d'indice impair (c'est-à-dire $a_{2k+1} = 0$ pour tout entier k tel que $2k+1 \leqslant n$) et
$$f(x) = a_0 + a_2 x^2 + \cdots + a_{2p} x^{2p} + \mathrm{o}(x^{2p+1}).$$
- Si f est impaire, alors le développement limité ne contient pas de termes d'indice pair (c'est-à-dire $a_{2k} = 0$ pour tout entier k tel que $2k \leqslant n$) et
$$f(x) = a_1 x + a_3 x^3 + \cdots + a_{2p+1} x^{2p+1} + \mathrm{o}(x^{2p+2}).$$

Démonstration. En remplaçant x par $-x$ dans le développement limité de f, on obtient :
$$f(-x) = a_0 - a_1 x + \cdots + n a_n (-x)^n + \mathrm{o}\big((-x)^n\big) = a_0 - a_1 x + \cdots + (-1)^n a_n x^n + \mathrm{o}(x^n).$$
Par suite, pour une fonction paire, on a l'égalité $f(-x) = f(x)$ et l'unicité du développement limité nous donne $\forall k \in [\![0, n]\!]$, $(-1)^k a_k = a_k$, soit $a_{2k+1} = 0$ pour tout k tel que $2k+1 \leqslant n$. Le développement à l'ordre $n := 2p+1$ est :
$$f(x) = a_0 + a_2 x^2 + \cdots + a_{2p} x^{2p} + a_{2p+1} x^{2p+1} + \mathrm{o}(x^{2p+1}),$$
c'est-à-dire $f(x) = a_0 + a_2 x^2 + \cdots + a_{2p} x^{2p} + \mathrm{o}(x^{2p+1})$, puisque $a_{2p+1} = 0$.
Lorsque f est impaire, l'égalité $f(-x) = -f(x)$ donne $a_{2k} = 0$ pour tout k tel que $2k \leqslant n$. On améliore d'une unité l'ordre du développement par le même argument que dans le cas pair. ∎

Une fonction f étant donnée, comment savoir si elle admet un développement limité, et trouver ses coefficients ? Il est évident qu'un polynôme admet un développement limité en tout point et à tout ordre, car, si P est de degré n : $P(x) = P(x) + 0 = P + \mathrm{o}(x-a)^n$. Mais que peut-on dire des autres fonctions ? Nous avons en fait déjà rencontré (sans le savoir...) un cas très important avec la formule de Taylor-Young.

Théorème 21. Soient $n \in \mathbb{N}$ et f une fonction n fois continûment dérivable sur un intervalle I. Alors pour tout point $a \in I$, f a un développement limité d'ordre n au point a, dont les coefficients sont $\alpha_k = f^{(k)}(a)/k!$, $k \in [\![0, n]\!]$.

Il n'y a rien à démontrer, car c'est simplement une autre façon d'énoncer la formule de Taylor-Young (*cf.* le théorème 54 de la page 655). Les coefficients doivent avoir la valeur indiquée, puisqu'ils sont uniques. Ainsi le polynôme de Taylor, lorsqu'il existe, est égal au polynôme d'approximation. En fait, en utilisant le théorème 55 de la page 656 , il suffit, dans l'énoncé du théorème ci-dessus, de supposer que f est n fois dérivable *au point* a.
On peut être tenté de penser que la réciproque est vraie, c'est-à-dire qu'une fonction qui admet un développement limité à l'ordre n en un point a est dérivable n fois en ce point. C'est effectivement le cas si $n = 0$ ou $n = 1$. Dans le premier cas, le développement limité s'écrit $f(x) = \alpha_0 + \mathrm{o}(1)$ et donc $\alpha_0 = f(a)$ par définition d'un $\mathrm{o}(1)$. Si $n = 1$ et $f(x) = \alpha_0 + \alpha_1(x-a) + \mathrm{o}(x-a)$, alors $\alpha_0 = f(a)$ pour la même raison, et par suite :
$$\frac{f(x) - f(a)}{x - a} = \alpha_1 + \mathrm{o}(1)$$

ce qui implique que le taux d'accroissement a une limite égale à α_1 : ainsi f est bien dérivable en a et $\alpha_1 = f'(a)$.
Mais la réciproque est fausse pour $n = 2$, comme le montre l'exercice ci-après.

Exemple. Soit f la fonction définie par $f(0) = 0$ et $f(x) = x^3 \sin(1/x^3)$ si $x \neq 0$. On a, pour tout $x \in \mathbb{R}$, $f(x) = x^2\varepsilon(x)$, avec $\varepsilon(x) := x\sin(1/x^3)$ si $x \in \mathbb{R}^*$ et $\varepsilon(0) := 0$. En utilisant $|\sin| \leqslant 1$, on montre que $\lim\limits_{x\to 0} \varepsilon(x) = 0$, donc $f(x) = x^2\varepsilon(x)$ est un développement limité à l'ordre 2, de partie régulière nulle. La fonction f est dérivable sur $\mathbb{R}^*$ et l'on a :

$$\forall x \in \mathbb{R}^*, \quad f'(x) = 3x^2 \sin(1/x^3) - \frac{3}{x}\cos(1/x^3).$$

On vérifie que la fonction $\cos(1/x^3)$ prend une infinité de fois la valeur 1 au voisinage de 0 et $\lim\limits_{x\to 0} 3x^2 \sin(1/x^3) = \lim\limits_{x\to 0} 3x\varepsilon(x) = 0$, donc la fonction f' n'est pas bornée au voisinage de 0. *A fortiori* f' n'est pas dérivable en 0, donc f n'est pas deux fois dérivable en 0.

Une fonction peut admettre un dévelopement limité à l'ordre 2 en un point sans être deux fois dérivable en ce point.

Exercice 8.

Soit $f : x \in \mathbb{R}^* \mapsto f(x) := \mathrm{e}^{-1/x^2}$. Montrer, *sans utiliser l'exercice* 1 de la page 863, que f admet, à tout ordre, un développement limité en 0 dont la partie régulière est nulle.
Solution. Montrons que, pour tout $n \in \mathbb{N}$, $f = \mathrm{o}(x^n)$. La fonction f étant paire, il suffit de considérer le cas $x > 0$. On a :

$$g_n(x) := x^{-n} f(x) = x^{-n}\mathrm{e}^{-\frac{1}{x^2}} = \mathrm{e}^{-\frac{1}{x^2} - n\ln x} = \mathrm{e}^{-\frac{1}{x^2}\left(1 + nx^2 \ln x\right)}.$$

Or, $\lim\limits_{x\to 0} 1 + nx^2 \ln x = 1$, donc $\lim\limits_{x\to 0} -\frac{1}{x^2}\left(1 + nx^2\ln x\right) = -\infty$ et $\lim\limits_{x\to 0} g_n(x) = 0$.

Utilisation du calcul formel

Les logiciels de calcul formel permettent d'écrire très facilement les développements limités de certaines fonctions. Par exemple en utilisant `Maple`, on peut utiliser la fonction **`taylor`** qui écrit le développement limité à l'ordre n d'une fonction f à l'origine : **`taylor(f(x),x=0,n+1)`**. Par exemple **`dl11:=taylor(sin(x),x=0,12);`** donne :

$$dl11 := x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \frac{x^9}{362880} - \frac{x^{11}}{39916800} + \mathrm{O}(x^{12}).$$

Si l'on souhaite obtenir seulement la partie régulière, on convertit le résultat en polynôme à l'aide de **`dl11:=convert(dl11,polynom)`** :

$$dl11 := x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \frac{x^9}{362880} - \frac{x^{11}}{39916800}.$$

On peut ensuite, si on le souhaite, comparer les graphes de f et de l'un de ses polynômes de Taylor en utilisant la fonction **`plot`** (*cf.* pour la fonction $f := \sin$, la figure IV.8.1 de la page 856).

$$
\begin{aligned}
\mathrm{e}^x &= 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \mathrm{o}(x^n) \\
\cos x &= 1 - \frac{x^2}{2} + \frac{x^4}{4!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \mathrm{o}(x^{2n+1}) \\
\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \mathrm{o}(x^{2n+2}) \\
\cosh x &= 1 + \frac{x^2}{2} + \frac{x^4}{4!} + \cdots + \frac{x^{2n}}{(2n)!} + \mathrm{o}(x^{2n+1}) \\
\sinh x &= x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} + \mathrm{o}(x^{2n+2}) \\
(1-x)^{-\alpha} &= 1 + \alpha x + \frac{\alpha(\alpha+1)}{2!} x^2 + \cdots + \frac{\alpha(\alpha+1)\ldots(\alpha+n-1)}{n!} x^n + \mathrm{o}(x^n) \\
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots + (-1)^{n+1} \frac{x^n}{n} + \mathrm{o}(x^n)
\end{aligned}
$$

TABLE IV.8.2 – Développements limités des fonctions usuelles au voisinage de 0. Ici $n \in \mathbb{N}$ et $\alpha \in \mathbb{R}$ sont quelconques.

4.2 Développements limités des fonctions usuelles

4.2.1 Fonctions transcendantes usuelles

La table IV.8.2 donne les développements limités, au voisinage de 0, de quelques fonctions transcendantes usuelles. Dans cette table, les développements limités de l'exponentielle, de $\sin$ et $\cos$ et celui de $\mapsto \ln(1+x)$ s'obtiennent par la formule de Taylor-Young. Ceux de $\sinh$ et $\cosh$ proviennent du fait que ces fonctions sont les parties impaires et paires de l'exponentielle.

Le développement limité de $(1-x)^{-\alpha}$ s'obtient par la formule du binôme lorsque $-\alpha \in \mathbb{N}$, car la fonction est alors un polynôme.

Pour le cas général $\alpha \in \mathbb{R}$, on démontre par récurrence que la $n^{ème}$ dérivée est :

$$\alpha(\alpha-1)\cdots(\alpha-n+1)(1-x)^{-\alpha-n}$$

et la formule de Taylor-Young donne alors le résultat. En particulier, si $\alpha = 1$, on obtient :

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots + x^n + \mathrm{o}(x^n). \tag{16}$$

On retrouve ce résultat en utilisant la formule donnant la somme des termes d'une suite géométrique :

$$\frac{1}{1-x} - (1 + x + x^2 + x^3 + \cdots + x^n) = \frac{x^{n+1}}{1-x} = \mathrm{o}(x^n)$$

($x^{n+1} = \mathrm{o}(x^n)$ et $1/(1-x)$ est bornée au voisinage de 0).

4.2.2 Fractions rationnelles

Une fonction fraction rationnelle qui admet un développement limité en 0 admet une limite en 0. Nous allons voir qu'inversement cette condition est suffisante et donner des méthodes de calcul du développement.

Commençons par une remarque simple. On a des variantes du développement limité :

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots + \mathrm{o}(x^n), \tag{17}$$

en *substituant* au polynôme x un polynôme T s'annulant en 0. En effet $T(x) = xS(x)$ et le polynôme S est borné au voisinage de 0, donc $\mathrm{o}(T^n) = \mathrm{o}(x^n)$.

Plus généralement si $T = x^k S$ ($k \in \mathbb{N}^*$), alors $\mathrm{o}(T^n) = \mathrm{o}(x^{kn})$.

Par exemple en posant $T := -x$, ou $T := -x^2$, on obtiendra de nouveau un développement limité au point 0. On trouve :

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \cdots + (-1)^n x^n + \mathrm{o}(x^n), \tag{18}$$

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots + (-1)^n x^{2n} + \mathrm{o}(x^{2n}). \tag{19}$$

Plus généralement, cela permet de montrer que toute fonction fraction rationnelle de la forme $1/B$, où B est un polynôme ne s'annulant pas à l'origine, admet un développement limité à tout ordre à l'origine et de le calculer. On note $b := B(0) \neq 0$ et $T := 1 - b^{-1}B$. Alors $\dfrac{1}{B} = b^{-1}\dfrac{1}{1-T}$ et l'on a :

$$\begin{aligned}\frac{1}{1-T} = 1 + T + T^2 + \cdots + T^n + \frac{T^{n+1}}{1-T} &= 1 + T + T^2 + \cdots + T^n + \mathrm{o}(T^n)\\ &= 1 + T + T^2 + \cdots + T^n + \mathrm{o}(x^n).\end{aligned} \tag{20}$$

En remplaçant dans (20) chaque T^p par le polynôme obtenu en supprimant les termes de degré strictement supérieur à n et en additionnant les résultats obtenus, on obtient un développement limité à l'ordre n à l'origine.

Si $T = \gamma_k t^k + \cdots$ avec $k \geqslant 2$ et $\gamma_k \neq 0$, on a, pour tout $p \in \mathbb{N}^*$, $T^p = t^{pk}(\gamma_k + \cdots)^p$. Pour obtenir un développement limité à l'ordre n de $\frac{1}{1-T}$, il est inutile de considérer les T^p pour $kp > n$ et les monômes de $(\gamma_k + \cdots)^p$ de degré strictement supérieur à $n - kp$.

Exemple. On cherche un développement limité en 0, à l'ordre 6, de :

$$f(x) := \frac{1}{1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!}} = \frac{1}{1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 - \frac{1}{720}x^6}.$$

On pose $T := \frac{1}{2}x^2 - \frac{1}{24}x^4 + \frac{1}{720}x^6 = x^2\left(\frac{1}{2} - \frac{1}{12}x^2 + \frac{1}{360}x^4\right)$. Il suffit de considérer $1+T+T^2+T^3$ ($2\times 4 = 8 > 6$) en négligeant les monômes de degré strictement supérieur à 4 du développement de $\left(\frac{1}{2} - \frac{1}{12}x^2 + \frac{1}{360}x^4\right)^2$ et ceux de degré strictement supérieur à 0 du développement de $\left(\frac{1}{2} - \frac{1}{12}x^2 + \frac{1}{360}x^4\right)^3$.

On obtient :

$$1+\frac{1}{2}x^2-\frac{1}{24}x^4+\frac{1}{720}x^6+x^4\left(\frac{1}{4}-\frac{1}{24}x^2\right)+\frac{1}{8}x^6=1+\frac{1}{2}x^2+\frac{5}{24}x^4+\frac{61}{720}x^6.$$

Le développement limité cherché est donc $f(x)=1+\frac{1}{2}x^2+\frac{5}{24}x^4+\frac{61}{720}x^6+\mathrm{o}(x^6)$.

Nous apprendrons plus loin à multiplier les développements limités. Ainsi, connaissant un développement limité de $1/B$ on pourra donc, pour tout polynôme A, obtenir un développement limité de $A/B=A(1/B)$. Nous allons donner une autre méthode pour obtenir un tel développement.

Exemple. Soient $A:=x^3+2x+1$ et $B:=2x^2+x+1$ deux polynômes. On veut montrer que la fonction fraction rationnelle $f:=A/B$ admet un un développement limité en 0 à l'ordre 3 et le calculer.

- Supposons d'abord qu'un tel développement existe : $f(x)=Q(x)+x^3\varepsilon(x)$. On a alors
$$A(x)=B(x)\big(Q(x)+x^3\varepsilon(x)\big)=B(x)Q(x)+x^3B(x)\varepsilon(x)=B(x)Q(x)+x^3\varepsilon_1(x).$$
La fonction $x^3\varepsilon_1(x)=A(x)-B(x)Q(x)$ est un polynôme, on en déduit que $\varepsilon_1(x)$ est un polynôme s'annulant en 0. Par suite $A(x)=B(x)Q(x)+x^4R(x)$, où R est un polynôme. On reconnaît la division de A par B suivant les puissances croissantes à l'ordre 4 (*cf.* le théorème 68 de la page 321).
- Nous disposons maintenant de l'idée qui permet de montrer l'existence d'un développement et de le calculer. On fait la division de A par B suivant les puissances croissantes à l'ordre 4 : $A(x)=B(x)Q(x)+x^4R(x)$, où R est un polynôme. Le calcul donne :
$$1+2x+x^3=(1+x+2x^2)(1+x-3x^2+2x^3)+x^4\left(4(1-x)\right).$$
D'où :
$$f(x)=1+x-3x^2+2x^3+4x^4\frac{1-x}{1+x+2x^2}=1+x-3x^2+2x^3+\mathrm{o}(x^3),$$
puisque $\lim\limits_{x\to0}4\,\frac{1-x}{1+x+2x^2}=4$.

La méthode utilisée dans l'exemple ci-dessus est facile à généraliser.

Proposition 22. Soient A et B deux polynômes avec $B(0)\neq0$. La fonction fraction rationnelle $f:=\frac{A}{B}$ admet un développement limité en 0 à un ordre $n\in\mathbb{N}$ arbitraire. La partie régulière de ce développement est le quotient dans la division du polynôme A par le polynôme B suivant les puissances *croissantes* à l'ordre $n+1$.

Démonstration. La fonction fraction rationnelle $f:=\frac{A}{B}$ est définie au voisinage de 0. On montre qu'elle admet un développement limité en 0 à un ordre $n\in\mathbb{N}$ arbitraire et l'on calcule ce développement en effectuant une division suivant les puissances croissantes de A par B à l'ordre $n+1$: $A(x)=B(x)\,Q(x)+x^{n+1}\,R(x)$. Alors :

$$\frac{A(x)}{B(x)}=Q(x)+x^{n+1}\,\frac{R(x)}{B(x)}=Q(x)+x^n\,\varepsilon(x),\quad\text{avec}\quad\lim_{x\to0}\varepsilon(x)=0,$$

puisque la fraction rationnelle R/B est *bornée* au voisinage de 0. ∎

Plus généralement, soient A et B deux polynômes. Pour que la fonction fraction rationnelle $f := \frac{A}{B}$ admette un développement limité en 0, il est *nécessaire* qu'elle soit bornée au voisinage de 0. Montrons que cette condition est aussi *suffisante*. On note $a_p x^p$ (resp. $b_q x^q$) le monôme non nul de plus bas degré de A (resp. B); alors $f := A/B \sim_0 \frac{a_p}{b_q} x^{p-q}$ (*cf.* l'exercice 4 de la page 865), donc f est bornée au voisinage de 0 si, et seulement si, $p \geqslant q$ (c'est-à-dire si son degré est positif). S'il en est ainsi, on pose $A_1 := x^{-p}A$ et $B_1 := x^{-q}B$; A_1 et B_1 sont des polynômes, $B_1(0) = b_q \neq 0$ et $f = x^{p-q}A_1/B_1$; on peut donc appliquer la proposition ci-dessus à A_1/B_1 et en déduire un développement limité en 0 de f.

4.3 Développements limités à droite et à gauche Développements limités à l'infini

4.3.1 Développements limités à droite et à gauche

Définition 5. On dit que f admet un développement limité à droite (resp. à gauche) à l'ordre n au point a si la restriction de f à $D_f \cap [a, +\infty[$ (resp. à $D_f \cap]-\infty, a]$) admet un développement limité à l'ordre n au point a.

On vérifie facilement que s'il existe $\eta > 0$ tel que $]a - \eta, a + \eta] \setminus \{a\} \subset D_f$, les deux conditions suivantes sont équivalentes :

(i) la fonction f admet un développement limité à l'ordre n au point a ;

(ii) la fonction f admet des développements limités à droite et à gauche à l'ordre n au point a et les parties régulières de ces deux développements sont égales.

Exercice 9.

Montrer que la fonction $f : x \in \mathbb{R} \mapsto \frac{1}{1+|x|^5}$ n'admet pas de développement limité à l'ordre 5 en 0.

Solution. On a $f_{|\mathbb{R}_+}(x) = \frac{1}{1+x^5}$, donc $f_{|\mathbb{R}_+}$ admet en 0 le développement limité à l'ordre 5 : $f_{|\mathbb{R}_+}(x) = 1 + x^5 + \mathrm{o}(x^5)$. On a $f_{|\mathbb{R}_-}(x) = \frac{1}{1-x^5}$, donc $f_{|\mathbb{R}_-}$ admet en 0 le développement limité à l'ordre 5 : $f_{|\mathbb{R}_-}(x) = 1 - x^5 + \mathrm{o}(x^5)$.

Les parties régulières $1 + x^5$ et $1 - x^5$ sont différentes donc f n'admet pas de développement limité à l'ordre 5 en 0.

4.3.2 Développements limités à l'infini.

Rappelons que l'on passe de la notion de développement limité en 0 à celle de développement limité en a et inversement par le changement de variable $x = t + a$.

On peut définir, de façon analogue, la notion de développement limité en $\pm\infty$ pour les fonctions de x, en utilisant le changement de variable $t = 1/x$ et les notions de développements limités à droite et à gauche en 0 pour les fonctions de t.

Soit $f : D_f \to \mathbb{R}$ définie au voisinage de $+\infty$. On note $\Delta := \{t \in \mathbb{R}_+^* \mid 1/t \in D_f\}$. Soit alors $g : t \in \Delta \mapsto f(1/t)$; cette fonction est définie au voisinage de 0. Si elle admet un développement limité d'ordre n à droite en 0 : $g(t) = P(t) + \mathrm{o}(t^n)$, on dira que f admet le développement limité d'ordre n en $+\infty$: $f(x) = P\left(\frac{1}{x}\right) + \mathrm{o}\left(\frac{1}{x^n}\right)$. On est donc amené à la définition suivante.

Définition 6. Soit f une fonction définie au voisinage de $+\infty$ (resp $-\infty$). On dit que f admet un développement limité à l'ordre n en $+\infty$ (resp. en $-\infty$) s'il existe un polynôme P tel que :

$$f(x) = P\left(\frac{1}{x}\right) + \mathrm{o}\left(\frac{1}{x^n}\right) \tag{21}$$

au voisinage de ce point.

Si la fonction f vérifie la définition ci-dessus, on dit aussi qu'elle admet un développement limité à l'ordre n *au voisinage de* $+\infty$ (resp. de $-\infty$).

Exemple. On cherche un développement limité de :

$$f(x) := \frac{1}{x}\left(\sqrt{3+x+x^2} - \sqrt{2-x+x^2}\right)$$

au voisinage de $+\infty$. On pose $t := 1/x$:

$$\begin{aligned}
f(x) &= t\left(\sqrt{3+\frac{1}{t}+\frac{1}{t^2}} - \sqrt{2-\frac{1}{t}+\frac{1}{t^2}}\right)\\
&= \sqrt{1+t+3t^2} - \sqrt{1-t+2t^2}\\
&= (1-u)^{\frac{1}{2}} - (1-v)^{\frac{1}{2}}, \qquad \text{avec } u := -t-3t^2 \text{ et } v := t-2t^2 \to 0\\
&= 1 - \frac{1}{2}u + \frac{1}{8}u^2 + \mathrm{o}(u^2) - 1 + \frac{1}{2}v - \frac{1}{8}v^2 + \mathrm{o}(v^2)\\
&\qquad\qquad \text{(en utilisant les développements de } (1-u)^{-\alpha} \text{ et } (1-v)^{-\alpha})\\
&= \frac{1}{2}t + \frac{3}{2}t^2 + \frac{1}{8}t^2 + \mathrm{o}(t^2) + \frac{1}{2}t - t^2 - \frac{1}{8}t^2 + \mathrm{o}(t^2)\\
&\qquad\qquad (\text{ car } u \sim -t, v \sim t \Rightarrow \mathrm{o}(u^2) = \mathrm{o}(t^2), \mathrm{o}(v^2) = \mathrm{o}(t^2))\\
&= t + \frac{1}{2}t^2 + \mathrm{o}(t^2) = \frac{1}{x} + \frac{1}{2x^2} + \mathrm{o}\left(\frac{1}{x^2}\right).
\end{aligned}$$

Les développements limités à l'infini permettent de trouver des limites (on voit, par exemple, ci-dessus que la limite de f est nulle, et que celle de $xf(x)$ est 1), ou d'obtenir des informations plus précises comme des précisions sur les branches infinies. Ainsi :

$$g(x) := x\left(\sqrt{3+x+x^2} - \sqrt{2-x+x^2}\right) = x^2 f(x) = x + \frac{1}{2} + \mathrm{o}(1)$$

au voisinage de $\pm\infty$. Cette expression n'est pas un développement limité selon notre définition, cependant elle permet de montrer que la droite d'équation $y = x + \frac{1}{2}$ est *asymptote* en $\pm\infty$ au graphe de g.

Exemple. Soient $A := 2x^3 - x^2 + 1$ et $B := x^3 + x^2 + x$. On veut montrer que $f := A/B$ admet un développement limité en $\pm\infty$ à l'ordre 4 et le calculer.

On utilise le changement de variable $x = 1/t$. On pose $\tilde{A}(t) = t^3\,A(1/t) = 2 - t + t^3$ et $\tilde{B}(t) = t^3\,B(1/t) = 1 + t + t^2$; $\tilde{A}$ et $\tilde{B}$ sont des polynômes en t et $\tilde{B}(0) = 0$. On

$$\begin{aligned}
\arctan x &= x - \frac{x^3}{3} + \frac{x^5}{5} + \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + \mathrm{o}(x^{2n+2}) \\
\operatorname{argtanh} x &= x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots + \frac{x^{2n+1}}{2n+1} + \mathrm{o}(x^{2n+2}) \\
\arcsin x &= x + \frac{x^3}{6} + \frac{3x^5}{40} + \cdots + \frac{(2n)!}{2^{2n}(n!)^2}\,\frac{x^{2n+1}}{2n+1} + \mathrm{o}(x^{2n+2}) \\
\operatorname{argsinh} x &= x - \frac{x^3}{6} + \frac{3x^5}{40} + \cdots + (-1)^n \frac{(2n)!}{2^{2n}(n!)^2}\,\frac{x^{2n+1}}{2n+1} + \mathrm{o}(x^{2n+2}) \\
\tan x &= x + \frac{x^3}{3} + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \mathrm{o}(x^8) \\
\tanh x &= x - \frac{x^3}{3} + \frac{2}{15}x^5 - \frac{17}{315}x^7 + \mathrm{o}(x^8).
\end{aligned}$$

TABLE IV.8.3 – Développements limités des fonctions usuelles au voisinage de 0. Ici $n \in \mathbb{N}$ est arbitraire.

a $\tilde{A}(t)/\tilde{B}(t) = A(x)/B(x)$. Nous savons (*cf.* la proposition 22 de la page 877) que $\tilde{A}/\tilde{B}$ admet un développement limité en 0 à tout ordre (en particulier à l'ordre 4). On le calcule par une division suivant les puissances croissantes à l'ordre 5 :

$$2 - t + t^3 = (1 + t + t^2)(2 - 3t + t^2 + 3t^3 - 4t^4) + t^5(1 + 4t)$$

On en déduit :

$$f(x) = 2 - \frac{3}{x} + \frac{1}{x^2} + \frac{3}{x^3} - \frac{4}{x^4} + \frac{1}{x^5}\,\frac{1 + \frac{4}{x}}{1 + \frac{1}{x} + \frac{1}{x^2}} = 2 - \frac{3}{x} + \frac{1}{x^2} + \frac{3}{x^3} - \frac{4}{x^4} + \frac{1}{x^5} + \mathrm{o}\left(\frac{1}{x^4}\right).$$

Il est facile de généraliser le résultat de l'exemple précédent.
Soient A et B deux polynômes avec $\deg A \leqslant \deg B$. D'après l'exercice 4 de la page 865, la fonction fraction rationnelle $f := \frac{A}{B}$ est équivalente à une fonction de la forme $cx^{\deg A - \deg B}$ ($c \neq 0$) en $\pm\infty$, donc la limite de f quand x tend vers $+\infty$ est finie.
On montre que f admet un développement limité en $+\infty$ à un ordre $n \in \mathbb{N}$ arbitraire et l'on calcule ce développement en se ramenant en 0 par le changement de variable $x = 1/t$:

$$A(x)/B(x) = A(1/t)B(1/t) = t^{\deg B - \deg A}\,\frac{\tilde{A}(t)}{\tilde{B}(t)},$$

où $\tilde{A}(t) := t^{\deg A}A(1/t)$ et $\tilde{B}(t) := t^{\deg B}B(1/t)$ sont des polynômes en t, avec $\tilde{B}(0) \neq 0$.
On obtient un développement limité en 0 de $\tilde{A}/\tilde{B}$, en effectuant une division suivant les puissances croissantes de t dont la partie régulière est $Q(t)$. On en déduit un développement limité de f en $\pm\infty$ dont la partie régulière est $x^{\deg B - \deg A}\,Q(1/x)$.

4.4 Opérations sur les développements limités

La table IV.8.3 donne les développements limités, au voisinage de 0, de quelques fonctions transcendantes usuelles. Elle complète la table IV.8.2 de la page 875.

Les opérations sur les développements limités présentées ci-dessous nous permettront de justifier les développements limités figurant dans la table IV.8.3. Ces opérations ont, bien sûr, beaucoup d'autres applications, en particulier pour les calculs de limites et d'équivalents.

4.4.1 Sommes et produits de développements limités

Théorème 23. Soient f et g deux fonctions ayant des développements limités à l'ordre n au point a : $f(x) = P(x-a) + \mathrm{o}(x-a)^n$, $g(x) = Q(x-a) + \mathrm{o}(x-a)^n$. Alors $f+g$, $f-g$, fg ont un développement limité à l'ordre n au point a, de la forme :

$$(f+g)(x) = P(x-a) + Q(x-a) + \mathrm{o}(x-a)^n,$$
$$(f-g)(x) = P(x-a) - Q(x-a) + \mathrm{o}(x-a)^n,$$
$$(fg)(x) = R(x-a) + \mathrm{o}(x-a)^n,$$

où R est le polynôme obtenu en ne gardant dans le produit PQ que les termes de degré au plus égal à n « en $(x-a)$ ».
Si :

$$P(x-a) = \sum_{k=0}^{n} \alpha_k (x-a)^k \quad \text{et} \quad Q(x-a) = \sum_{k=0}^{n} \beta_k (x-a)^k,$$

on a $R(x-a) = \sum\limits_{k=0}^{n} \gamma_m (x-a)^m$, avec, pour tout $m \in [\![0, n]\!]$, $\gamma_m = \sum\limits_{k=0}^{m} \alpha_k \beta_{m-k}$.

Démonstration. Les deux premières formules ont déjà été justifiées (*cf.* la proposition 19 de la page 872).
Pour la multiplication, on a :

$$\begin{aligned}(fg)(x) &= f(x)g(x) = (P(x-a) + \mathrm{o}(x-a)^m)\,(Q(x-a) + \mathrm{o}(x-a)^n)\\ &= P(x-a)Q(x-a) + P(x-a)\mathrm{o}(x-a)^n + Q(x-a)\mathrm{o}(x-a)^n + \mathrm{o}(x-a)^{2n}.\end{aligned}$$

Or P et Q sont continues en tout point (polynômes), donc bornées : $P = \mathrm{O}(1)$, $Q = \mathrm{O}(1)$. Il en résulte que $P(x-a)\mathrm{o}(x-a)^n = \mathrm{o}(x-a)^n$ et $Q(x-a)\mathrm{o}(x-a)^n = \mathrm{o}(x-a)^n$.
Par suite $P(x-a)\mathrm{o}(x-a)^n + Q(x-a)\mathrm{o}(x-a)^n + \mathrm{o}(x-a)^{2n} = \mathrm{o}(x-a)^n$ et $(fg)(x) = P(x-a)Q(x-a) + \mathrm{o}(x-a)^n$. Par définition de R, on a :

$$R(x-a) = (x-a)^{n+1} S(x-a) \qquad \text{donc} \qquad (fg)(x) = R(x-a) + \mathrm{o}(x-a)^n. \qquad \blacksquare$$

Exemple. Au voisinage de 0, on a $\sin x = x - x^3/6 + \mathrm{o}(x^4)$ et $\mathrm{e}^x = 1 + x + x^2/2 + \mathrm{o}(x^2)$ (voir la table IV.8.3 de la page précédente). On en déduit que :

$$\begin{aligned}\sin x + \mathrm{e}^x &= x + \mathrm{o}(x^2) + 1 + x + \frac{x^2}{2} + \mathrm{o}(x^2)\\ &= 1 + 2x + \frac{x^2}{2} + \mathrm{o}(x^2).\end{aligned}$$

De même :

$$\begin{aligned}
\sin x\,\mathrm{e}^x &= \left(x - \frac{x^3}{6} + \mathrm{o}(x^4)\right)\left(1 + x + \frac{x^2}{2} + \mathrm{o}(x^2)\right)\\
&= x\left(1 - \frac{x^2}{6} + \mathrm{o}(x^3)\right)\left(1 + x + \frac{x^2}{2} + \mathrm{o}(x^2)\right)\\
&= x + x^2 + \frac{x^3}{2} + \mathrm{o}(x^3) - \frac{x^3}{6} - \frac{x^4}{6} - \frac{x^5}{12} + \mathrm{o}(x^5)\\
&\qquad\qquad + \mathrm{o}(x^4) + \mathrm{o}(x^5) + \mathrm{o}(x^6) + \mathrm{o}(x^6)\\
&= x + x^2 + \frac{x^3}{3} + \mathrm{o}(x^3).
\end{aligned}$$

À la troisième ligne, on a développé la seconde, en tenant compte des règles sur les petits-o données précédemment, comme $x\mathrm{o}(x^2) = \mathrm{o}(x^3)$ ou $-\frac{x^3}{6}\mathrm{o}(x^2) = \mathrm{o}(x^5)$. On voit dans la quatrième que la plupart des termes sont négligeables devant le petit-o de plus faible puissance (ici $\mathrm{o}(x^3)$) qui « absorbe » ceux-ci et les autres petits-o. Dans la pratique, il est utile d'anticiper cela et *de ne pas calculer tous les termes du produit*. On remarque aussi dans ce cas que nous avons obtenu un développement limité à l'ordre 3, alors que le théorème ne nous garantissait que l'ordre 2. Cette bonne surprise provient de l'annulation du sinus au point considéré. On peut généraliser en utilisant la remarque suivante.

Remarque.
On reprend les notations du théorème 23. On note $\alpha_p(x-a)^p$ (resp. $\beta_q(x-a)^q$) le monôme non nul de plus bas degré de $P(x-a)$ (resp. $Q(x-a)$). Pour obtenir le développement limité à l'ordre n de fg, il suffit de disposer du développement limité de f à l'ordre $n-q$ et de celui de g à l'ordre $n-p$.

4.4.2 Intégration de développements limités

On a :

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \cdots + (-1)^n x^n + \mathrm{o}(x^n), \tag{22}$$

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots + (-1)^n x^{2n} + \mathrm{o}(x^{2n+1}). \tag{23}$$

Pour obtenir (23), on remplace x par x^2 dans (22), puis l'on utilise la parité.
Les deux fonctions $\frac{1}{1+x}$ et $\frac{1}{1+x^2}$ sont des dérivées de fonctions transcendantes classiques (respectivement $\ln(1+x)$ et $\arctan x$). On peut obtenir un développement limité de ces dernières grâce au théorème suivant.

Théorème 24. Soit f une fonction continue sur un intervalle I, admettant un développement limité $f(x) = P(x-a) + \mathrm{o}(x-a)^n$ en un point $a \in I$, avec $P(x-a) = \alpha_0 + \alpha_1(x-a) + \cdots + \alpha_n(x-a)^n$. Alors toute primitive F de f a un développement limité au même point, de la forme :

$$F(x) = F(a) + \alpha_0(x-a) + \alpha_1\frac{(x-a)^2}{2} + \cdots + \alpha_n\frac{(x-a)^{n+1}}{n+1} + \mathrm{o}(x-a)^{n+1}. \tag{24}$$

En d'autres termes, on peut primitiver un développement limité, et l'on gagne un ordre de développement.
En particulier, l'application du théorème 24 à (22) et (23) respectivement donne les développements limités de $\ln(1+x)$ et $\arctan x$ figurant dans la table IV.8.3.

Démonstration.

- On considère d'abord un cas particulier. On suppose que la partie régulière du développement limité de f en a est nulle ($P=0$) et que $F(a)=0$. Alors $f=(x-a)^n\varphi$, avec $\lim\limits_{x\to a}\varphi(x)=0$. Montrons que F admet en a un développement limité à l'ordre $n+1$ dont la partie régulière est nulle. On a $F(x)=\displaystyle\int_a^x f(t)dt$. Soit $\varepsilon>0$, il existe $\eta>0$ tel que, si $|t-a|\leqslant\eta$, alors $|\varphi(t)|\leqslant\varepsilon$ et $|f(t)|\leqslant\varepsilon|t-a|^n\leqslant\varepsilon|x-a|^n$. Alors :
$$\big|F(x)\big|\leqslant\left|\int_a^x f(t)\,\mathrm{d}t\right|\leqslant\varepsilon|x-a|^n\left|\int_a^x \mathrm{d}t\right|=\varepsilon|x-a|^{n+1}.$$
Ainsi $F=\mathrm{o}(x-a)^{n+1}$.
- Passons au cas général. Soient $g(x):=f(x)-P(x-a)$ et G la primitive de g qui s'annule en a. On a $g(x)=\mathrm{o}(x-a)^n$ et, d'après ce qui précède, $G(x)=\mathrm{o}(x-a)^{n+1}$. Donc :
$$\begin{aligned}F(x)-F(a)&=\int_a^x f(t)\,\mathrm{d}t=\sum_{k=0}^n\alpha_k\frac{(x-a)^{k+1}}{k+1}+\int_a^x g(t)\,\mathrm{d}t\\&=\sum_{k=0}^n\alpha_k\frac{(x-a)^{k+1}}{k+1}+G(x)=\sum_{k=0}^n\alpha_k\frac{(x-a)^{k+1}}{k+1}+\mathrm{o}(x-a)^{n+1}.\end{aligned}$$ ■

Toutes les fonctions transcendantes dont la dérivée est une fraction rationnelle « simple » n'ayant pas 0 pour pôle peuvent être développées de cette façon.
Voici un exemple plus compliqué, celui de la fonction $\arcsin$. Son développement limité en 0 est donné dans la table IV.8.3 de la page 880, nous allons expliquer comment l'obtenir. La dérivée de $x\mapsto\arcsin x$ est $x\mapsto 1/\sqrt{1-x^2}=(1-x^2)^{-1/2}$. Cette dernière fonction est la composée de $y\mapsto(1-y)^{-1/2}$ et de $x\mapsto x^2$. On obtient le développement de la première dans la table, en utilisant $\alpha=1/2$:
$$(1-y)^{-\frac12}=1+\frac12 y+\frac12\,\frac32\,\frac{y^2}{2}+\cdots+\frac12\,\frac32\cdots\frac{2n-1}{2}\,\frac{y^n}{n!}+\mathrm{o}(y^n).$$
Le coefficient qui intervient devant y^n, que nous noterons c_n, peut être « simplifié » en multipliant par les nombres pairs qui manquent entre les impairs :
$$c_n=\frac12\,\frac32\cdots\frac{2n-1}{2}\,\frac{1}{n!}=\frac{1.3.5\ldots(2n-1)}{2^n n!}=\frac{(2n)!}{2.4.6\ldots(2n)\,2^n n!}=\frac{(2n)!}{2^{2n}(n!)^2}.$$
En posant $y=x^2$, l'on reste au voisinage de 0 et en tenant compte du fait que la fonction est paire, on trouve :
$$(1-x^2)^{-1/2}=c_0+c_1x^2+\cdots+c_nx^{2n}+\mathrm{o}(x^{2n+1}).$$
Il ne reste plus qu'à primitiver pour obtenir le développement limité de $\arcsin x$.

4.4.3 Divisions de développements limités

Une autre opération importante sur les développements limités est la division.
Nous allons d'abord traiter un cas particulier, celui de la fonction tangente au point $a = 0$.

Exemple.
On part des développements limités du sinus et du cosinus donnés dans la table IV.8.2 :

$$\tan x = \frac{\sin x}{\cos x} = \frac{x - \frac{x^3}{6} + \mathrm{o}(x^4)}{1 - \frac{x^2}{2} + \mathrm{o}(x^3)}.$$

Le dénominateur est de la forme $1-T$, en posant $T := \dfrac{x^2}{2} + \mathrm{o}(x^3)$. En particulier, lorsque x tend vers 0, T aussi, et par conséquent on peut utiliser la formule (17).
Comme $T = \mathrm{O}(x^2)$, on a $T^2 = \mathrm{O}(x^4) = \mathrm{o}(x^3)$ et donc toutes les puissances de T supérieures ou égales à 2 sont absorbées par le $\mathrm{o}(x^3)$ apparaissant dans T :

$$\begin{aligned}\tan x &= \left(x - \frac{x^3}{6} + \mathrm{o}(x^4)\right)\left(1 + T + \mathrm{O}(T^2)\right)\\ &= \left(x - \frac{x^3}{6} + \mathrm{o}(x^4)\right)\left(1 + \frac{x^2}{2} + \mathrm{o}(x^3)\right) = x + \frac{x^3}{3} + \mathrm{o}(x^4).\end{aligned}$$

En partant de développements plus précis du sinus et cosinus, on peut obtenir la formule donnée dans la table IV.8.3 de la page 880. Nous l'établirons plus loin par une autre méthode. On peut en fait obtenir explicitement le développement limité de $\tan$ en 0 à un ordre arbitraire (29) en utilisant des nombres rationnels appelés nombres de Bernoulli (*cf.* l'exercice 11 de la page 886).

Théorème 25. Soient f, g deux fonctions ayant au point a un développement limité à l'ordre n. On suppose en outre que $g(a) \neq 0$. Alors :

(i) f/g a un développement limité à l'ordre n au point a ;

(ii) si $a := 0$, la partie régulière du développement limité de f/g en 0 est le quotient, dans la division suivant les puissances croissantes, des parties régulières de f et g.

Démonstration. Les preuves sont de simples variantes des méthodes données pour les fractions rationnelles dans la partie 4.2.2.

(i) Il suffit de faire la preuve dans le cas où $a := 0$. Par hypothèse $g(0) \neq 0$, donc nous pouvons multiplier le développement limité de g par $1/g(0)$, obtenant ainsi une formule du type :

$$\frac{g(x)}{g(0)} = 1 + \gamma_1 x + \cdots + \gamma_n x^n + \mathrm{o}(x^n).$$

Notons $T := 1 - g(0)^{-1} g = -\sum_{k=1}^{n} \gamma_k x^k$. Alors $T = x\,\mathrm{O}(1)$.

On a :

$$\frac{1}{1-T} = 1 + T + T^2 + \cdots + T^n + \frac{T^{n+1}}{1-T}, \quad \text{et} \quad \frac{T^{n+1}}{1-T} = \mathrm{o}(T^n) = \mathrm{o}(x^n). \quad (25)$$

On en déduit que :

$$\begin{aligned}\frac{1}{1-T} &= 1-(\gamma_1 x+\cdots+\gamma_n x^n)+(\gamma_1 x+\cdots+\gamma_n x^n)^2+\cdots\\ &\qquad +(-1)^n(\gamma_1 x+\cdots+\gamma_n x^n)^n+\mathrm{o}(x^n)\\ &= P(x)+\mathrm{o}(x^n).\end{aligned}$$

En supprimant les termes de degré strictement supérieur à n dans le polynôme $P(x)$, on obtient un polynôme $Q(x)$ et l'on a : $\frac{1}{1-T} = Q(x)+\mathrm{o}(x^n)$, avec $\deg Q \leqslant n$. C'est le développement à l'ordre n de $\frac{1}{1-T}$. En le multipliant par le développement de $f(x)/g(0)$ on obtient celui de f/g .

(ii) Soient :

$$f(x) = A(x)+\mathrm{o}(x^n) \qquad \text{et} \qquad g(x) = B(x)+\mathrm{o}(x^n)$$

les développements limités à l'ordre n en 0 de f et g. On a $B(0) \neq 0$. On va donner une autre preuve de (i) et calculer la partie régulière du développement limité de f/g.

On effectue la division suivant les puissances croissantes de A par B à l'ordre $n+1$: $A(x) = B(x)\,Q(x)+x^{n+1}\,R(x)$. Montrons que $Q(x)$ est la partie régulière du développement limité de f/g à l'ordre n.

Remarquons tout d'abord que :

$$g(x) = B(x)\left(1+\frac{\mathrm{o}(x^n)}{B(x)}\right) \quad \text{et} \quad f(x) = B(x)\left(Q(x)+\frac{x^{n+1}\,R(x)+\mathrm{o}(x^n)}{B(x)}\right).$$

Comme $B(0) \neq 0$, la fonction $1/B$ est bornée au voisinage de 0, ce qui donne :

$$\frac{\mathrm{o}(x^n)}{B(x)} = \mathrm{o}(x^n) \qquad \text{ainsi que} \qquad \frac{x^{n+1}\,R(x)+\mathrm{o}(x^n)}{B(x)} = \mathrm{o}(x^n).$$

On obtient donc :

$$\frac{f(x)}{g(x)} = \frac{Q(x)+\mathrm{o}(x^n)}{1+\mathrm{o}(x^n)} = Q(x)+\mathrm{o}(x^n).$$

■

Revenons à la preuve de (i). Comme on l'a vu dans la démonstration, l'idée consiste à mettre $g(0)$ en facteur, et à écrire le dénominateur sous la forme $1-T$ avant d'utiliser la formule (25). Cette idée se généralise même dans le cas où $g(0) = 0$. Il faut alors factoriser le terme « dominant » du développement limité du dénominateur (c'est-à-dire le premier terme non nul, si ceux-ci ont été rangés par puissances croissantes). Le terme factorisé doit « se simplifier » avec le numérateur, sinon la fonction f/g n'est pas bornée au voisinage de 0, et ne peut donc pas avoir de développement limité. On peut ainsi se ramener au cas $g(0) \neq 0$.

Exemple. Voici un exemple : le calcul du développement limité en 0, à l'ordre 4, de $\sinh^2 x/(1-\cos x)$. Le terme dominant dans $1-\cos x$ est $x^2/2$; c'est donc lui que l'on

factorise :

$$\begin{aligned}
\frac{\sinh^2 x}{1-\cos x} &= \frac{\left(x+\frac{x^3}{6}+\frac{x^5}{5!}+\mathrm{o}(x^5)\right)^2}{\frac{x^2}{2}-\frac{x^4}{24}+\frac{x^6}{720}+\mathrm{o}(x^6)} = \frac{x^2+\frac{x^4}{3}+\frac{2}{45}x^6+\mathrm{o}(x^6)}{\frac{x^2}{2}\left(1-\frac{x^2}{12}+\frac{x^4}{360}+\mathrm{o}(x^4)\right)} \\
&= \frac{2+\frac{2}{3}x^2+\frac{4}{45}x^4+\mathrm{o}(x^4)}{1-T}, \qquad \left(\text{avec} \quad T := \frac{x^2}{12}-\frac{x^4}{360}+\mathrm{o}(x^4)\right) \\
&= \left(2+\frac{2}{3}x^2+\frac{4}{45}x^4+\mathrm{o}(x^4)\right)\left(1+T+T^2+\mathrm{o}(T^2)\right) \\
&\qquad \text{(par la formule (17))} \\
&= \left(2+\frac{2}{3}x^2+\frac{4}{45}x^4+\mathrm{o}(x^4)\right)\left(1+\frac{x^2}{12}-\frac{x^4}{360}+\frac{x^4}{144}+\mathrm{o}(x^4)\right) \\
&\qquad \text{(car} \quad T=\mathrm{O}(x^2)\text{, donc } \mathrm{o}(T^2)=\mathrm{o}(x^4)) \\
&= 2+\frac{5}{6}x^2+\frac{11}{72}x^4+\mathrm{o}(x^4).
\end{aligned}$$

La fonction étant paire, on pourrait remplacer $\mathrm{o}(t^4)$ par $\mathrm{o}(t^5)$ dans le résultat final.
Noter l'ordre des développements limités des fonction $\sinh$ et $\cos$ dont on a eu besoin pour obtenir à la fin l'ordre 4. Cela peut facilement se deviner en anticipant la simplification par t^2 et donc la nécessité d'un ordre 6 au numérateur et au dénominateur (l'ordre 5 pour le sinus hyperbolique suffit pour avoir $\sinh^2 t$ à l'ordre 6).

Exercice 10.

Déterminer le développement limité de $x \mapsto \tan x$ à l'ordre 8 en 0 en utilisant une division suivant les puissances croissantes.

Solution. Comme la fonction tangente est impaire, il suffit de déterminer le développement limité à l'ordre 7. On a :

$$\sin(x) = x-\frac{x^3}{6}+\frac{x^5}{120}-\frac{x^7}{5040}+\mathrm{o}(x^7) \quad \text{et} \quad \cos(x) = 1-\frac{x^2}{2}+\frac{x^4}{24}-\frac{x^6}{720}+\mathrm{o}(x^7).$$

On pose la division suivant les puissances croissantes en ne gardant que les termes de degré inférieur ou égal à 7 (voir la table IV.8.4 de la page suivante).

On en déduit : $\tan x = x+\dfrac{1}{3}x^3+\dfrac{2}{15}x^5+\dfrac{17}{315}x^7+\mathrm{o}(x^8)$.

On montrerait, par un calcul similaire : $\tanh x = x-\dfrac{1}{3}x^3+\dfrac{2}{15}x^5-\dfrac{17}{315}x^7+\mathrm{o}(x^8)$.

L'exercice suivant permet de calculer le développement limité de $\tanh x$ à un ordre arbitraire en utilisant les nombres de Bernoulli. Pour une autre application de ces nombres, *cf.* l'exercice IV.1.61 de la page 612.

Exercice 11.

1. a) Montrer que la fonction définie par $f : x \in \mathbb{R}^* \mapsto \frac{x}{2}+\frac{x}{\mathrm{e}^x-1}$ et $f(0)=1$ est paire. Montrer que $f(x) = \frac{x}{2}\operatorname{cotanh}\frac{x}{2}$.

$$
\begin{array}{rrrr|l}
x & -\dfrac{x^3}{6} & +\dfrac{x^5}{120} & -\dfrac{x^7}{5040} & 1-\dfrac{x^2}{2}+\dfrac{x^4}{24}-\dfrac{x^6}{720} \\ \cline{5-5}
-\Big(x & -\dfrac{x^3}{2} & +\dfrac{x^5}{24} & -\dfrac{x^7}{720}\Big) & x+\dfrac{x^3}{3}+\dfrac{2x^5}{15}+\dfrac{17x^7}{315} \\ \cline{1-4}
 & \dfrac{x^3}{3} & -\dfrac{x^5}{30} & +\dfrac{x^7}{840} & \\
 & -\Big(\dfrac{x^3}{3} & -\dfrac{x^5}{6} & +\dfrac{x^7}{72}\Big) & \\ \cline{2-4}
 & & \dfrac{2x^5}{15} & -\dfrac{4x^7}{315} & \\
 & & -\Big(\dfrac{2x^5}{15} & -\dfrac{x^7}{15}\Big) & \\ \cline{3-4}
 & & & \dfrac{17x^7}{315} &
\end{array}
$$

TABLE IV.8.4 – Division suivant les puissances croissantes

b) Montrer que la fonction f définie en a) admet un développement limité à tout ordre en 0. Pour tout $p \in \mathbb{N}^*$, on écrit la partie régulière à l'ordre $2p$ sous la forme $\sum\limits_{k=1}^{p} (-1)^{k-1} \frac{B_k}{(2k)!} x^{2k}$.

Montrer que les nombres B_k ainsi définis sont *rationnels*[1].

c) Montrer que, pour tout $n \in \mathbb{N}^*$, $n \geqslant 2$:

$$\sum_{k=1}^{n-1} \binom{2n}{2k} (-1)^{k+1} B_k = n-1 \tag{26}$$

et pour tout $n \in \mathbb{N}^*$:

$$\sum_{k=1}^{n} \binom{2n+1}{2k} (-1)^{k+1} B_k = n - \frac{1}{2}\cdot \tag{27}$$

Calculer B_k pour $k \in [\![1, 8]\!]$.

2. a) Montrer que $\tanh t = 2 \operatorname{cotanh} 2t - \operatorname{cotanh} t$.

b) Donner les formules des développements limités de $\tanh$ en 0 à un ordre arbitraire en utilisant les nombres B_n.

« Deviner » les formules analogues pour $\tan$.

[1] Ils sont appelés *nombres de Bernoulli*. Introduits par Jacques Bernoulli vers 1700 pour calculer la somme des puissances $k^{\text{èmes}}$ des n premiers entiers, ils ont été ainsi baptisés plus tard par Abraham de Moivre.

Solution.

1. a) On a, pour tout $x \in \mathbb{R}^*$, $f(x) = \dfrac{x}{2}\dfrac{\mathrm{e}^x+1}{\mathrm{e}^x-1}$, d'où :

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

On a :

$$f(x) = \frac{x}{2}\frac{\mathrm{e}^x+1}{\mathrm{e}^x-1} = \frac{x}{2}\frac{\mathrm{e}^{x/2}+\mathrm{e}^{-x/2}}{\mathrm{e}^{x/2}-\mathrm{e}^{-x/2}} = \frac{x}{2}\operatorname{cotanh}\frac{x}{2}\cdot$$

b) On a $f - \frac{x}{2} = \frac{1}{1-T}$, avec $T(x) = \frac{1+x-\mathrm{e}^x}{x} = -\frac{x}{2!} - \frac{x^3}{3!} - \cdots - \frac{x^n}{(n+1)!} + \mathrm{o}(x^n)$. Ainsi la partie régulière de T est, à tout ordre, un polynôme à coefficients rationnels. Par suite, pour tout $k \in \mathbb{N}^*$, la partie régulière de T^k est aussi, à tout ordre, un polynôme à coefficients rationnels. Le résultat s'en déduit.

c) On définit $\Phi : \mathbb{R} \to \mathbb{R}$ par $\Phi(x) := f(x) - \dfrac{x}{2} = \dfrac{x}{\mathrm{e}^x-1}$, pour $x \in \mathbb{R}^*$ et $\Phi(0) := 1$ et l'on note[2] $\displaystyle\sum_{k=0}^{n} \frac{b_k}{k!}x^k$ la partie régulière à l'ordre n du développement limité de Φ en 0. On a $b_0 = 1$, $b_1 = -\frac{1}{2}$, et, pour tout $n \in \mathbb{N}^*$, $b_{2n} = (-1)^{n+1}b_n$, $b_{2n+1} = 0$. De $\Phi(x)(\mathrm{e}^x - 1) = x$ l'on déduit, en utilisant le produit des développements limités :

$$0 = \sum_{k=0}^{m-1} \frac{b_k}{k!(m-k)!} = \frac{1}{m!}\sum_{k=0}^{m-1}\binom{m}{k}b_k, \tag{28}$$

donc $\sum\limits_{k=0}^{m-1}\binom{m}{k}b_k = 0$ et $\sum\limits_{k=2}^{m-1}\binom{m}{k}b_k = -b_0 - \binom{m}{1}b_1 = -b_0 - mb_1$.

En séparant les cas pairs et impairs ($m = 2n$, $m = 2n+1$) et en remplaçant les b_k par leurs valeurs, on obtient les deux identités (26) et (27). Ces identités permettent de calculer les B_n par récurrence. On en déduit (sans difficulté mais moyennant un peu de patience ou en utilisant un système de calcul formel) :

$$B_1 = \frac{1}{6} \qquad B_2 = \frac{1}{30} \qquad B_3 = \frac{1}{42} \qquad B_4 = \frac{1}{30}$$
$$B_5 = \frac{5}{66} \qquad B_6 = \frac{691}{2730} \qquad B_7 = \frac{7}{6} \qquad B_8 = \frac{3617}{510}\cdot$$

2. a) Notons $\tau := \tanh t$. D'après les formules (13) et (14) de la page 700 on a :

$$2\operatorname{cotanh} 2t - \operatorname{cotanh} t = 2\frac{1+\tau^2}{2\tau} - \frac{1}{\tau} = \tau.$$

b) Pour tout $p \in \mathbb{N}^*$, on a :

$$x\operatorname{cotanh} x = f(2x) = 1 + \sum_{k=1}^{p}(-1)^{k+1}\frac{B_k}{(2k)!}2^{2k}x^{2k} + \mathrm{o}(x^{2p+1})$$

[2] On trouve dans certains ouvrages une définition des nombres de Bernoulli différente de la nôtre : « $B_n := b_{2n}$ ».

et, en remplaçant x par $2x$ dans ce développement limité :

$$2x \operatorname{cotanh} 2x = f(2x) = 1 + \sum_{k=1}^{p} (-1)^{k+1} \frac{B_k}{(2k)!} 2^{4k} x^{2k} + \mathrm{o}(x^{2p+1})$$

En utilisant $\tanh x = 2 \operatorname{cotanh} 2x - \operatorname{cotanh} x$, on obtient, pour tout $p \in \mathbb{N}^*$:

$$\tanh x = \sum_{k=1}^{p} (-1)^{k+1} 2^{2k} (2^{2k} - 1) \frac{B_k}{(2k)!} x^{2k-1} + \mathrm{o}(x^{2p}).$$

En regardant la table IV.8.3, on devine que, pour tout $p \in \mathbb{N}^*$:

$$\tan x = \sum_{k=1}^{p} 2^{2k} (2^{2k} - 1) \frac{B_k}{(2k)!} x^{2k-1} + \mathrm{o}(x^{2p}). \tag{29}$$

Ce résultat peut être justifié en utilisant la relation :

$$\sum_{k=1}^{p} (-1)^{k+1} 2^{2k} (2^{2k} - 1) \frac{B_k}{(2k)!} (ix)^{2k-1} = i \sum_{k=1}^{p} 2^{2k} (2^{2k} - 1) \frac{B_k}{(2k)!} x^{2k-1}$$

(*cf.* l'exercice IV.8.18 de la page 908).

Exercice 12.

Montrer que la fonction $f : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \arcsin \frac{x^2}{\sqrt{2(x^4-2x^2+2)}} \end{array}$ est bien définie.

Trouver un développement limité de f à l'ordre 10 au voisinage de 0.

Solution. On a, pour tout $x \in \mathbb{R}$, $x^4 - 2x^2 + 2 = (x^2 - 1)^2 + 1 > 0$.

On note $g(x) := \frac{x^2}{\sqrt{2(x^4-2x^2+2)}}$. C'est une fonction paire continûment dérivable sur $\mathbb{R}$.

Pour tout $x \in \mathbb{R}$, $1 - g(x)^2 = \frac{(2-x^2)^2}{2(x^4-2x^2+2)} > 0$, donc, pour tout $x \in \left]-\sqrt{2}, \sqrt{2}\right[$, l'on a $|g(x)| < 1$. Ainsi f est paire et continûment dérivable sur $\left]-\sqrt{2}, \sqrt{2}\right[$.

Calculons f'. On a $f'(x) = \frac{g'(x)}{\sqrt{1-g(x)^2}}$. On note $u := x^4 - 2x^2 + 2$.

On a $\sqrt{1 - g(x)^2} = \frac{2-x^2}{\sqrt{2u}}$ et $g' = \frac{\sqrt{2}}{u^{3/2}}(2x - x^3)$, donc $f'(x) = \frac{2x}{u} = \frac{2x}{x^4-2x^2+2}$.

La fonction f' est impaire.

On obtient un développement limité de f' en 0 à l'ordre 9 par une division suivant les puissances croissantes :

$$f'(x) = x + x^3 + \frac{x^5}{2} - \frac{x^9}{4} + \mathrm{o}(x^9).$$

D'où, par intégration, en utilisant $f(0) = 0$:

$$f(x) = \frac{x^2}{2} + \frac{x^4}{4} + \frac{x^6}{12} - \frac{x^{10}}{40} + \mathrm{o}(x^{10}).$$

4.4.4 Composition de développements limités

Théorème 26. Soit $n \in \mathbb{N}$. Soit f une fonction ayant un développement limité à l'ordre n au point a : $f(x) = P(x-a)+\mathrm{o}_a(x-a)^n$. Soient $b = f(a)$ et g une fonction ayant un développement limité au point b, à l'ordre n : $g(y) = Q(y-b) + \mathrm{o}_b(y-b)^n$. Alors $g \circ f$ a un développement limité à l'ordre n au point a de la forme $g \circ f(x) = R(x-a) + \mathrm{o}_a(x-a)^n$, où $R(x-a)$, est le polynôme (en $(x-a)$) obtenu en ne gardant dans $Q\big(P(x-a)-b\big)$ que les monômes $c_p(x-a)^p$ de degré p au plus égal à n.

Démonstration. Posons $y = f(x)$. Puisque f a un développement limité, elle est continue, et de ce fait, lorsque x tend vers a, y tend vers b. Ainsi :

$$g \circ f(x) = Q\big(f(x)-b\big) + \mathrm{o}_a(f(x)-b)^n.$$

Posons $P(x-a) = \alpha_0 + \cdots + \alpha_n(x-a)^n$, et notons d'abord que $\alpha_0 = P(a) = f(a) = b$. Si $n \geqslant 1$, $f(x) - b = \alpha_1(x-a) + \cdots + \alpha_n(x-a)^n + \mathrm{o}(x-a)^n$ est un $\mathrm{O}(x-a)$. De ce fait, $\mathrm{o}_a\,(f(x)-b)^n = \mathrm{o}_a(x-a)^n$.

Posons de même $Q(y-b) = \beta_0 + \cdots + \beta_n(y-b)^n$, de sorte que $Q\big(f(x)-b\big)$ est une somme de termes de la forme $\beta_k\big(P(x-a) + \mathrm{o}(x-a)^n\big)^k$. Si $k \geqslant 1$, ce terme est de la forme $\beta_k P(x-a)^k + \mathrm{o}(x-a)^n$, en appliquant k fois la règle de multiplication des développements limités.

Finalement, on trouve :

$$\begin{aligned} g \circ f(x) &= \beta_0 + \cdots + \beta_n P(x-a)^n + \mathrm{o}(x-a)^n + \mathrm{o}(x-a)^n \\ &= Q\big(P(x-a)-b\big) + \mathrm{o}_a(x-a)^n. \end{aligned}$$

Par définition :

$$Q\big(P(x-a)-b\big) = R(x-a) + (x-a)^{n+1}S(x-a),$$

donc $g \circ f(x) = R(x-a) + \mathrm{o}_a(x-a)^n$. ∎

Remarque. Supposons $a = b = 0$. Si le monôme de plus bas degré de P est $\alpha_p x^p$, avec $p \geqslant 2$, on a $f \sim \alpha_p x^p$ et $\mathrm{o}(f^q) = \mathrm{o}(x^{pq})$. Pour avoir un développement de $g \circ f$ à l'ordre n, il *suffit* donc de développer g à un ordre q tel que $pq \geqslant n$.

Exemple. Développement limité à l'ordre 4 en 0 de $\varphi : x \mapsto \ln\left(\frac{\sinh x}{x}\right)$.

On note $f(x) := \frac{\sinh x}{x} - 1$ et $g(y) := \ln(1+y)$. On a $\varphi = g \circ f$. On écrit le développement limité à l'ordre 4 en 0 de f : $f(x) = \frac{1}{6}x^2 + \frac{1}{120}x^4 + \mathrm{o}(x^4)$. On utilise la remarque ci-dessus. Le monôme de plus bas degré de ce développement est $\frac{1}{6}x^2$, donc $\mathrm{o}(f^2) = \mathrm{o}(x^4)$ et il suffit de développer g à l'ordre 2 : $\ln(1+y) = y - \frac{1}{2}y^2 + \mathrm{o}(y^2)$. On obtient :

$$\varphi(x) = \left(\frac{1}{6}x^2 + \frac{1}{120}x^4\right) - \frac{1}{2}\left(\frac{1}{6}x^2 + \frac{1}{120}x^4\right)^2 + \mathrm{o}(x^4) = \frac{1}{6}x^2 - \frac{1}{180}x^4 + \mathrm{o}(x^4).$$

Exercice 13.
Calculer le développement limité à l'ordre 6 en 0 de $\varphi : x \mapsto \frac{1}{\cos x}$.
Solution. On peut utiliser le développement limité d'un quotient en effectuant une division suivant les puissances croissantes. Nous allons employer un argument, plus rapide, de fonction composée (on pourra comparer avec l'exemple de la page 876).
On pose $f(x) := 1 - \cos x$ et $g(y) := \frac{1}{1-y}$. On a $\varphi(x) = g \circ f(x)$. On écrit le développement limité à l'ordre 6 en 0 de f : $f(x) = \frac{1}{2}x^2 - \frac{1}{24}x^4 + \frac{1}{720}x^6 + \mathrm{o}(x^6)$. On utilise la remarque ci-dessus. Le monôme de plus bas degré de ce développement est $\frac{1}{2}x^2$, donc $\mathrm{o}\big(f(x)^3\big) = \mathrm{o}(x^6)$ et il suffit de développer g à l'ordre 3 : $\frac{1}{1-y} = 1 + y + y^2 + y^3 + \mathrm{o}(y^3)$.
On obtient :

$$\begin{aligned}\varphi(x) &= 1 + \left(\frac{1}{2}x^2 - \frac{1}{24}x^4 + \frac{1}{720}x^6\right) + \left(\frac{1}{2}x^2 - \frac{1}{24}x^4 + \frac{1}{720}x^6\right)^2 \\ &\quad + \left(\frac{1}{2}x^2 - \frac{1}{24}x^4 + \frac{1}{720}x^6\right)^3 + \mathrm{o}(x^6) \\ &= 1 + \frac{1}{2}x^2 + \frac{5}{24}x^4 + \frac{61}{720}x^6 + \mathrm{o}(x^6).\end{aligned}$$

4.5 Calculs de limites et d'équivalents

Les développements limités sont très utiles pour calculer des limites ou pour trouver des équivalents simples de certaines fonctions (souvent appelés *parties principales* de ces fonctions). On utilise essentiellement le point méthode de la page 872. En pratique, le principal problème est de trouver à quel ordre il faut calculer les développements limités : un ordre trop bas ne permet pas de conclure et un ordre trop haut entraîne des calculs inutiles. Nous allons donner quelques exemples (*cf.* aussi l'exercice type corrigé IV.8.0 de la page 903).

Exemple. Cherchons la limite, quand x tend vers 0, de :

$$f(x) := \frac{2(1-\cos x)\sin x - x^3(1-x^2)^{1/4}}{\sin^5 x - x^5}.$$

On pose $\varphi(x) = 2(1-\cos x)\sin x - x^3(1-x^2)^{1/4}$ et $\psi(x) = \sin^5 x - x^5$.
On a $\lim\limits_{x\to 0} \varphi(x) = 0$ et $\lim\limits_{x\to 0} \psi(x) = 0$. Dans ce cas l'on dit que l'on a pour $f = \frac{\varphi}{\psi}$ une *forme indéterminée* du type $\frac{0}{0}$.
On cherche des équivalents simples du numérateur φ et du dénominateur ψ. Commençons par le cas de ψ, qui est plus simple. On écrit $\psi(x) = x^5\left(\left(\frac{\sin x}{x}\right)^5 - 1\right)$ et l'on fait un développement limité à l'ordre 2 de $(\sin x/x)^5$: $\left(\frac{\sin x}{x}\right)^5 = \left(1 - \frac{x^2}{6}\right)^5 + \mathrm{o}(x^2)$. On en déduit :

$$\psi(x) = x^5\left(\left(1 - \frac{x^2}{6}\right)^5 - 1 + \mathrm{o}(x^2)\right) = x^5\left(-\frac{5x^2}{6}\right) + \mathrm{o}(x^7),$$

et $\psi \sim -\frac{5}{6}x^7$.

Compte tenu du résultat précédent, on cherche ensuite un développement limité *à l'ordre* 7 de φ. Si $x \neq 0$, on peut écrire :

$$2(1-\cos x)\sin x = x^3 \frac{2(1-\cos x)}{x^2}\frac{\sin x}{x}.$$

Nous sommes ainsi ramenés à considérer les développements limités *à l'ordre* 4 de $\frac{2(1-\cos x)}{x^2}$, $\frac{\sin x}{x}$ et $(1-x^2)^{1/4}$:

$$\frac{2(1-\cos x)}{x^2} = 1 - \frac{x^2}{12} + \frac{x^4}{360} + \mathrm{o}(x^4), \qquad \frac{\sin x}{x} = 1 - \frac{x^2}{6} + \frac{x^4}{120} + \mathrm{o}(x^4)$$

$$(1-x^2)^{1/4} = 1 + \frac{1}{4}(-x^2) + \frac{1}{4}\left(-\frac{3}{4}\right)\frac{(-x^2)^2}{2!} = 1 - \frac{x^2}{4} - \frac{3x^4}{32} + \mathrm{o}(x^4).$$

On en déduit, par multiplication : $\frac{2(1-\cos x)}{x^2}\frac{\sin x}{x} = 1 - \frac{x^2}{4} + \frac{x^4}{40} + \mathrm{o}(x^4)$, puis $\varphi(x) = \left(\frac{1}{40} + \frac{3}{32}\right)x^7 + \mathrm{o}(x^7) = \frac{19}{160}x^7 + \mathrm{o}(x^7)$ et $\varphi \sim \frac{19}{160}x^7$. Ainsi $f \sim -\frac{19}{160}\frac{6}{5} = -\frac{57}{400}$, donc $\lim\limits_{x\to 0} f(x) = -\frac{57}{400}$.

Exercice 14.

On définit : $f : x \in]0, \pi/2[\mapsto f(x) := \dfrac{(\sin x)^{\sinh x} - (\sinh x)^{\sin x}}{(\tan x)^{\tanh x} - (\tanh x)^{\tan x}}$.

Montrer que f admet une limite quand x tend vers 0 et la calculer.

Solution.

On écrit $f = \frac{\varphi}{\psi}$, avec

$$\varphi(x) := \mathrm{e}^{\sinh x \ln \sin x} - \mathrm{e}^{\sin x \ln \sinh x} \quad \text{et} \quad \psi(x) := \mathrm{e}^{\tanh x \ln \tan x} - \mathrm{e}^{\tan x \ln \tanh x}.$$

Quand x tend vers 0, $\sin x$, $\sinh x$, $\tan x$ et $\tanh x$ tendent vers 0, donc, d'après la proposition 10 de la page 866, $\ln x \sim \ln \sin x \sim \ln \sinh x \sim \ln \tan x \sim \ln \tanh x$. Par suite :

$$\lim_{x\to 0} \sin x \ln \sinh x = \lim_{x\to 0} \sinh x \ln \sin x = \lim_{x\to 0} \tan x \ln \tanh x = \lim_{x\to 0} \tanh x \ln \tan x = 0,$$

donc, en utilisant (10), page 865 :

$$\varphi \sim \sinh x \ln \sin x - \sin x \ln \sinh x \quad \text{et} \quad \psi \sim \tanh x \ln \tan x - \tan x \ln \tanh x.$$

On en déduit $f \sim g$, avec $g := \frac{\sinh x \ln \sin x - \sin x \ln \sinh x}{\tanh x \ln \tan x - \tan x \ln \tanh x}$.

On note ν le numérateur de g et δ son dénominateur. On a :

$$\nu(x) := \sinh x \ln \frac{\sin x}{x} - \sin x \ln \frac{\sinh x}{x} + (\sinh x - \sin x)\ln x,$$

$$\delta(x) := \tanh x \ln \frac{\tan x}{x} - \tan x \ln \frac{\tanh x}{x} + (\tanh x - \tan x)\ln x.$$

La fonction ν est donc la somme de trois termes qui sont respectivement de la forme :

$$-\frac{x^3}{6} + \mathrm{o}(x^3), \ -\frac{x^3}{6} + \mathrm{o}(x^3) \quad \text{et} \quad \left(\frac{x^3}{3} + \mathrm{o}(x^3)\right)\ln x.$$

On a $x^3 = \mathrm{o}(x^3 \ln x)$, donc $\nu = \frac{x^3}{3}\ln x + \mathrm{o}(x^3 \ln x)$ et $\nu \sim \frac{x^3}{3}\ln x$.

De même $\delta \sim -\frac{2x^3}{3}\ln x$ et $\lim\limits_{x\to 0} f(x) = \lim\limits_{x\to 0} g(x) = -\frac{1}{2}$.

Exercice 15.
Étant donnée $f : \begin{array}{rcl} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \mathrm{e}^{-3(x-1)} + 3x \end{array}$ on définit $g : \begin{array}{rcl} \mathbb{R}^* & \longrightarrow & \mathbb{R} \\ x & \longmapsto & f(x) - f(1/x). \end{array}$

Trouver un équivalent simple de g en 1.

Solution. On a $g(1) = 0$. On pourrait, *a priori*, utiliser des développements limités au voisinage de 1. Il est plus simple d'utiliser les changements de variable $t := x-1$ et $u := \frac{1}{x}-1$: $\varphi(t) := f(x) = \mathrm{e}^{-3t} + 3(t+1)$ et $\psi(u) := f(1/x) = \mathrm{e}^{-3u} + 3(u+1)$ et d'utiliser des développements limités au voisinage de 0 (en t et en u). On a $u = -\frac{t}{1+t}$, donc $u \sim -t$ et, pour tout $k \in \mathbb{N}^*$, $\mathrm{o}(u^k) = \mathrm{o}(t^k)$.

Pour tout $k \in [\![1,4]\!]$, on trouve après quelques calculs simples, $g(x) = \mathrm{o}(t^k)$, il est donc nécessaire d'utiliser des développements à l'ordre 5. On a :

$$f(x) = \varphi(t) = 4 + \frac{9}{2}t^2 - \frac{9}{2}t^3 + \frac{27}{8}t^4 - \frac{81}{40}t^5 + \mathrm{o}(t^5),$$
$$f(1/x) = \psi(u) = 4 + \frac{9}{2}u^2 - \frac{9}{2}u^3 + \frac{27}{8}u^4 - \frac{81}{40}u^5 + \mathrm{o}(u^5).$$

Donc :

$$\begin{aligned} g(x) &= f(x) - f(1/x) = \varphi(t) - \psi(u) \\ &= \frac{9}{2}(t^2 - u^2) - \frac{9}{2}(t^3 - u^3) + \frac{27}{8}(t^4 - u^4) - \frac{81}{40}(t^5 - u^5) + \mathrm{o}(t^5). \end{aligned}$$

En utilisant :

$$u = -\frac{t}{1+t} = -t + t^2 - t^3 + t^4 - t^5 + \mathrm{o}(t^5),$$

on obtient :

$$u^2 = t^2 - 2t^3 + 3t^4 - 4t^5 + \mathrm{o}(t^5), \qquad u^3 = -t^3 + 3t^4 - 6t^5 + \mathrm{o}(t^5)$$
$$u^4 = t^4 - 4t^5 + \mathrm{o}(t^5), \qquad u^5 = -t^5 + \mathrm{o}(t^5).$$

Après quelques calculs simples laissés au lecteur, on obtient $g(x) = \frac{9}{20}t^5 + \mathrm{o}(t^5)$ et $g \sim \frac{9}{20}(x-1)^5$ au voisinage de $x = 1$.

---

5 Méthodes de calcul approché d'intégrales

Soient $a, b \in \mathbb{R}$, tels que $a < b$ et f une fonction *continue* sur $I := [a,b]$. On s'intéresse au *calcul numérique* de l'intégrale $\mathcal{J} := \displaystyle\int_a^b f(x)\,\mathrm{d}x$. Dans certains cas, on sait calculer une primitive F de f en termes de « fonctions élémentaires » et le calcul numérique de $\mathcal{J} = F(b) - F(a)$ est possible, par exemple, en utilisant des tables de valeurs de ces fonctions. Dans d'autres cas l'intégrale $\mathcal{J}$ peut être *exactement* calculée en utilisant des nombres transcendants « bien connus », comme π, e, $\ln 2$... Mais tous ces cas sont exceptionnels et, le plus souvent, il faut recourir à des méthodes numériques permettant un *calcul approché* de $\mathcal{J}$ (on dit que ce sont des *méthodes de quadrature*). Le but de cette partie est de décrire de telles méthodes, en se limitant aux plus « traditionnelles ».

Ces méthodes reposent toutes sur le même principe. On choisit $n \in \mathbb{N}^*$, on partage le segment $[a, b]$ à l'aide d'une subdivision *à pas constant* $\mathbf{u} := \left(x_i := a + i\frac{b-a}{n}\right)_{i\in[\![0,n]\!]}$ et l'on utilise la relation de Chasles en remplaçant sur chaque segment $[x_i, x_{i+1}]$ de la subdivison ($i \in [\![0, n-1]\!]$) la fonction f par une *fonction polynôme* P_i convenable, obtenue par « interpolation » (dont on sait calculer l'intégrale) : on obtient ainsi $\displaystyle\sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} P_i$ comme valeur approchée de $\mathcal{J}$. Ces méthodes sont dites *de Newton-Cotes* (NC_n) car elles ont été découvertes par Isaac Newton et son collaborateur Roger Cotes au début du XVIIIe siècle.

Pour chaque méthode on supposera f de classe $\mathcal{C}^p$ (avec $p \in \mathbb{N}^*$ convenable), on notera M_p un majorant de $|f^{(p)}|$ sur $[a, b]$ et l'on donnera un majorant de la valeur absolue de la différence entre l'intégrale et la valeur approchée en fonction de M_p et de n. On appelle un tel majorant *incertitude de la méthode*. Si ce majorant est de la forme $C\frac{M_p}{n^m}$ ($m \in \mathbb{N}^*$), on dit que la méthode est *d'ordre* m.

Dans chaque cas, la fonction polynôme choisie coïncide avec f en un certain nombre N de points de chaque segment $[x_i, x_{i+1}]$. On parle alors de *méthode à N points*. Nous nous limiterons à des méthodes *à un point* (méthodes des rectangles et des rectangles médians), *à deux points* (méthode des trapèzes) et *à trois points* (méthode de Simpson).

Les formules pour l'incertitude permettent de prévoir que si n est plus grand ou si, pour un n donné, l'ordre est plus grand, la méthode est meilleure. Mais, dans la pratique, en plus de l'erreur due à la méthode, il faut aussi tenir compte pour le calcul de l'erreur totale des erreurs de calcul et d'arrondi sur les valeurs de f en tous les points utilisés. Le nombre de ces points augmente avec n et il en est donc de même des erreurs de calcul et d'arrondi. Par suite, si l'on gagne d'un côté, on perd de l'autre...

Dans le module « Méthodes numériques » du livre L2, nous reviendrons sur l'intégration numérique. Nous donnerons des compléments et nous décrirons des généralisations.

5.1 Méthode des rectangles

La méthode des rectangles est une méthode à un point. On remplace sur chaque intervalle $[x_{i-1}, x_i]$ ($i \in [\![1, n]\!]$) la fonction f par la fonction constante $f(x_i)$. Pour être plus précis, on peut parler de méthode des rectangles à droite. Il y a aussi une méthode des rectangles à gauche : on remplace $f(x_i)$ par $f(x_{i-1})$. Les résultats sont similaires *mutatis mutandis*.

FIGURE IV.8.2 – méthode des rectangles, $n := 6$

En appliquant la méthode des rectangles, on obtient une fonction en escalier, notons son intégrale :

$$R_n(f) := \frac{b-a}{n} \sum_{i=1}^{n} f(x_i)$$

C'est la somme de Riemann à droite de f.

Si $f \geqslant 0$, l'intégrale $R_n(f)$ est une somme d'aires de rectangles, d'où le nom de la méthode.

Le lemme suivant va nous permettre de montrer que la méthode des rectangles est d'ordre 1.

Lemme 27. Soient $f \in \mathcal{C}^1(I)$ et M_1 un majorant de $|f'|$ sur I. Pour tous $c, d \in I$, tels que $c \leqslant d$, on a :

$$\left| \int_c^d f(x)\,\mathrm{d}x - (d-c)f(c) \right| \leqslant \frac{M_1}{2}(d-c)^2.$$

Démonstration. En appliquant l'inégalité de Taylor-Lagrange à l'ordre 2 à une primitive F de f, on obtient :

$$\left| \int_c^d f(x)\,\mathrm{d}x - (d-c)f(c) \right| = \left| F(d) - F(c) - (d-c)F'(c) \right| \leqslant \frac{M_1(c-d)^2}{2}.$$

On a utilisé $\sup\limits_{x\in I} \left|F''(x)\right| = \sup\limits_{x\in I} \left|f'(x)\right| \leqslant M_1$. ∎

Incertitude de la méthode

On suppose f de classe $\mathcal{C}^1$ sur I. On utilise la relation de Chasles et on applique sur chaque segment $[x_{i-1}, x_i]$ le lemme 27 (avec $c := x_{i-1}$ et $d := x_i$). On a :

$$\begin{aligned}
\left| \int_a^b f(x)\,\mathrm{d}x - R_n(f) \right| &= \left| \sum_{i=1}^n \left(\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{b-a}{n} f(x_i) \right) \right| \\
&\leqslant \sum_{i=1}^n \left| \int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{b-a}{n} f(x_i) \right| \\
&\leqslant n \frac{M_1}{2} \left(\frac{b-a}{n} \right)^2 = \frac{M_1(b-a)^2}{2n}.
\end{aligned}$$

Ainsi :

$$\left| \int_a^b f(x)\,\mathrm{d}x - R_n(f) \right| \leqslant \frac{M_1(b-a)^2}{2n}. \tag{30}$$

On peut montrer que l'ordre 1 obtenu ne peut pas être amélioré (*cf.* l'exercice IV.8.22 de la page 911).

5.2 Méthode des rectangles médians

La méthode des rectangles médians est une méthode à un point. On remplace sur chaque intervalle $[x_{i-1}, x_i]$ ($i \in [\![1, n]\!]$) la fonction f par la constante $f\left(\frac{x_{i-1}+x_i}{2}\right)$.

En appliquant la méthode des rectangles médians, on obtient une fonction en escalier. Notons son intégrale :

$$A_n(f) := \frac{b-a}{n} \sum_{i=1}^n f\left(\frac{x_{i-1}+x_i}{2} \right).$$

C'est une somme de Riemann de f.

FIGURE IV.8.3 – méthode des rectangles médians, $n := 6$

Le lemme suivant va nous permettre de montrer que la méthode des rectangles médians est d'ordre 2.

Lemme 28. Soient $f \in \mathcal{C}^2(I)$ et M_2 un majorant de $|f''|$ sur I. Pour tous $c, d \in I$, tels que $c \leqslant d$, on a :

$$\left| \int_c^d f(x)\,\mathrm{d}x - (d-c) f\left(\frac{c+d}{2}\right) \right| \leqslant \frac{M_2}{24}(d-c)^3.$$

Démonstration. Notons $\ell := \frac{c+d}{2}$. On a $\displaystyle\int_c^d (x-\ell)\,\mathrm{d}x = 0$, donc

$$(d-c)f(\ell) = \int_c^d \big(f(\ell) + (x-\ell)f'(\ell)\big)\,\mathrm{d}x \tag{31}$$

et, par suite, en utilisant l'inégalité de Taylor-Lagrange à l'ordre 2 pour f :

$$\begin{aligned}
\left| \int_c^d f(x)\,\mathrm{d}x - (d-c)f(\ell) \right| &= \left| \int_c^d \big(f(x) - f(\ell) - (x-\ell)f'(\ell)\big)\,\mathrm{d}x \right| \\
&\leqslant \int_c^d \big|f(x) - f(\ell) - (x-\ell)f'(\ell)\big|\,\mathrm{d}x \\
&\leqslant \int_c^d \frac{M_2}{2}(x-\ell)^2\,\mathrm{d}x = \frac{M_2}{24}(d-c)^3.
\end{aligned}$$

■

Remarque. Si $f \geqslant 0$, l'égalité (31) peut être interprétée géométriquement. Elle signifie que l'aire du rectangle de base $[c, d]$ et de hauteur $f(\ell)$ est égale à l'aire du trapèze de même base dont le côté supérieur est tangent en $(\ell, f(\ell))$ au graphe de f.

Incertitude de la méthode

On suppose $f \in \mathcal{C}^2(I)$. On utilise la relation de Chasles et on applique sur chaque segment $[x_{i-1}, x_i]$ le lemme 28 (avec $c := x_{i-1}$ et $d := x_i$). On a :

$$\begin{aligned}
\left| \int_a^b f(x)\,\mathrm{d}x - A_n(f) \right| &= \left| \sum_{i=1}^n \left(\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{b-a}{n} f\left(\frac{x_{i-1}+x_i}{2}\right) \right) \right| \\
&\leqslant \sum_{i=1}^n \left| \int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{b-a}{n} f\left(\frac{x_{i-1}+x_i}{2}\right) \right| \\
&\leqslant n\frac{M_2}{24}\left(\frac{b-a}{n}\right)^3 = \frac{M_2(b-a)^2}{24n^2}.
\end{aligned}$$

Ainsi :

$$\left| \int_a^b f(x)\,\mathrm{d}x - A_n(f) \right| \leqslant \frac{M_2(b-a)^3}{24n^2}. \tag{32}$$

On peut montrer que l'ordre 2 obtenu ne peut pas être amélioré (*cf.* l'exercice IV.8.22 de la page 911).

5.3 Méthode des trapèzes

La méthode des trapèzes est une méthode à deux points. On remplace sur chaque segment $[x_{i-1}, x_i]$ ($i \in [\![1, n]\!]$) la fonction f par l'unique *fonction affine* coïncidant avec f en x_{i-1} et x_i : on fait une *interpolation linéaire* (*cf.* l'exercice IV.8.3 de la page 905). Si $f \geqslant 0$, l'intégrale $T_n(f)$ de la fonction affine par morceaux obtenue est une somme d'aires de trapèzes, d'où le nom de la méthode.
On prouve facilement le résultat suivant.

FIGURE IV.8.4 – méthode des trapèzes, $n := 5$

Lemme 29. Soient $c, d \in I$, avec $c < d$. On désigne par φ l'unique fonction affine telle que $\varphi(c) = f(c)$ et $\varphi(d) = f(d)$. Alors :

$$\varphi = \frac{f(d)(x-c) + f(c)(d-x)}{d-c} \quad \text{et} \quad \int_c^d \varphi(x)\,\mathrm{d}x = \frac{d-c}{2}\left(f(c) + f(d)\right).$$

On en déduit, en utilisant la relation de Chasles :

$$T_n(f) := \frac{b-a}{n} \sum_{i=1}^{n} \frac{f(x_{i-1}) + f(x_i)}{2}.$$

Le lemme suivant va nous permettre de montrer que la méthode des trapèzes est d'ordre 2.

Lemme 30. Soient $f \in \mathcal{C}^2(I)$ et M_2 un majorant de $|f''|$ sur I. Pour tous $c, d \in I$, tels que $c \leqslant d$, on a :

$$\left| \int_c^d f(x)\,\mathrm{d}x - (d-c)\frac{f(c)+f(d)}{2} \right| \leqslant \frac{M_2}{12}(d-c)^3.$$

Démonstration. On va utiliser les propriétés des fonctions convexes (*cf.* la partie 2.8 du module IV.2).

On définit une fonction α sur $[c, d]$ par $\alpha : x \mapsto \alpha(x) := f(x) - \frac{M_2}{2}(x-c)(d-x)$. Elle est deux fois dérivable et $\alpha''(x) = f''(x) + M_2 \geqslant 0$, donc la fonction α est *convexe* (*cf.* le corollaire 69 de la page 667). De plus $\alpha(c) = f(c)$ et $\alpha(d) = f(d)$, par suite, pour tout $x \in [c, d]$, on a, en utilisant la définition de la convexité (*cf.* définition 19 de la page 658) :

$$\begin{aligned}\alpha(x) = \alpha\left(\frac{d-x}{d-c}c + \frac{x-c}{d-c}d\right) &\leqslant \frac{d-x}{d-c}\alpha(c) + \frac{x-c}{d-c}\alpha(d) \\ &= \frac{d-x}{d-c}f(c) + \frac{x-c}{d-c}f(d) = \varphi(x),\end{aligned}$$

donc $\alpha(x) \leqslant \varphi(x)$ (la corde est *en-dessus* de l'arc).

On définit une fonction β sur $[c,d]$ par $\beta : x \mapsto \beta(x) := f(x) + \frac{M_2}{2}(x-c)(d-x)$. Elle est deux fois dérivable et $\beta''(x) = f''(x) - M_2 \leqslant 0$, donc la fonction β est *concave*. De plus $\beta(c) = f(c)$ et $\beta(d) = f(d)$, par suite, pour tout $x \in [c,d]$, on a $\varphi(x) \leqslant \beta(x)$ (la corde est *au-dessous* de l'arc).
Ainsi :

$$\forall x \in [c,d], \quad f(x) - \frac{M_2}{2}(x-c)(d-x) \leqslant \varphi(x) \leqslant f(x) + \frac{M_2}{2}(x-c)(d-x).$$

On en déduit : $\big|f(x) - \varphi(x)\big| \leqslant \frac{M_2}{2}(x-c)(d-x)$.
D'après le lemme 29 : $\displaystyle\int_c^d \varphi(x)\,\mathrm{d}x = (d-c)\frac{f(c)+f(d)}{2}\cdot$, donc :

$$\begin{aligned}\left|\int_c^d f(x)\,\mathrm{d}x - (d-c)\frac{f(c)+f(d)}{2}\right| &= \left|\int_c^d f(x)\,\mathrm{d}x - \int_c^d \varphi(x)\,\mathrm{d}x\right| \\ &\leqslant \int_c^d \big|f(x) - \varphi(x)\big|\,\mathrm{d}x \\ &\leqslant \int_c^d \frac{M_2}{2}(x-c)(d-x) = \frac{M_2}{12}(d-c)^3\cdot\end{aligned}$$

■

Incertitude de la méthode.

On suppose $f \in \mathcal{C}^2(I)$. On utilise la relation de Chasles et on applique sur chaque segment $[x_{i-1}, x_i]$ le lemme 30 (avec $c := x_{i-1}$ et $d := x_i$).
On a :

$$\begin{aligned}\left|\int_a^b f(x)\,\mathrm{d}x - T_n(f)\right| &= \left|\sum_{i=1}^n \left(\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{b-a}{n}\frac{f(x_{i-1}) + f(x_i)}{2}\right)\right| \\ &\leqslant \sum_{i=1}^n \left|\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{b-a}{n}\frac{f(x_{i-1}) + f(x_i)}{2}\right| \\ &\leqslant n\frac{M_2}{12}\left(\frac{b-a}{n}\right)^3 = \frac{M_2(b-a)^2}{12n^2}\cdot\end{aligned}$$

Ainsi :

$$\left|\int_a^b f(x)\,\mathrm{d}x - T_n(f)\right| \leqslant \frac{M_2(b-a)^3}{12n^2}\cdot \tag{33}$$

On peut montrer que l'ordre 2 obtenu ne peut pas être amélioré (*cf.* l'exercice IV.8.22 de la page 911).

5.4 Méthode de Simpson

La méthode de Simpson[3] est une méthode à trois points. On remplace sur chaque intervalle $[x_{i-1}, x_i]$ ($i \in [\![1, n]\!]$) la fonction f par l'unique fonction polynôme de degré inférieur ou égal à 2 coïncidant avec f en x_{i-1}, $\frac{x_{i-1}+x_i}{2}$ et x_i. On note $S_n(f)$ l'intégrale de la fonction continue ainsi obtenue. Si $f \geqslant 0$, c'est une somme d'aires de domaines « sous des arcs de parabole ».

FIGURE IV.8.5 – méthode de Simpson, $n := 2$

Lemme 31. Soient $c, d \in I$, avec $c < d$. On désigne par ψ l'unique fonction polynôme de degré inférieur ou égal à 2 telle que :

$$\psi(c) = f(c), \quad \psi\left(\frac{c+d}{2}\right) = f\left(\frac{c+d}{2}\right) \quad \text{et} \quad \psi(d) = f(d).$$

Alors :

$$\int_c^d \psi(x)\,\mathrm{d}x = \frac{d-c}{6}\left(f(c) + f(d) + 4f\left(\frac{c+d}{2}\right)\right).$$

Démonstration. La fonction ψ est donnée par la méthode du *polynôme d'interpolation de Lagrange* (*cf.* théorème 23 de la page 298). On a :

$$\psi(x) = 2f(c)\frac{(x-d)\left(x-\frac{c+d}{2}\right)}{(c-d)^2} + 2f(d)\frac{(x-c)\left(x-\frac{c+d}{2}\right)}{(c-d)^2} - 4f\left(\frac{c+d}{2}\right)\frac{(x-c)(x-d)}{(c-d)^2}.$$

En utilisant $\frac{(x-c)(x-d)}{(c-d)^2} = \frac{(x-c)^2}{(c-d)^2} + \frac{x-c}{c-d}$, on obtient : $\displaystyle\int_c^d \frac{(x-c)(x-d)}{(c-d)^2}\,\mathrm{d}x = \frac{c-d}{6}$.

De façon analogue on obtient :

$$\int_c^d \frac{(x-d)(x-\frac{c+d}{2})}{(c-d)^2}\,\mathrm{d}x = \frac{d-c}{12} \quad \text{et} \quad \int_c^d \frac{(x-c)(x-\frac{c+d}{2})}{(c-d)^2}\,\mathrm{d}x = \frac{d-c}{12}.$$

D'où le résultat. ∎

On obtient donc, pour valeur approchée de l'intégrale :

$$S_n(f) := \frac{b-a}{6n}\sum_{i=1}^{n}\left(4f\left(\frac{x_{i-1}+x_i}{2}\right) + f(x_{i-1}) + f(x_i)\right).$$

On vérifie facilement que $S_n(f) = \frac{1}{3}\left(2A_n(f) + T_n(f)\right)$. On en déduit que, si $f \in \mathcal{C}^2(I)$, la méthode de Simpson est d'ordre au moins 2. Nous allons voir que, si $f \in \mathcal{C}^4(I)$, elle est en fait d'ordre 4. Nous utiliserons le lemme suivant.

[3] Due au mathématicien anglais Thomas Simpson 1710-1761.

Lemme 32. Soient $f \in \mathcal{C}^4(I)$ et M_4 un majorant de $|f^{(4)}|$ sur I. Pour tous $c \in I$ et $h > 0$, tels que $[c-h, c+h] \subset I$, on a :

$$\left| \int_{c-h}^{c+h} f(x)\,\mathrm{d}x - \frac{h}{3}\left(f(c+h) + f(c-h) + 4f(c)\right) \right| \leqslant \frac{M_4}{90} h^5. \tag{34}$$

Démonstration. L'idée de la preuve est simple. On définit une fonction auxiliaire Φ en remplaçant dans le premier membre de l'inégalité (34) ci-dessus le réel h par une variable t et en supprimant la valeur absolue. On majore ensuite Φ en utilisant ses trois premières dérivées.

On désigne par F une primitive de f et l'on définit :

$$\Phi : t \in [0,h] \mapsto \Phi(t) := F(c+t) - F(c-t) - \frac{t}{3}\left(f(c+t) + f(c-t) + 4f(c)\right).$$

La fonction f étant de classe $\mathcal{C}^3$, la fonction Φ est de classe $\mathcal{C}^4$. Ses trois premières dérivées sont :

$$\begin{aligned}
\Phi'(t) &:= \frac{2}{3}\left(f(c+t) + f(c-t)\right) - \frac{t}{3}\left(f'(c+t) - f'(c-t)\right) - \frac{4}{3}f(c),\\
\Phi''(t) &:= \frac{1}{3}\left(f'(c+t) - f'(c-t)\right) - \frac{t}{3}\left(f''(c+t) + f''(c-t)\right),\\
\Phi^{(3)}(t) &:= -\frac{t}{3}\left(f^{(3)}(c+t) - f^{(3)}(c-t)\right).
\end{aligned}$$

On a $\Phi(0) = \Phi'(0) = \Phi''(0) = 0$ et, en utilisant l'inégalité des accroissements finis pour la fonction $f^{(3)}$: $\left|\Phi^{(3)}(t)\right| \leqslant \frac{2M_4}{3} t^2$.

Par suite :

$$\left|\Phi''(t)\right| = \left| \int_0^t \Phi^{(3)}(u)\,\mathrm{d}u \right| \leqslant \frac{2M_4}{3} \int_0^t u^2\,\mathrm{d}u = \frac{2M_4}{9} t^3,$$

donc :

$$\left|\Phi'(t)\right| = \left| \int_0^t \Phi''(u)\,\mathrm{d}u \right| \leqslant \frac{2M_4}{9} \int_0^t u^3\,\mathrm{d}u = \frac{M_4}{18} t^4$$

et :

$$\left|\Phi(t)\right| \leqslant \frac{M_4}{18} \int_0^t u^4\,\mathrm{d}u = \frac{M_4}{90} t^5.$$

On obtient (34) en faisant $t := h$ dans l'inégalité ci-dessus. ■

Incertitude de la méthode

On suppose $f \in \mathcal{C}^4(I)$.

Soit $i \in [\![1, n]\!]$. Notons $c := \frac{x_{i-1}+x_i}{2}$ et $h := \frac{b-a}{2n}$. On a $x_{i-1} = c - h$ et $x_i = c + h$. En appliquant (34), on obtient :

$$\left| \int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{h}{3}\left(f(x_i) + f(x_{i-1}) + 4f\left(\frac{x_{i-1}+x_i}{2}\right)\right) \right| \leqslant \frac{M_4 (b-a)^5}{2880\, n^5}.$$

En sommant sur $i \in [\![1, n]\!]$ les inégalités ci-dessus et en utilisant la relation de Chasles et l'inégalité triangulaire, on en déduit :

$$\left| \int_a^b f(x)\,\mathrm{d}x - S_n(f) \right| \leqslant \frac{M_4(b-a)^5}{2880\,n^4}. \tag{35}$$

On peut montrer, en considérant le cas des polynômes de degré 4, que l'ordre 4 obtenu ci-dessus ne peut pas être amélioré (*cf.* l'exercice IV.8.22 de la page 911).

Un exemple

On peut montrer que l'on ne peut pas exprimer une primitive de la fonction $x \mapsto \mathrm{e}^{-x^2}$ en utilisant les fonctions transcendantes « élémentaires » et des opérations algébriques (c'est un résultat difficile...). Cela a justifié l'introduction d'une nouvelle fonction « spéciale », la fonction d'erreur : $erf : x \mapsto \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t$. Cette fonction est utile en probabilités et statistiques (pour exprimer la marge d'erreur d'une évaluation) et pour l'étude de la diffusion de la chaleur. Il existe des tables de la fonction d'erreur et l'on peut aussi, pour en obtenir des valeurs, utiliser un système de calcul formel ou numérique (comme *Scilab*).

À titre d'exemple, nous allons calculer des valeurs approchées de $E := \displaystyle\int_0^1 \mathrm{e}^{-x^2}\,\mathrm{d}x$ à l'aide des quatre méthodes que nous venons de présenter.

Dans ce paragraphe, les calculs seront faits à l'aide d'une calculette 4 opérations. Nous utiliserons ensuite, au paragraphe suivant, le système `Maple`.

Dans tous les cas, on choisira la même subdivision : $n := 1$ et $(x_0 = a := 0,\ x_1 = b := 1)$.

On utilisera une valeur exacte et des valeurs approchées, avec 5 décimales exactes, de la fonction $f : x \mapsto \mathrm{e}^{-x^2}$ aux points $0, 1/2$ et 1 :

$$f(0) = 1,\ f(1/2) = 0{,}77880,\ f(1) = 1/\mathrm{e} = 0{,}36788.$$

La valeur approchée de E avec 5 décimales exactes est $\displaystyle\int_0^1 \mathrm{e}^{-x^2}\,\mathrm{d}x \approx 0{,}74682$, ce qui permet de « calculer » les erreurs dans chaque cas.

Nous laissons le détail des calculs au lecteur. Les résultats sont présentés dans le tableau ci-dessous.

Méthode	des rectangles	des rectangles médians	des trapèzes	de Simpson
E approchée	0,36788	0,77880	0,68394	0,74718
Erreur	0,37894	-0,03198	0,06288	-0,00036

On constate que l'erreur (absolue) faite avec la méthode des trapèzes est environ le double de celle faite avec la méthode des rectangles médians comme on peut le prévoir en considérant les inégalités (32) et (33) et que la meilleure méthode est celle de Simpson : avec trois points on obtient une approximation à 4×10^{-4} près.

Utilisation du calcul formel

En utilisant un système (convenable...) de calcul formel, on peut facilement écrire, pour f, a ,b, n donnés et pour chaque méthode, l'approximation écrite formellement, la comparaison graphique entre celle-ci et la fonction et le calcul numérique de l'approximation. Par exemple, en utilisant `Maple` (version 2017, **`Student[Calculus1]`**), on dispose de la commande **`ApproximateInt(Int(f(x), x = a..b), opts)`** avec, en particulier les options suivantes :

```
ApproximateInt (Int(f(x), x = a..b, method= ,
                                    output= , partition = ).
```

- L'option **`partition`** $= n$ permet de choisir le nombre n de sous-segments.
- L'option **`method`** permet de choisir la méthode d'approximation : **`right`** (rectangles à droite), **`left`** (rectangles à gauche), **`midpoint`** (rectangles médians), **`trapezoid`** (trapèzes), **`simpson`** (Simpson).
- L'option **`output`** permet de choisir la **`sortie`** souhaitée : **`plot`** (superposition du graphe de la fonction et de celui de l'approximation), **`sum`** (expression formelle de l'approximation sous forme de somme $\Sigma := \sum_{i=1}^{n-1} \star$). On peut ensuite obtenir la valeur numérique de l'approximation en utilisant **`evalf`** : $\int_a^b f \approx$ **`evalf`**(Σ).

À titre d'exemple, reprenons le calcul approché de l'intégrale $E := \int_0^1 \mathrm{e}^{-x^2}\,\mathrm{d}x$ (*cf.* la page 901) en utilisant ces outils.
La commande :

```
ApproximateInt(exp(-x^2),x=0..1,method=midpoint,
                                    output=plot,partition=6)
```

donne la figure IV.8.6 de la page suivante.
La commande :

```
ApproximateInt(exp(-x^2),x=0..1,method=midpoint,
                                    output=sum,partition=6)
```

donne $\Sigma := \frac{1}{6}\sum_{i=0}^{5} \mathrm{e}^{-(\frac{i}{6}+\frac{1}{12})^2}$ et et **`evalf(`**Σ**`)`** donne l'approximation $0{,}7476770835$.
En variant les méthodes et la valeur de **`partition`**, on obtient le tableau suivant :

	right	midpoint	trapezoid	simpson	partition
E	0,3678794412	0,7788007831	0,6839397206	0,7471804289	1
E	0,6346330487	0,7502523495	0,7399864753	0,7468303915	3
E	0,6924426991	0,7476770835	0,7451194124	0,7468245264	6
E	0,7146047681	0,7471308778	0,7462107961	0,7468241839	10
E	0,7404784019	0,7468363957	0,7467996072	0,7468241329	50
E	0,7436573984	0,7468271984	0,7468180015	0,7468241328	100

FIGURE IV.8.6 – méthode des rectangles médians, $n := 6$

En utilisant **`evalf(int(exp(-x^2)),x=0..1),15)`**, `Maple` donne pour E l'approximation $0{,}746824132812427$, avec 15 décimales exactes, ce qui permet de trouver les décimales exactes (soulignées) dans les approximations du tableau. On peut aussi utiliser **`evalf(sqrt (Pi)/2*erf(1),15)`**, qui donne les mêmes résultats.

Exercice type corrigé

IV.8.0 Soient deux fonctions de classe $\mathcal{C}^\infty$ au voisinage de 0 et $p \in \mathbb{N}^*$, $p \geqslant 2$. On suppose que f et g admettent au voisinage de 0 des développements limités de la forme :

$$f(t) = t + at^p + bt^{2p-1} + \mathrm{o}(t^{2p-1}) \quad \text{et} \quad g(t) = t + a't^p + b't^{2p-1} + \mathrm{o}(t^{2p-1}).$$

1. Montrer que $g \circ f - f \circ g$ admet un développement limité à l'ordre $2p-1$ au voisinage de 0.
2. Montrer que la fonction $h : x \mapsto \tan(\sin x) - \sin(\tan x)$ est équivalente, lorsque x tend vers 0, à un monôme de degré 7 ; le calculer.

Solution.

1. On a :

$$f(t) = t + at^p + bt^{2p-1} + \mathrm{o}(t^{2p-1}) = t\left(1 + at^{p-1} + bt^{2p-2} + \mathrm{o}(t^{2p-2})\right) = t\left(1 + u(t)\right),$$

avec $u(t) := at^{p-1} + bt^{2p-2} + \mathrm{o}(t^{2p-2})$.

On en déduit, en utilisant la formule du binôme pour $(1+u(t))^p$:

$$f(t)^p = t^p\left(1 + pat^{p-1} + \mathrm{o}(t^{p-1})\right) = t^p + pat^{2p-1} + \mathrm{o}(t^{2p-1}) \quad \text{et} \quad f(t)^{2p-1} = t^{2p-1} + \mathrm{o}(t^{2p-1}),$$

puis :

$$g \circ f(t) = t + (a+a')t^p + (b + b' + paa')t^{2p-1} + \mathrm{o}(t^{2p-1}).$$

En permutant f et g, on obtient :

$$f \circ g(t) = t + (a + a')t^p + (b + b' + paa')t^{2p-1} + \mathrm{o}(t^{2p-1}).$$

D'où : $g \circ f(t) - f \circ g(t) = \mathrm{o}(t^{2p-1})$. Ainsi $g \circ f - f \circ g$ admet un développement limité à l'ordre $2p - 1$ au voisinage de 0 dont la partie régulière est nulle.

2. On applique 1, avec $f := \sin$, $g := \tan$ et $p := 3$ (les conditions sont satisfaites car $\sin$ et $\tan$ sont des fonctions impaires). La fonction $h := \tan \circ \sin - \sin \circ \tan$ admet un développement limité en 0 à tout ordre. D'après la question 1, la partie régulière de son développement limité à l'ordre $5 = 2 \times 3 - 1$ est nulle. La fonction h est *impaire*, donc son développement limité à l'ordre 7 est de la forme $h(t) = \alpha\, t^7 + \mathrm{o}(t^7)$. Il reste à calculer α.
On développe $\sin$ et $\tan$ à l'ordre 7 en 0 :

$$\sin t = t - \frac{t^3}{6} + \frac{t^5}{120} - \frac{t^7}{5040} + \mathrm{o}(t^7) = t\left(1 - \frac{t^2}{6} + \frac{t^4}{120}\right) - \frac{t^7}{5040} + \mathrm{o}(t^7)$$
$$\tan t = t + \frac{t^3}{3} + \frac{2t^5}{15} + \frac{17t^7}{315} + \mathrm{o}(t^7) = t\left(1 + \frac{t^2}{3} + \frac{2t^4}{15}\right) + \frac{17t^7}{315} + \mathrm{o}(t^7)$$

Notons $T : t \mapsto \tan(\sin t)$ et $S : t \mapsto \sin(\tan t)$. On calcule les monômes en t^7 dans les parties régulières des développements limités de T et S.
Pour T, on obtient $-\frac{1}{5040}t^7$ (pour $\sin t$), $\frac{t^3}{3}\frac{13t^4}{120} = \frac{13}{360}t^7$ (pour $\sin^3 t$), $\frac{2t^5}{15}\frac{-5t^2}{6} = -\frac{1}{9}t^7$ (pour $\sin^5 t$) et $\frac{17}{315}t^7$ (pour $\sin^7 t$).
Pour S, on obtient $\frac{17}{315}t^7$ (pour $\tan t$), $-\frac{t^3}{6}\frac{11t^4}{15} = -\frac{11}{90}t^7$ (pour $\tan^3 t$), $\frac{t^5}{120}\frac{5t^2}{3} = \frac{1}{72}t^7$ (pour $\tan^5 t$) et $-\frac{1}{5040}t^7$ (pour $\tan^7 t$).
Par suite :

$$\alpha = \frac{13}{360} - \frac{1}{9} + \frac{11}{90} - \frac{1}{72} = \frac{1}{30}\cdot$$

La fonction h est donc équivalente à $\frac{t^7}{30}$ au voisinage de 0.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

Formules de Taylor

IV.8.1 Calculer l'intégrale : $I := \displaystyle\int_0^1 \frac{(1+t)(1-t)^2}{(1+t^2)^2}\,\mathrm{d}t$ en appliquant la formule de Taylor avec reste intégral à la fonction $f : t \mapsto \ln(1 + t^2)$.

IV.8.2

1. Pour tout $n \in \mathbb{N}^*$, on considère la fonction $g_n : t \in [0,1] \mapsto (1-t)^n \mathrm{e}^t$. Montrer que $0 \leqslant g_n \leqslant 1$.
2. En utilisant une formule de Taylor avec reste intégral retrouver l'encadrement :

$$1 + \sum_{k=1}^{n} \frac{1}{k!} \leqslant e \leqslant 1 + \sum_{k=1}^{n} \frac{1}{k!} + \frac{1}{n!}\cdot$$

IV.8.3 $*$ *Autour de l'interpolation linéaire.*

La méthode d'interpolation linéaire consiste à calculer une *valeur approchée* d'une fonction f définie sur un segment $[a,b]$ en un point $c \in]a,b[$ en prenant la valeur en c de l'unique fonction affine P qui coïncide avec f en a et b. Avant l'utilisation systématique des calculettes et des ordinateurs, cette méthode a été utilisée pendant des siècles pour calculer la valeur approchée d'une fonction en un point intermédiaire entre deux entrées d'une table numérique de fonctions élémentaires.

Soit f une fonction de classe $\mathcal{C}^2$ sur un intervalle $[a,b]$ d'intérieur non vide. On se propose d'encadrer l'erreur $P(c)-f(c)$ dans la méthode d'interpolation linéaire et d'en donner une expression intégrale en fonction de f''.

1. a) On note $P(x) := \dfrac{b-x}{b-a}f(a) + \dfrac{x-a}{b-a}f(b)$. Montrer, en utilisant deux formules de Taylor-Lagrange, que, pour tout $c \in [a,b]$, il existe $d_1, d_2 \in [a,b]$, tels que :

$$P(c)-f(c) = \frac{(c-a)(b-c)}{2}\left(\frac{c-a}{b-a}f''(d_1) + \frac{b-c}{b-a}f''(d_2)\right). \tag{36}$$

En déduire qu'il existe $d \in [a,b]$, tel que :

$$P(c)-f(c) = \frac{(c-a)(b-c)}{2}f''(d). \tag{37}$$

b) Montrer que, pour tout $c \in [a,b]$:

$$\frac{(c-a)(b-c)}{2}\inf_{x\in[a,b]} f''(x) \leqslant P(c)-f(c) \leqslant \frac{(c-a)(b-c)}{2}\sup_{x\in[a,b]} f''(x). \tag{38}$$

En déduire : $\left|f(c)-P(c)\right| \leqslant \frac{(c-a)(b-c)}{2}\sup\limits_{x\in[a,b]}\left|f''(x)\right| \leqslant \frac{(b-a)^2}{8}\sup\limits_{x\in[a,b]}\left|f''(x)\right|$.

c) On suppose que $f'' \geqslant 0$ (resp. $f'' \leqslant 0$). Montrer, en utilisant la question précédente, que f est convexe (resp. concave).

Inversement, retrouver les inégalités (38) en utilisant les propriétés des fonctions convexes et concaves.

2. Pour tout $t \in [0,1]$, on définit une fonction $K_t : [0,1] \to \mathbb{R}$ par :

$$K_t(t) = t(1-t) \quad \text{et} \quad K_t(s) = \begin{cases} (1-t)s & \text{si} \quad s \in [0,t[\\ t(1-s) & \text{si} \quad s \in]t,1]. \end{cases}$$

Montrer que K_t est une fonction continue positive et calculer $\displaystyle\int_0^1 K_t(s)\,\mathrm{d}s$.

3. Soit $c \in [a,b]$. On note $t := \frac{c-a}{b-a}\cdot$ Montrer :

$$P(c)-f(c) = (1-t)f(a)+tf(b)-f(c) = (b-a)^2\int_0^1 f''\left((1-s)a+sb\right)K_t(s)\,\mathrm{d}s. \tag{39}$$

Retrouver les inégalités (38).

4. Dans la pratique, quand on applique la méthode d'interpolation linéaire pour calculer une valeur approchée de $f(c)$ avec $c \in [a,b]$, on ne dispose que de *valeurs approchées* de $f(a)$ et $f(b)$. Ainsi à l'erreur de méthode s'ajoutent une erreur due à la fabrication de la table et l'erreur d'arrondi. Les tables sont en général faites de telle sorte que, pour une table à m décimales, les deux dernières sont majorées (en module) par $5.10^{-(m+1)}$.
   (a) Majorer l'erreur totale pour la méthode d'interpolation linéaire quand a et b correspondent à deux entrées successives d'une table de fonctions numériques et quand $t := \frac{c-a}{b-a}$ est un nombre décimal.
   (b) Calculer l'erreur de la méthode dans le cas d'une table de logarithmes décimaux à 5 décimales[4] pour $a \in [10^3, 10^4[$ et $b := a+1$.
   (c) On utilise une table à 8 décimales de la fonctions sinus, la variable x étant *en degrés*. On note $sin(x) := \sin\left(\frac{\pi}{180}x\right)$. La table donne $sin(40) \approx 0{,}64278761$ et $sin(41) \approx 0{,}65605903$. En déduire une valeur approchée de $sin(40{,}25)$ et étudier l'erreur commise.

IV.8.4 $*$ Soit $g : [-1,1] \to \mathbb{R}$ une fonction impaire de classe $\mathcal{C}^5$.

1. Écrire les formules de Taylor avec reste intégral en $x = 0$ à l'ordre 4 pour g et à l'ordre 3 pour g'. En déduire une formule pour $g - \frac{x}{3}g'$.
2. Montrer que pour tout $x \in]0,1]$, il existe $d_x \in]0,x[$ tel que :

$$\int_0^x g^{(5)}(t)\left(\frac{(x-t)^4}{24} - \frac{x(x-t)^3}{18}\right)\mathrm{d}t = -\frac{x^5}{180}\,g^{(5)}(d_x).$$

En déduire :

$$\exists\, d \in]0,1[, \quad g(1) = \frac{1}{3}\left(2g'(0) + g'(1)\right) - \frac{x^5}{180}g^{(5)}(d). \tag{40}$$

Développements limités et calculs de limites

IV.8.5 Écrire le développement limité en 0 de $\dfrac{x+1}{x-2}$ à l'ordre 3, puis celui de $\dfrac{x^2-x}{2x^2-x+5}$ à l'ordre 2.

IV.8.6 Calculer les développements limités en 0, à l'ordre 3, des fonctions suivantes :

$$\sinh x \sin x, \qquad \frac{1}{(1+x)(1-x^4)}, \qquad \sin x \ln(1+x).$$

IV.8.7 Calculer le développement limité de $x^4 - 1$ à l'ordre 3 aux points 0 et -1.

IV.8.8 On considère la fonction f définie par : $\forall x \in \mathbb{R},\ f(x) = \ln\left(x + \sqrt{x^2+1}\right)$. Calculer f'. En déduire le développement limité de f à l'ordre 5 en 0.

IV.8.9 Écrire le développement limité de $\dfrac{1}{1+\frac{x}{2}}$ à l'ordre 3 en 0. En déduire le développement limité de $1/x$ à l'ordre 3 en 2.

[4] Comme, par exemple, la table de Bouvard et Ratinet, aux éditions Hachette, « populaire » chez les lycéens et les étudiants des années 60.

IV.8.10 Écrire le développement limité de $\ln\left(1+\dfrac{x}{e}\right)$ à l'ordre 2 en 0.
En déduire celui de $\ln x$ à l'ordre 2 en e.

IV.8.11 Calculer les limites suivantes :

$$\lim_{x\to 0}\frac{\sin x - x}{x^2},\qquad \lim_{x\to 0}\frac{\sin x-\tan x}{\tanh x-\tan x},$$

$$\lim_{x\to+\infty}\sqrt[3]{x^3+x^2+7}-x,\qquad \lim_{x\to\frac{\pi}{2}}\frac{\sqrt[4]{1+2\cos x}-1}{x-\frac{\pi}{2}}.$$

IV.8.12 Écrire le développement limité à l'ordre 2 en 0 des fonctions suivantes :

$$\frac{\sin x - x}{x^2},\qquad \frac{\sin x-\tan x}{\tanh x-\tan x}.$$

IV.8.13 Soit f définie sur $\mathbb{R}$ par :

$$\forall x\geqslant 0,\ f(x)=\cosh\sqrt{x}\qquad\text{et}\qquad \forall x<0,\ f(x)=\cos\sqrt{-x}.$$

1. Écrire le développement limité de $\cos x$ et $\cosh x$ à l'ordre 4 en 0, puis montrer que f admet un développement limité à l'ordre 2 en 0 et le calculer.
2. Montrer que f est continûment dérivable sur $\mathbb{R}$.

IV.8.14 Soit f une fonction deux fois continûment dérivable sur $\mathbb{R}$. Calculer :

$$\lim_{h\to 0}\frac{2f(x+3h)-3f(x+2h)+f(x)}{3h^2}.$$

IV.8.15

1. Soit la fonction $f:\mathbb{R}\setminus\{2\}\to\mathbb{R}$ définie par $f(x):=\dfrac{2x^2-7x-9}{x-2}$.
   En utilisant un développement limité en 0 adéquat, écrire $f(x)$ sous la forme :
   $$f(x)=ax+b+\frac{c}{x}+\frac{1}{x}\,\varphi\Big(\frac{1}{x}\Big)\qquad\text{avec }\lim_{t\to 0}\varphi(t)=0.$$
   En déduire que le graphe de f admet une asymptote en $+\infty$ (resp. $-\infty$) et déterminer sa position par rapport à cette asymptote.
2. Mêmes questions avec les fonctions :
   $$\sqrt[4]{x^4-x^3-1}+x,\qquad x^2\arctan\left(\frac{1}{1+x}\right).$$

IV.8.16 Montrer comment appliquer les théories des équivalences et des développements limités pour étudier les limites éventuelles des suites de termes généraux :

1. $u_n:=(\mathrm{e}^{1/n}-1)\cdot\dfrac{n^2+1}{n^4+1}$, $\quad v_n:=\left(1+\dfrac{5}{n}\right)^{3n+1}$, $\quad w_n:=\left(1+\dfrac{3}{n}\right)^{1/n}$,
   $$z_n:=(n+3)^2\left(\left(\frac{1}{n}\right)^{1/n^2}-1\right).$$

2. $u_n := \left(\sin\dfrac{1}{n}\right)\cdot(n^2+1)$, $\qquad u_n := \left(1-\cos\dfrac{1}{\sqrt{n}}\right)\cdot n$

   $u_n := \dfrac{n\cos(1/n)}{2\sinh(1/n)}$, $\qquad u_n := \left[\arctan\left(\dfrac{n}{n^2+1}\right)\right]\cdot\left[n^2\tan\dfrac{1}{n}\right]$.

3. $u_n := n^2\left(\sinh\dfrac{1}{\sqrt{n}}\right)\left(\arctan\dfrac{2}{n^2}\right)$.

4. $u_n := \left(\cos\left(\dfrac{1}{n}\right)-1\right)^2\cdot\left(n^4+n+1\right)$.

5. $v_n := \left(\mathrm{e}^{1/n}-1-\dfrac{1}{n}-\dfrac{1}{2n^2}-\dfrac{1}{6n^3}\right)\cdot n^4$.

IV.8.17

Trouver des équivalents simples en 0 des fonctions $\sinh\circ\sin-\sin\circ\sinh$ et $\tanh\circ\tan-\tan\circ\tanh$.

IV.8.18

1. Pour tout $p\in\mathbb{N}$, on note $\tilde{A}_p$ (resp. A_p) la partie régulière du développement limité de $\sinh x$ (resp. $\sin x$) en 0 à l'ordre p et $\tilde{C}_p$ (resp. C_p) la partie régulière du développement limité de $\cosh x$ (resp. $\cos x$) en 0 à l'ordre p.

   On note $\tilde{A}_p = \tilde{Q}_p\tilde{C}_p + x^{p+1}\tilde{S}_p$ (resp. $A_p = Q_pC_p + x^{p+1}S_p$) la division suivant les puissances croissantes de $\tilde{A}_p$ par $\tilde{C}_p$ (resp. A_p par C_p) à l'ordre p.

   a) Montrer que $\tilde{A}_p(ix) = iA_p(x)$ et $\tilde{C}_p(ix) = C_p(x)$.

   b) Montrer que $\tilde{Q}_p(ix) = iQ_p(x)$.

2. Calculer le développement limité de $\tan x$ en 0 à l'ordre p en utilisant les nombres de Bernoulli (*cf.* (29) à la page 889).

Formules de Taylor et dérivation numérique

IV.8.19 $*$ *Dérivation numérique : dérivée première.*

Soient n un entier supérieur ou égal à 2 et (x_p) la subdivision régulière d'un intervalle $[a,b]$ définie par $x_p := a+ph$ où $h := \frac{b-a}{n}$ et $p\in[\![0,n]\!]$. On suppose que f est une fonction admettant sur cet intervalle une dérivée troisième continue f''', bornée par un réel M.

On suppose que f est « uniquement connue » par la donnée des $n+1$ nombres $(u_p = f(x_p))_{p\in[\![0,n]\!]}$ et l'on se propose de donner des valeurs approchées des $n+1$ nombres $(v_p = f'(x_p))_{p\in[\![0,n]\!]}$ en fonction des $(u_p)_{p\in[\![0,n]\!]}$ avec une majoration de l'erreur en fonction de h et M. D'après la définition de la dérivée en un point, on peut approximer celle-ci par un taux d'accroissement, c'est-à-dire $v_p \approx \frac{1}{h}(u_{p+1}-u_p)$ (ou $v_p \approx \frac{1}{h}(u_p-u_{p-1})$), mais nous nous proposons de donner une « meilleure approximation » de la dérivée. Plus précisément, l'objet de l'exercice est de démontrer

les relations :

$$\left| v_p - \frac{1}{2h}\,(u_{p+1} - u_{p-1}) \right| = \left| \frac{h^2}{6}\, f'''(\omega) \right| \leqslant \frac{Mh^2}{6} \quad \text{pour} \quad 0 < p < n, \tag{41}$$

$$\left| v_0 - \frac{1}{2h}\,(-3u_0 + 4u_1 - u_2) \right| = \left| \frac{h^2}{3}\, f'''(\theta) \right| \leqslant \frac{Mh^2}{3}, \tag{42}$$

$$\left| v_n - \frac{1}{2h}\,(-u_{n-2} + 4u_{n-1} - 3u_n) \right| = \left| \frac{h^2}{3}\, f'''(\tau) \right| \leqslant \frac{Mh^2}{3}\cdot \tag{43}$$

où ω, θ et τ sont des réels convenables dans $]a, b[$ (ω dépendant de p).
On montrera ainsi que l'on a des valeurs approchées :

$$v_p \approx \frac{1}{2h}\,(u_{p+1} - u_{p-1}), \quad v_0 \approx \frac{1}{2h}\,(-3u_0 + 4u_1 - u_2), \quad v_n \approx \frac{1}{2h}\,(-u_{n-2} + 4u_{n-1} - 3u_n),$$

et que l'erreur d'approximation est majorée par $\frac{Mh^2}{6}$ ou $\frac{Mh^2}{3}$ selon les cas.

1. a) Montrer, en utilisant deux formules de Taylor-Young, que, pour $0 < p < n$:

$$f'(x_p) - \frac{1}{2h}\,(f(x_{p+1}) - f(x_{p-1})) = -\frac{h^2}{6}\, f'''(x_p) + \mathrm{o}(h^2).$$

En déduire qu'il existe $\omega \in]a, b[$ tel que l'on ait les relations (41). On pourra utiliser la fonction auxiliaire φ définie sur $[0, h]$ par :

$$\varphi(t) := 2t\, f'(x_p) - f(x_p + t) + f(x_p - t) + \lambda\, \frac{t^3}{3},$$

où la constante λ est choisie de telle sorte que $\varphi(h) = 0$.

b) Interpréter géométriquement l'approximation (41) démontrée ci-dessus (on notera A_p le point de coordonnées (x_p, u_p)).

2. Montrer, en utilisant deux formules de Taylor-Young, que :

$$f'(x_0) - \frac{1}{2h}\big(- f(x_0) + 4f(x_1) - f(x_2)\big) = \frac{h^2}{3}\, f'''(x_1) + \mathrm{o}(h^2).$$

En déduire qu'il existe $\theta \in]a, b[$ tel que l'on ait les relations (42).
On pourra utiliser la fonction auxiliaire ψ définie sur $[0, h]$ par :

$$\psi(t) := 2t\, f'(0) + 3f(0) - 4f(t) + f(2t) - 2\mu\, \frac{t^3}{3},$$

où la constante μ est choisie de telle sorte que $\psi(h) = 0$.

3. Interpréter géométriquement l'approximation (42) démontrée ci-dessus.
4. Justifier le choix des coefficients $-3, 4, -1$ dans le premier membre de (42).
5. Prouver (43).
6. Comparer les approximations (41), (42), (43) et les approximations $v_p \approx \frac{1}{h}(u_{p+1} - u_p)$.

IV.8.20 $*$ *Dérivation numérique : dérivée seconde.*
Soient n un entier supérieur ou égal à 2 et $(x_p)_{p\in[\![0,n]\!]}$ la subdivision régulière d'un intervalle $[a,b]$ définie par $x_p := a + ph$ où $h := \frac{b-a}{n}$ et $p \in [\![0,n]\!]$. On suppose que f est une fonction admettant sur cet intervalle une dérivée quatrième continue $f^{(4)}$, bornée par un réel M'.
On suppose que f est « uniquement connue » par la donnée des $n+1$ nombres $(u_p = f(x_p))$, et l'on se propose de donner des valeurs approchées des $n-1$ nombres $(w_p = f''(x_p))_{p\in[\![1,n-1]\!]}$ en fonction des $\{u_p\}_{p\in[\![0,n]\!]}$ avec une majoration de l'erreur en fonction de h et M'.
Plus précisément, l'objet de l'exercice est de démontrer des relations :

$$\left| w_p - \frac{1}{h^2}\left(u_{p+1} - 2u_p + u_{p-1}\right)\right| = \frac{h^2}{12}\left|f^{(4)}(\rho)\right| \leqslant \frac{M'h^2}{12}\cdot, \tag{44}$$

où ρ est un réel convenable dans $[a,b]$ (dépendant de p).
On montrera ainsi que les valeurs approchées sont : $w_p \approx \frac{1}{h^2}(u_{p+1} - 2u_p + u_{p-1})$ et que l'erreur d'approximation est majorée par $\frac{M'h^2}{12}$.
Montrer que, pour $0 < p < n$:

$$f''(x_p) - \frac{1}{h^2}\left(f(x_{p+1}) - 2f(x_p) + f(x_{p-1})\right) = -\frac{h^2}{12}f^{(4)} + \mathrm{o}(h^2).$$

En déduire qu'il existe $\rho \in \,]-x_{p-1}, x_{p+1}[$ tel que l'on ait les relations :

$$\left| w_p - \frac{1}{h^2}\left(u_{p+1} - 2u_p + u_{p-1}\right)\right| = \frac{h^2}{12}\left|f^{(4)}(\rho)\right| \leqslant \frac{M'h^2}{12}\cdot$$

On pourra utiliser la fonction auxiliaire χ définie sur $[0,h]$ par :

$$t \mapsto \chi(t) := t^2 f''(x_p) - f(x_p + t) + 2f(x_p) - f(x_p - t) + \lambda\frac{t^4}{12},$$

où λ est un réel convenablement choisi.

Méthodes de calcul approché d'intégrales

IV.8.21 $*$ Soient $I := [a,b]$ et $f \in \mathcal{C}^4(I)$.
Montrer, en utilisant la formule (40) de l'exercice IV.8.4 de la page 906, que, pour tous $c \in I$ et $h > 0$, tels que $[c-h, c+h] \subset I$, il existe $\xi \in]c-h, c+h[$ tel que :

$$\int_{c-h}^{c+h} f(x)\,\mathrm{d}x - \frac{h}{3}\left(f(c+h) + f(c-h) + 4f(c)\right) = -\frac{h^5}{90}f^{(4)}(\xi).$$

En déduire une nouvelle preuve du lemme 32 de la page 900.

IV.8.22 On utilise les notations de la partie 5 (page 893).

1. Montrer que l'ordre 1, obtenu pour la méthode des rectangles, ne peut pas être amélioré.
2. Montrer que l'ordre 2, obtenu pour la méthode des rectangles médians ne peut pas être amélioré. *On utilisera l'exercice IV.1.24.*
3. Montrer que l'ordre 2, obtenu pour la méthode des trapèzes, ne peut pas être amélioré.
4. Montrer que si la fonction f est convexe et de classe $\mathcal{C}^2$, on a l'encadrement :

$$A_n(f) \leqslant \int_a^b f(x)\,\mathrm{d}x \leqslant T_n(f).$$

5. Montrer que la méthode de Simpson donne un résultat exact si f est un polynôme de degré au plus 3.
6. Trouver un polynôme de degré 4 tel que, en choisissant convenablement M_4, (32) soit une égalité.

   Soit f un polynôme de degré 4. On note $m_4 := f^{(4)} \in \mathbb{R}$. Montrer que, pour tous $c \in I$ et $h > 0$, tels que $[c-h, c+h] \subset I$, on a :

$$\int_{c-h}^{c+h} f(x)\,\mathrm{d}x - \frac{h}{3}\left(f(c+h) + f(c-h) + 4f(c)\right) = -\frac{m_4}{90}h^5. \tag{45}$$

   En déduire : $\displaystyle\int_a^b f(x)\,\mathrm{d}x - S_n(f) = -\frac{m_4(b-a)^5}{2880\,n^4}$ et conclure.

Probabilités statistiques

Partie V

Probabilités et statistique ont de très nombreuses applications : en biologie, médecine, physique, sciences politiques, finance… Elles restent malheureusement mal comprises. Qui sait définir l'expression « espérance de vie » prononcée, en aveugle, presque chaque jour dans les médias, ou comprend l'importance de l'échantillonnage dans les sondages ?

Ces deux disciplines sont nées au même moment : la statistique vers 1650, quand sont publiées des tables de mortalité (J. Graunt, puis E. Halley en fin du siècle), théorisées par L. Euler (travaillant par exemple sur un problème de pensions des veuves) ; le calcul des probabilités (finies) en 1654 avec B. Pascal à Paris et P. de Fermat à Toulouse (succédant à quelques essais : L. B. Pacioli au XVe, J. Cardan au XVIe, Galilée). Il s'agit au début d'une tentative d'approche scientifique du hasard présent dans les jeux, dont ceux de dés (az-zahr : dé à jouer en arabe).

Les probabilités finies ont été codifiées par C. Huygens et J. Bernoulli. Les probabilités continues furent notamment l'œuvre de A. De Moivre, J.-L. Lagrange, P.-S. Laplace, S. D. Poisson, couronnée par H. Lebesgue (comme application de sa théorie de l'intégration de 1901), et aussi E. Borel et P. Lévy. A. Kolmogorov les a formalisées de manière décisive (1933). Le travail de N. Wiener sur le mouvement brownien (1923) est aussi une clé pour le développement moderne du sujet.

Peu à peu les statistiques ont fait place à « la statistique », en partie autonome des probabilités, même si les deux ont toujours, comme à leur naissance, un aspect appliqué, important et performant. Leur spectre est maintenant très large, avec une forte composante théorique et aussi de nouveaux paradigmes avec l'apparition des mégadonnées et l'apprentissage automatique. Ce domaine des mathématiques est lié à d'autres comme l'analyse (théorie de la mesure, équations aux dérivées partielles) ou la théorie des nombres. Des outils théoriques sophistiqués sont apparus à la fin du XXe siècle : équations différentielles stochastiques (K. Ito, W. Döblin), calcul de Malliavin. Ils ont d'importantes retombées applicatives, en finance par exemple (formule de Black-Scholes pour la valeur des options).

Ce volume initie au sujet : d'un côté la statistique descriptive, de l'autre les probabilités finies. La première est un outil de description, classification et simplification de données pouvant être très volumineuses. L'essentiel est de comprendre comment, partant d'une situation concrète, on en extrait des propriétés significatives avec des outils comme la moyenne, la variance, la corrélation… Pour les probabilités, le cas fini suffit à introduire de nombreuses idées sans avoir besoin de sortir l'artillerie lourde qu'on utilisera dans les volumes suivants. Il est fondamental de surmonter une trop étroite

vision « combinatoire » et de faire l'effort de comprendre les concepts fondamentaux de probabilité conditionnelle et d'indépendance.

Dans le module V.1, des outils de statistique descriptive sont présentés. Il s'agit de montrer comment on peut représenter de manière synthétique, par des tableaux, des graphiques ou des grandeurs moyennes, des données brutes qui peuvent être volumineuses et difficiles à interpréter *a priori*. Dans le modules V.2 et V.3 une théorie élémentaire des probabilités est développée. Il s'agit de présenter comment une analyse mathématique et rigoureuse peut être appliquée à des phénomènes aléatoires. On parle de phénomène aléatoire lorsque le résultat d'une expérience n'est pas prévisible, comme lors du lancer d'un dé ou du tirage du loto. On se limitera dans ce module à des expériences aléatoires qui ne peuvent donner qu'un nombre fini de résultats possibles, ce qui simplifie grandement la théorie, mais limite aussi quelque peu sa portée. La généralisation à des expériences plus complexes (comportant un nombre infini de résultats possibles) sera menée dans le cours L3. Le but de ce module est de montrer que ces phénomènes, bien qu'aléatoires, sont soumis aux règles de la logique et de la raison, et que leur analyse est possible. Le module V.2 est consacré à l'étude des événements qui peuvent arriver ou non lors d'une expérience aléatoire (faire au moins un Pile lors d'une suite de lancers de pièces par exemple) et sur la manière dont on peut évaluer quantitativement leur vraisemblance. Le module V.3 est consacré à l'étude des variables aléatoires, qui sont des quantités dont les valeurs prises dépendent du résultat d'une expérience aléatoire (le nombre de Piles lors d'une suite de lancers de pièces par exemple).

Statistique descriptive V.1

1 Introduction à la statistique descriptive

La vision moderne de la statistique est intimement liée à la prise de décision. On peut dire que la théorie statistique est un outil d'aide à la décision car elle aide à visualiser, à comprendre et à analyser des données. Une branche de la statistique, appelée statistique descriptive, a effectivement pour but de décrire, classer et simplifier des données qui peuvent être volumineuses, de les représenter de manière synthétique sous forme de tableaux ou de graphiques compréhensibles, et d'extraire quelques valeurs importantes qui décrivent les propriétés essentielles des données telles que la moyenne, la variance, la corrélation, etc. Cette partie du livre est consacrée à cette discipline.

Il existe une autre branche de la statistique qui consiste à comparer un jeu de données à un modèle (probabiliste) et à tirer des conclusions de ce jeu de données : c'est la statistique inférentielle. Par exemple, vous savez tous ce qu'est un sondage : on interroge $1\,000$ personnes, 550 déclarent vouloir voter pour M. X, et 450 pour M. Y. Tout le monde est capable d'en déduire l'estimation « 55% de la population a pour intention de voter X ». La vraie question qui se pose est : quelle confiance puis-je avoir en cette estimation ? Comment puis-je quantifier ou minimiser cette incertitude ? C'est là que la théorie des probabilités intervient et permet de répondre précisément à ces questions. Ce sera le sujet d'une partie du cours L3.

1.1 Données statistiques

1.1.1 Populations, caractères

À la base de toute étude statistique, il y a une *population*, formée d'entités élémentaires ou individus. Les individus sont souvent des personnes, mais ils peuvent être des choses, comme des entreprises. Sur ces individus on peut observer un ou des *caractères*. Par exemple, considérons la population constituée de l'ensemble des Français. Les caractères qu'on peut observer sont le poids, le revenu, le sexe, l'intention de vote à la prochaine élection… Si la population est constituée de l'ensemble des entreprises françaises, alors les caractères qu'on peut observer sont le nombre d'employés, le chiffre d'affaires, le statut juridique…

On distingue trois grands types de caractères, selon le type de valeurs qui peuvent être prises par ces caractères. Un caractère est dit :
– qualitatif quand les valeurs ne peuvent être ni ordonnées, ni ajoutées (par exemple, sexe, intention de vote) ;
– ordinal, quand les valeurs peuvent être ordonnées mais pas ajoutées (par exemple, opinion sur l'action du gouvernement exprimée sur une échelle de valeurs) ;
– quantitatif, quand les valeurs sont numériques (par exemple, mesures physiques, économiques).
Les valeurs que peut prendre un caractère s'appellent les modalités. Le caractère qualitatif « sexe » a deux modalités : masculin et féminin. La caractère ordinal « opinion sur l'action du gouvernement » peut avoir par exemple quatre modalités : pas satisfait, peu satisfait, satisfait, très satisfait. La caractère quantitatif « poids d'une personne » a pour modalités les entiers naturels. On peut encore distinguer deux types de caractères quantitatifs. Un caractère quantitatif peut avoir un nombre infini de valeurs possibles dans un intervalle réel : on parle alors de caractère continu. Un caractère quantitatif peut aussi ne prendre qu'un petit nombre de valeurs, par exemple le nombre d'enfants d'une famille, on parle alors de caractère discret. Cependant, la frontière entre continu et discret est beaucoup moins claire en pratique qu'en théorie : le poids d'une personne, s'il est arrondi au kilogramme près, peut être considéré comme un caractère discret. Cependant, on peut trouver plus pratique, vu le grand nombre de valeurs possibles, d'adopter le point de vue qu'il s'agit d'un caractère continu.

1.1.2 Données, échantillons

Observons un certain caractère sur des individus d'une population. Lorsqu'on collecte ces observations, on obtient des *données*. Deux types d'observations peuvent être menées. On dira qu'on effectue un *recensement* si les observations portent sur tous les individus de la population. Par exemple, on peut recenser les âges de tous les élèves d'une classe. Mais il arrive souvent qu'on ne puisse pas obtenir toutes les données d'une population, par exemple les âges de tous les Français. On peut cependant souvent obtenir un ensemble partiel de données. Un tel sous-ensemble incomplet s'appelle un *échantillon*. Le cardinal de l'échantillon est couramment appelé *taille* de l'échantillon.

1.1.3 Échantillonnage et erreur d'échantillonnage

On dispose en général seulement d'échantillons. En effet, il est rare de faire un recensement complet pour collecter toutes les données d'une population. Des considérations économiques peuvent dicter le nombre d'observations qu'on peut faire. Mais il existe des motivations non économiques.
1) Un échantillonnage peut être plus précis, car effectué sur une durée plus courte. Songez par exemple aux sondages (dits média-métriques) faits sur les émissions de télévision. Il faudrait des mois, voire des années pour collecter les avis de tous les téléspectateurs. Or, d'une part, les directeurs de chaînes de télévision veulent avoir très rapidement une idée de la manière dont est perçue une émission pour pouvoir réagir efficacement. D'autre part, les avis des téléspectateurs évoluent au cours du temps. Si l'on fait un recensement complet, celui-ci s'étalera sur des mois, et les premières personnes interrogées peuvent avoir changé d'avis pendant ce laps de temps.

2) Une autre raison justifiant un échantillonnage et pas un recensement a trait au fait que l'action de faire une observation peut endommager l'individu. Par exemple, considérez un lot d'ampoules électriques sorties d'une usine. On souhaite connaître la durée de vie de ces ampoules pour pouvoir la faire figurer sur les emballages. Or la seule méthode de mesure consiste à laisser allumées les ampoules jusqu'à ce qu'elles grillent. On ne va évidemment pas détruire tout le lot d'ampoules ! Dans cette situation, on ne vas pas tester toutes les ampoules, mais seulement un petit nombre d'entre elles.

Malheureusement, un échantillon peut ne pas être le reflet de la population. Les statisticiens sont bien conscients de cette possibilité d'*erreur d'échantillonnage*, que est due au choix particulier des individus sélectionnés pour les observations. Tout l'art de la statistique consiste à quantifier et minimiser cette erreur. Par exemple, vous souhaitez connaître les intentions de vote des Français pour une certaine élection. Vous faites ce sondage par téléphone fixe dans la journée. Cela introduit un biais dans le sens où vous n'allez collecter des réponses que de personnes qui ont un téléphone fixe, qui ne travaillent pas et qui sont chez elles à ce moment-là. La sous-représentation des individus qui sont difficiles à observer conduit à un biais de non réponse. Il faut être aussi conscient de l'existence d'autres sources d'erreurs qui ne sont pas liées à l'échantillonnage lui-même. Par exemple, toujours dans le cas d'un sondage d'opinion, certaines personnes peuvent être tentées de cacher la vérité pour impressionner leur interlocuteur, ou pour se fondre dans la masse, ou pour tout autre motif.

1.1.4 Échantillonnage aléatoire

Comme on vient de le voir, la détermination d'un *échantillon représentatif* d'une population est un problème. Il convient de sélectionner les individus sur lesquels les observations seront effectuées avec rigueur et honnêteté. On peut envisager des méthodes rapides mais pas très honnêtes : si vous voulez des informations sur l'âge des Français, il vaut mieux éviter de demander uniquement à vos amis, car vous risquez d'avoir une sur-représentation de votre propre classe d'âge. Parmi toutes les méthodes possibles, la méthode d'*échantillonnage aléatoire* est la plus couramment utilisée. Elle consiste à organiser une loterie géante dans laquelle tous les individus sont des numéros qui sont ensuite tirés au hasard (en utilisant un générateur de nombres aléatoires) jusqu'à obtenir un échantillon de la taille voulue. Ainsi tous les individus ont la même chance d'être sélectionnés pour observation. Cette méthode présente deux avantages. Premièrement, cette méthode est la plus honnête qu'on puisse envisager car elle ne présente aucun biais psychologique ou systématique. Deuxièmement, on peut quand même par malchance tomber sur un échantillon non représentatif, mais le gros avantage de la méthode est qu'on peut quantifier l'erreur d'échantillonnage en utilisant la théorie des probabilités. C'est surtout ce dernier point qui intéresse les statisticiens.

1.2 Représentation des données

Il arrive très souvent, en particulier avec les moyens informatiques modernes, qu'un échantillonnage fournisse une quantité de données très importante. Pour pouvoir communiquer les résultats d'une étude de ces données, il faut réduire ces données sous une forme utile. Par forme utile, on entend une forme qui soit compréhensible par tout le monde et qui illustre bien les caractéristiques importantes de l'échantillon.

1.2.1 Tableaux

Lorsque l'on dépouille des données quantitatives, on se trouve en général face à un énorme tableau de nombres. L'ordre dans lequel on a recueilli les données ne présentant pas d'intérêt particulier, on peut essayer d'organiser un peu les données brutes en groupant les observations par paquets, qu'on appelle *classes*. Dans le cas d'un caractère discret, ces classes sont simplement les modalités. On peut alors associer deux nombres à chaque classe. L'*effectif* d'une classe est le nombre d'observations qui tombent dans cette classe. La *fréquence* d'une classe est le rapport de l'effectif de cette classe sur le nombre total d'observations.
Prenons par exemple les nombres d'enfants par famille collectés dans un immeuble de 80 familles :

| | | |

Le nombre de classes ne doit pas être trop grand pour que le nouveau tableau soit suffisamment clair, mais pas trop petit pour qu'il n'y ait pas de perte d'information trop importante. Il faut que toutes les observations soient recouvertes par ces classes. Très souvent les classes sont des intervalles consécutifs disjoints, dont l'union contient les valeurs minimum et maximum des observations. Afin d'assurer leur caractère disjoint, on peut supposer que les intervalles sont semi-ouverts à droite et semi-fermés à gauche. Souvent, mais pas toujours (voir ci-dessous), les intervalles ont la même longueur. L'*amplitude* d'une classe est la largeur de l'intervalle. L'*effectif* d'une classe est le nombre d'observations qui tombent dans cette classe. La *fréquence* d'une classe est le rapport de l'effectif de cette classe sur le nombre total d'observations. Par exemple, on prend les données de l'exemple 2 et on les regroupe en 8 classes :

Durée de vie (en heures)	Effectif	Fréquence
De 1 200 à 1 400	3	0,0250
De 1400 à 1 600	27	0,2250
De 1 600 à 1 800	8	0,0667
De 1 800 à 2 000	7	0,0583
De 2 000 à 2 200	23	0,1917
De 2 200 à 2 400	28	0,2333
De 2 400 à 2 600	23	0,1917
De 2 600 à 2 800	1	0,0083
	120	1

Exemple 2 : Distribution des ampoules selon leur durée de vie.

Ce simple regroupement met en évidence l'existence d'un creux dans la distribution des durées de vie entre 1 600 et 2 000.

1.2.2 Graphiques

Dans le cas de données discrètes on trace la plupart du temps un *diagramme en bâtons*. Des segments de droite verticaux sont dessinés. Chaque segment correspond à une classe (i.e. une modalité). La valeur de la classe est l'abscisse du segment, l'ordonnée de l'extrémité inférieure du segment est 0 et l'ordonnée de l'extrémité supérieure est la fréquence de la classe. L'exemple 1 peut ainsi se représenter sous la forme du diagramme en bâtons donné sur la figure V.1.1. On peut aussi trouver des diagrammes en bâtons dont la hauteur des bâtons est l'effectif de la classe plutôt que la fréquence.

FIGURE V.1.1 – Distribution (diagramme en bâtons) du nombre d'enfants par famille.

Dans le cas de données continues, on regroupe d'abord les données par classes comme dans le tableau « Exemple 2 : Distribution des ampoules selon leur durée de vie ». On trace alors un *histogramme* constitué de rectangles verticaux. Les bases des rectangles sont déterminées par les classes. Les hauteurs de rectangles doivent être telles que les surfaces des rectangles sont proportionnelles aux fréquences des classes correspondantes. Le travail est simple lorsque la largeur de chaque classe est la même : la hauteur d'un rectangle est alors prise égale à la fréquence de la classe correspondante. C'est le cas de l'exemple 2, et on obtient alors la figure V.1.2. On peut aussi trouver des histogrammes dont la hauteur des rectangles est l'effectif de la classe plutôt que la fréquence.

FIGURE V.1.2 – Histogramme des durées de vie des ampoules.

Il arrive qu'on ait affaire à des classes non régulières. Ainsi, dans l'exemple 2, on peut subdiviser les trois classes de 2 000 à 2 600 en six classes. Le nouveau tableau est :

Durée de vie (en heures)	Effectif	Fréquence	Hauteur des rectangles
De 1 200 à 1 400	3	0,0250	0,0250
De 1 400 à 1 600	27	0,2250	0,2250
De 1 600 à 1 800	8	0,0667	0,0667
De 1 800 à 2 000	7	0,0583	0,0583
De 2 000 à 2 100	9	0,0750	0,1500
De 2 100 à 2 200	14	0,1167	0,2333
De 2 200 à 2 300	11	0,0917	0,1834
De 2 300 à 2 400	17	0,1417	0,2834
De 2 400 à 2 500	13	0,1083	0,2166
De 2 500 à 2 600	10	0,0830	0,1660
De 2 600 à 2 800	1	0,0083	0,0083
	120	1	

Exemple 2 : Distribution des ampoules selon leur durée de vie.

Le tracé d'un histogramme doit alors prendre en compte la non uniformité des largeurs des classes. Pour cela, on prend la plus grande des largeurs (ou amplitudes) comme largeur de référence, et on multiplie la fréquence de chaque classe par le rapport de cette largeur maximale sur la largeur de la classe. On obtient ainsi la dernière colonne du tableau précédent, et l'histogramme correspondant est tracé sur la figure V.1.3, qui ressemble beaucoup à l'histogramme de la figure V.1.2, mais avec un niveau de détail plus fin.

FIGURE V.1.3 – Histogramme des durées de vie des ampoules.

1.2.3 Données qualitatives

Dans le cas de données qualitatives, on peut encore tracer des diagrammes en bâtons. Par exemple, considérons le statut marital du chef de famille relevé dans les 80 familles habitant un immeuble. Il existe cinq catégories : célibataire (C), marié (M), pacsé (P), divorcé (D), et veuf (V).

M	C	C	C	M	V	D	P	C	C	C	M	M	M	C	D
M	P	V	M	M	C	C	C	M	M	C	D	M	P	V	C
C	C	M	P	M	C	M	C	C	C	D	P	V	C	P	M
P	C	C	V	V	C	P	M	P	M	C	V	C	M	M	C
C	M	C	D	C	M	V	P	M	V	V	D	D	C	D	M

Exemple 3 : Statut marital du chef de famille dans les 80 familles de l'immeuble.

Statut marital	Effectif	Fréquence
C	29	0,3625
P	10	0,1125
M	23	0,2875
D	8	0,1000
V	10	0,1125
	80	1

Exemple 3 : Distribution des familles selon le statut marital du chef de famille.

On obtient alors le diagramme en bâtons tracé sur la figure V.1.4(a). Il existe cependant beaucoup d'autres manières de visualiser les résultats. On peut ainsi tracer des *diagrammes à secteurs* (camemberts). La distribution des familles selon le statut marital du chef de famille peut ainsi être représentée selon la figure V.1.4(b), qui est une représentation équivalente au diagramme en bâtons de la figure V.1.4(a).

FIGURE V.1.4 – Distribution du statut marital des individus.
Diagramme en bâtons (a) et diagramme en secteurs (b).

2 Statistique descriptive univariée

Les tableaux ou les diagrammes présentés dans la section précédente sont utiles, mais ils ne sont que des outils de visualisation. On cherche souvent, à partir des données quantitatives collectées, à extraire des caractéristiques chiffrées simples, des nombres qui révèlent les propriétés importantes de l'échantillon ou de la population. Donnons nous un échantillon de n valeurs réelles que nous notons $x = (x_i)_{i\in[\![1,n]\!]}$.

2.1 Mesures de tendance centrale

On recherche souvent en particulier un nombre représentant la tendance centrale d'une distribution, c'est-à-dire la moyenne. On peut définir de plusieurs manières une notion de valeur « moyenne ». Grosso modo, il s'agit de déterminer la valeur qui est la plus représentative de toutes les données.

2.1.1 Médiane

La médiane est une valeur M telle qu'il y ait autant d'observations supérieures ou égales à M que d'observations inférieures ou égales à M. Grosso modo, la valeur médiane est la valeur qui se trouve « au milieu » des observations. Cependant, ainsi définie, il peut y avoir plusieurs valeurs possibles pour la médiane : prenez un échantillon comportant uniquement deux observations $x_1 < x_2$; n'importe quelle valeur prise entre x_1 et x_2 convient. Pour calculer précisément la médiane, on commence par ordonner l'échantillon $(x_i)_{i\in[\![1,n]\!]}$ par ordre croissant, et on note $(x_{(i)})_{i\in[\![1,n]\!]}$ l'échantillon ordonné tel que :

$$\{x_1, \ldots, x_n\} = \left\{x_{(1)}, \ldots, x_{(n)}\right\} \text{ avec } x_{(1)} \leqslant \cdots \leqslant x_{(n)} .$$

Ainsi, si l'échantillon est constitué de la suite d'entiers $(5, 2, 4, 2, 6)$, alors l'échantillon ordonné est $(2, 2, 4, 5, 6)$.
Si l'échantillon comporte un nombre impair $2p + 1$ d'observations, alors la médiane est :

$$\mathrm{Med}(x) := x_{(p+1)} .$$

Si l'échantillon comporte un nombre pair $2p$ d'observations, alors la médiane est :

$$\mathrm{Med}(x) := \frac{x_{(p)} + x_{(p+1)}}{2} .$$

Ainsi, dans l'exemple 1, la médiane est 1 (car on a $x_{(32)} = \cdots = x_{(44)} = 1$, et la taille de l'échantillon est 80, donc $\mathrm{Med}(x) = (x_{(40)} + x_{(41)})/2 = 1$). On voit que la médiane est peu sensible aux valeurs extrêmes. En effet, si au lieu de 5 familles à 5 enfants, on avait 5 familles avec 15 enfants, cela ne changerait rien à la médiane. On voit aussi que la médiane n'est pas forcément une modalité. En effet, si au lieu de 31 familles à 0 enfants et 14 familles à trois enfants, on avait 40 familles à 0 enfants et 5 familles à trois enfants, alors la médiane serait $0{,}5$ (car alors on aurait $x_{(40)} = 0$ et $x_{(41)} = 1$).

2.1.2 Mode (ou classe modale)

Le mode est la valeur la plus fréquente que l'on trouve dans l'échantillon.
Dans le cas d'un caractère discret, il suffit de regarder le tableau des fréquences pour déterminer la classe la plus nombreuse. Ainsi, dans l'exemple 2, le mode est 0. Le mode n'est

pas défini de manière unique. On peut très bien trouver plusieurs classes avec le même effectif. On parle alors de distribution multi-modale. Le mode, comme la médiane, est peu sensible aux valeurs extrêmes.
Dans le cas de caractères continus, on considère plutôt la classe modale. Le résultat dépend donc des classes choisies, ce qui rend cette notion peu pratique à utiliser. Dans le cas où les classes n'ont pas les mêmes largeurs, il faut comparer non pas les effectifs des classes, mais leur fréquence normalisée par la largeur des classes. Ainsi, dans l'exemple 2, la classe modale est la classe de 2300 à 2400.

2.1.3 Moyenne arithmétique

On peut définir la moyenne arithmétique d'un échantillon comme étant le barycentre des données, affectées de coefficients égaux pour chaque individu.

Définition 1. La moyenne arithmétique d'un échantillon est définie par :

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{1}$$

On appelle communément $\bar{x}$ la moyenne des x_i. Sur l'exemple 1, on a $\bar{x} = 1{,}575$.

Exercice 1.
Montrer que, dans le cas de données discrètes prenant les valeurs $\{\alpha_1, \ldots, \alpha_p\}$, le tableau des fréquences permet de calculer la moyenne par la formule :

$$\bar{x} = \sum_{i=1}^{p} f_i \alpha_i \,,$$

où f_i est la fréquence de la classe α_i.
Solution. Cette formule se démontre en regroupant les termes de (1) par classes :

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{p}\sum_{i\,|\,x_i=\alpha_j} x_i \\
&= \frac{1}{n}\sum_{j=1}^{p}\sum_{i\,|\,x_i=\alpha_j} \alpha_j = \frac{1}{n}\sum_{j=1}^{p} \alpha_j \operatorname{card}\left\{i \in [\![1,n]\!] \mid x_i = \alpha_j\right\} \\
&= \sum_{j=1}^{p} \alpha_j f_j \,.
\end{aligned}$$

Ainsi, sur l'exemple 1, on a :

$$\bar{x} = \frac{1}{80}\left(0\times 31 + 1\times 13 + 2\times 11 + 3\times 14 + 4\times 6 + 5\times 5\right) = 1{,}575\,.$$

Dans le cas de données continues, il est nécessaire d'avoir toutes les données pour calculer la moyenne exactement. Dans l'exemple 2, on utilise les données complètes du tableau de la page 918 et on trouve $\bar{x} \simeq 2\,031$. Si on ne dispose que du tableau des fréquences et qu'on a perdu les données brutes, alors on estime la moyenne par la formule

$$\bar{x} \simeq \sum_i f_i c_i \,,$$

où c_i est le centre de la i-ème classe et f_i est sa fréquence. Par exemple, si on ne dispose que du tableau des fréquences de la page 919 dans l'exemple 2, alors on trouve

$$\begin{aligned}\bar{x} \simeq\, & 0{,}0250 \times 1300 + 0{,}2250 \times 1500 + 0{,}0667 \times 1700 + 0{,}0583 \times 1900 \\ & + 0{,}1917 \times 2100 + 0{,}2333 \times 2300 + 0{,}1917 \times 2500 + 0{,}0083 \times 2700\,,\end{aligned}$$

et on trouve $\bar{x} \simeq 2\,035$, qui est différent mais tout de même très proche de la valeur $\bar{x} \simeq 2\,031$ obtenue avec toutes les données.
Que ce soit pour des données discrètes ou continues, la moyenne se comporte de manière satisfaisante lorsqu'on additionne ou lorsqu'on fusionne des données.

Proposition 1 (Linéarité de la moyenne).

1. Considérons la situation où on dispose de deux échantillons $x^{(1)} = (x_i^{(1)})_{i\in[\![1,n]\!]}$ et $x^{(2)} = (x_i^{(2)})_{i\in[\![1,n]\!]}$ de même taille n et de moyennes respectives $\bar{x}^{(1)}$ et $\bar{x}^{(2)}$. Construisons l'échantillon $x = (x_i)_{i\in[\![1,n]\!]}$ en posant $x_i = ax_i^{(1)} + bx_i^{(2)}$ pour tout $i \in [\![1, n]\!]$, où a et b sont deux constantes réelles.
   La moyenne de l'échantillon x est :
$$\bar{x} = a\bar{x}^{(1)} + b\bar{x}^{(2)}\,.$$
2. Considérons la situation où on dispose de deux échantillons $x^{(1)} = (x_i^{(1)})_{i\in[\![1,n_1]\!]}$ et $x^{(2)} = (x_i^{(2)})_{i\in[\![1,n_2]\!]}$ de tailles n_1 et n_2 et de moyennes respectives $\bar{x}^{(1)}$ et $\bar{x}^{(2)}$. Réunissons les deux échantillons et construisons l'échantillon global $x = (x_i)_{i\in[\![1,n]\!]}$ de taille $n = n_1+n_2$ en posant $x_i = x_i^{(1)}$ si $i \leqslant n_1$ et $x_i = x_{i-n_1}^{(2)}$ si $n_1 + 1 \leqslant i \leqslant n$. La moyenne de l'échantillon x est :
$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}\,.$$

La seconde propriété est importante en pratique. Elle permet, lorsqu'on fusionne les résultats issus d'échantillons différents, d'obtenir la moyenne de l'échantillon global sans avoir à refaire tous les calculs. La médiane et le mode ne vérifient pas ces propriétés de linéarité (voir l'exercice V.1.0 de la page 935).
Il existe une autre raison pour laquelle la moyenne arithmétique est souvent utilisée. Comme le montre la proposition 2 de la page suivante, c'est le nombre réel qui minimise l'erreur quadratique :

$$\mathcal{E}(\mu) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\,. \tag{2}$$

L'erreur quadratique $\mathcal{E}(\mu)$ est la moyenne des carrés des distances entre les x_i et le nombre réel μ. L'erreur $\mathcal{E}(\mu)$ est d'autant plus petite que les données sont toutes concentrées autour de μ.

Proposition 2. L'erreur quadratique (2) atteint son unique minimum en $\mu = \bar{x}$.

En d'autres termes, la moyenne $\bar{x}$ est la constante qui approche au mieux les données au sens des moindres carrés.

Démonstration. En fait, l'erreur quadratique, en tant que fonction de μ, est une fonction quadratique. Il suffit de développer la formule définissant $\mathcal{E}(\mu)$ et d'identifier les coefficients :

$$\mathcal{E}(\mu) = \mu^2 + 2a_1\mu + a_0 \,,$$

avec :

$$a_1 = -\frac{1}{n}\sum_{i=1}^{n} x_i = -\bar{x} \,, \qquad a_0 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \,.$$

Une telle fonction quadratique admet un unique minimum en $\mu = -a_1 = \bar{x}$. ■

Mais attention, la moyenne n'a pas que des avantages, elle a aussi quelques inconvénients qu'il convient de garder en mémoire. La moyenne est en effet sensible aux valeurs extrêmes d'un échantillon. Par exemple, si vous cherchez la fortune moyenne des Français à partir d'un échantillon de $1\,000$ personnes, et que l'une d'entre elles possède un milliard d'euros, alors vous trouvez que, quelles que soient les fortunes des 999 autres, la fortune moyenne est supérieure à un million d'euros, puisqu'elle vérifie :

$$\bar{x} = \frac{1}{1\,000}\sum_{i=1}^{1\,000} x_i \geqslant \frac{1}{1\,000} \times 10^9 = 10^6 \,.$$

2.2 Mesures de dispersion

On vient d'examiner différentes mesures de la tendance centrale d'un échantillon, et on a pu déterminer que la moyenne arithmétique et le mode étaient souvent les mesures de tendance centrale les plus pratiques et manipulables. On va maintenant chercher une mesure de la variabilité d'un échantillon, c'est-à-dire un nombre qui est d'autant plus grand que les données de l'échantillon sont dispersées. On va en fait proposer et discuter deux mesures de dispersion, une liée à la moyenne arithmétique et l'autre liée à la médiane.

2.2.1 Variance et écart-type

La dispersion d'un échantillon peut se visualiser en considérant les écarts à la moyenne, c'est-à-dire l'échantillon $(x'_i)_{i\in[\![1,n]\!]}$ avec $x'_i = x_i - \bar{x}$. On cherche à obtenir une valeur unique représentative de ces écarts. Notez que ce sont des nombres positifs ou négatifs. On ne va pas prendre la moyenne arithmétique des écarts, car elle est nulle quel que soit l'échantillon d'après la propriété de linéarité de la moyenne arithmétique. On va donc prendre les carrés des écarts et calculer leur moyenne arithmétique. On obtient ainsi la *variance* de l'échantillon.

Définition 2. La variance d'un échantillon x est définie par :

$$\mathrm{Var}(x) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \,. \tag{3}$$

La variance $\mathrm{Var}(x)$ n'est autre que la valeur minimale de l'écart quadratique moyen comme le montre la proposition 2 de la page précédente.

Dans certains ouvrages, vous trouverez une définition légèrement différente de la variance, où le facteur $1/n$ dans (3) est remplacé par $1/(n-1)$. Il y a une raison à cela, qui a trait à des résultats de statistique inférentielle reliant la variance d'un échantillon à la variance théorique de la loi statistique sous-jacente qu'on verra dans le cours L3. Cependant il n'y a pas lieu de se compliquer la vie à ce niveau, la définition (3) est celle qui est la plus couramment utilisée, et c'est celle que nous utiliserons dans la suite.

Proposition 3 (Théorème de Koenig).
La variance d'un échantillon x est égale à :

$$\mathrm{Var}(x) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 \,. \tag{4}$$

Cette proposition signifie que la variance est la moyenne des carrés moins le carré de la moyenne. En fait, l'expression (3) est utile pour montrer que la variance est une quantité positive qui augmente avec la dispersion des données. L'expression (4) est utile pour calculer pratiquement la variance d'un échantillon donné.

Démonstration. On développe les carrés qui interviennent dans l'expression de la variance :

$$\begin{aligned}
\mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \frac{2}{n}\left(\sum_{i=1}^{n} x_i\bar{x}\right) + \frac{1}{n}\left(\sum_{i=1}^{n}\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}\frac{2}{n}\left(\sum_{i=1}^{n} x_i\right) + \bar{x}^2\frac{1}{n}\left(\sum_{i=1}^{n} 1\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 \,.
\end{aligned}$$

La variance n'est pas linéaire, mais elle satisfait les propriétés suivantes.

1. Si l'on connaît la moyenne $\bar{x}$ et la variance $\mathrm{Var}(x)$ d'un échantillon $x = (x_i)_{i\in[\![1,n]\!]}$, et si $\lambda \in \mathbb{R}$, alors la moyenne $\bar{x}^{(\lambda)}$ et la variance $\mathrm{Var}(x^{(\lambda)})$ de l'échantillon $x^{(\lambda)} = (\lambda x_i)_{i\in[\![1,n]\!]}$ sont :

$$\bar{x}^{(\lambda)} = \lambda\bar{x}\,, \qquad \mathrm{Var}(x^{(\lambda)}) = \lambda^2\mathrm{Var}(x)\,. \tag{5}$$

2. Si l'on connaît les moyennes $\bar{x}^{(1)}$ et $\bar{x}^{(2)}$ et les variances $V_1 = \mathrm{Var}(x^{(1)})$ et $V_2 = \mathrm{Var}(x^{(2)})$ de deux échantillons $x^{(1)}$ et $x^{(2)}$ de tailles n_1 et n_2, alors la variance de l'échantillon global de taille $n = n_1 + n_2$ peut se calculer uniquement à

partir de ces valeurs. On part de (4) :

$$\begin{aligned}\mathrm{Var}(x) &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 \\ &= \frac{n_1}{n}\left(V_1 + [\bar{x}^{(1)}]^2\right) + \frac{n_2}{n}\left(V_2 + [\bar{x}^{(2)}]^2\right) - \left(\frac{n_1}{n}\bar{x}^{(1)} + \frac{n_2}{n}\bar{x}^{(2)}\right)^2 .\end{aligned}$$

3. La variance d'un échantillon est nulle si, et seulement si, toutes les données sont égales (car une somme de termes positifs ou nuls est nulle si, et seulement si, tous les termes sont nuls).

La variance présente un inconvénient visible dans l'équation (5) : si les données s'expriment en une unité physique (des heures dans l'exemple 2), la moyenne arithmétique s'exprime aussi dans cette unité, mais la variance s'exprime dans l'unité carrée (en heure^2 dans l'exemple). C'est pourquoi on a introduit la notion d'écart-type qui est simplement la racine carrée de la variance.

Définition 3. L'écart-type d'un échantillon x est défini par :

$$\sigma := \sqrt{\mathrm{Var}(x)}\,.$$

Si les données s'expriment toutes dans une certaine unité, alors l'écart-type s'exprime dans la même unité.

2.2.2 Fractiles

Les fractiles sont un autre moyen de quantifier la dispersion de données quantitatives. Le fractile à θ % d'un échantillon est la valeur qui sépare la fraction θ % des plus petites données de la fraction $(100-\theta)$ % des plus grandes données. Ainsi, le fractile à 50 % n'est autre que la médiane.

Les fractiles à 25 %, 50 % et 75 % sont les trois quartiles. Sur l'exemple 2, les trois quartiles sont : $1\,602$, $2\,131$ et $2\,341$. Une mesure de la dispersion d'un échantillon est l'espace inter-quartile qui est la différence entre le troisième quartile et le premier quartile. C'est donc la largeur de l'intervalle qui contient l'échantillon duquel on a retiré les 25 % plus grandes valeurs et les 25 % plus petites valeurs. Qualitativement, plus l'espace inter-quartile est grand, plus la dispersion des données est grande. C'est donc bien une mesure de la dispersion, à l'instar de l'écart-type. On peut aussi voir que l'espace inter-quartile est moins sensible aux valeurs extrêmes que l'écart-type. On peut donc dire que l'espace inter-quartile est à la médiane ce que l'écart-type est à la moyenne arithmétique. Sur l'exemple 2, on trouve une médiane de $2\,131$ et un espace inter-quartile de 739. La moyenne est égale à $2\,031$ et l'écart-type est 389 environ.

Un visualisation très populaire des fractiles d'un échantillon peut être obtenue avec une boîte à moustaches (traduction française du terme « Box and Whiskers Plot », ou en abrégé « Box Plot »). Il s'agit d'une représentation schématique utile pour connaître leur allure (équilibrée ou non, concentrée ou étalée, symétrique ou non…) : les bords de la boîte correspondent aux premier et troisième quartile. La boîte est divisée en deux au niveau du second quartile (médiane). Les extrémités des moustaches correspondent aux fractiles à 5

et $95\,\%$. On peut ensuite terminer le graphique, en faisant figurer par des points les données qui sont en dehors des moustaches (on appelle ces données les outliers).

Attention, les boîtes à moustaches ne sont pas encore bien normalisées, vous pouvez trouver d'autres définitions pour les positions des moustaches, on peut prendre le premier et le 99ème centiles, ou le min et le max de l'échantillon (auquel cas il n'y a pas d'outlier).

Une boîte avec des moustaches courtes indique que l'échantillon est assez concentré autour de sa médiane. Au contraire des moustaches longues indiquent que l'échantillon est assez dispersé. Sur la figure V.1.5 on dessine la boîte à moustaches correspondant à l'exemple 2. On constate que l'échantillon n'est pas équilibré, la médiane n'est pas vraiment au milieu des premier et troisième quartiles.

FIGURE V.1.5 – Boîte à moustaches des durées de vie des ampoules.

3 Statistique descriptive bivariée

On a étudié jusqu'à maintenant des échantillons simples dans le sens où les données portaient sur un seul caractère des individus sur lesquels les observations ont été effectuées. On va maintenant étudier des séries statistiques doubles : les observations portent sur deux caractères, par exemple poids et taille d'une personne, ou salaire et ancienneté d'un employé, etc. Les données $(x_i, y_i)_{i\in[\![1,n]\!]}$ se présentent alors sous la forme d'un tableau à deux colonnes, comme par exemple :

Ancienneté (années)	Salaires annuels (kE)	Ancienneté (années)	Salaires annuels (kE)
1	19,41	3	19,74
1	18,49	4	20,81
2	18,98	6	20,31
2	18,57	5	20,78
10	20,54	7	21,36
1	17,96	6	20,04
10	21,12	9	20,58
5	21,34	1	18,71
5	20,13	9	21,29
10	19,55	3	18,25

Exemple 4 : Ancienneté et salaire de 20 employés d'une entreprise (kE veut dire kilo-euro, qui est une unité équivalent à mille euros).

On peut bien sûr mener une étude statistique de chacun des caractères séparément, calculer la moyenne et la variance de chacune des séries simples. Sur l'exemple 4, on trouve :

$$\bar{x} = 5\,, \quad \mathrm{Var}(x) \simeq 10{,}7\,, \quad \bar{y} = 19{,}898\,, \quad \mathrm{Var}(y) \simeq 1{,}223\,.$$

On veut ici extraire des informations sur la distribution jointe des deux caractères et étudier leur dépendance. Comme dans le cas de séries simples, on peut commencer par établir des tables de fréquences. Une table de fréquences prend l'aspect d'un tableau à double entrée :

Salaire / Ancienneté	17,5-18,5	18,5-19,5	19,5-20,5	20,5-21,5	Total
1-2	0,1	0,2	0	0	0,3
3-4	0,05	0	0,05	0,05	0,15
5-6	0	0	0,15	0,1	0,25
7-8	0	0	0	0,05	0,05
9-10	0	0	0,05	0,2	0,25
Total	0,15	0,2	0,25	0,4	1

Table des fréquences pour l'exemple 4.

On lit sur ce tableau (première case en haut à gauche) que $10\ \%$ des employés ont entre 1 et 2 ans d'ancienneté et un salaire compris entre $17{,}5$ et $18{,}5$ kilo-euros. La dernière ligne et la dernière colonne représentent les distributions marginales, c'est-à-dire les distributions des caractères x et y seuls. Le 1 en bas à droite peut être vu comme un contrôle des calculs faits (la somme des fréquences des distributions marginales doit être égale à 1).
On veut maintenant mener une analyse quantitative de la dépendance des deux caractères d'une série statistique double. Cette analyse peut prendre essentiellement deux formes qui sont intimement liées. L'analyse de régression donne des outils de prédiction du comportement d'un caractère si on connaît la valeur d'un autre caractère. L'analyse de corrélation a pour objet la force de la relation entre les deux caractères. Ces analyses sont utiles par exemple dans des applications en finance ou en biologie. Dans ces applications on cherche à identifier les facteurs qui influent sur d'autres, dans quel sens et dans quelle mesure le poids d'une personne influe sur son taux de cholestérol, etc.

3.1 Ajustement linéaire par moindre carrés

Dans toute cette partie on considère un échantillon du type $(x_i, y_i)_{i\in[\![1,n]\!]}$. On peut associer à chaque donnée (x_i, y_i) un point du plan, et on peut représenter un échantillon de n données comme un nuage de n points. Le nuage correspondant à l'exemple 4 est dessiné sur la figure V.1.6. On constate que le nuage a une forme allongée, et on peut alors essayer de dessiner une droite passant au milieu de ces points. Cette droite, appelée droite de régression linéaire, est un moyen de représenter la dépendance des deux caractères et la détermination de son équation constitue l'objet de ce paragraphe.

FIGURE V.1.6 – Distribution jointe du salaire et de l'ancienneté des employés et droite de régression linéaire.

La détermination de la droite de régression linéaire peut être effectuée empiriquement, à la main, par ajustement graphique, mais on peut

aussi donner des bases rigoureuses pour une méthode d'ajustement linéaire. La méthode des moindre carrés permet de déterminer la meilleure droite passant par le nuage de points constitué par une série statistique double. On cherche la droite de la forme $y = ax + b$ minimisant l'erreur quadratique :

$$\mathcal{E}(a,b) := \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2\,, \tag{6}$$

qui est la somme des distances au carré entre les points (x_i, y_i) et les points $(x_i, ax_i + b)$ de mêmes abscisses situés sur la droite $y = ax + b$.

Proposition 4. Si $n \geqslant 2$ et les x_i ne sont pas tous égaux, alors il existe une unique droite $y = ax+b$ minimisant l'erreur quadratique d'ajustement linéaire (6). Cette droite est appelée droite de régression linéaire et elle a pour équation $y = a_0x + b_0$ avec :

$$a_0 = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}\,, \qquad b_0 = \bar{y} - a_0\bar{x}\,,$$

où $\mathrm{Cov}(x,y)$ est la covariance de x et y :

$$\mathrm{Cov}(x,y) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\,. \tag{7}$$

La droite de régression linéaire s'écrit donc $y = a_0(x - \bar{x}) + \bar{y}$. On peut noter que la droite passe forcément par le point $(\bar{x}, \bar{y})$. On va étudier de manière fine les propriétés de la covariance dans le prochain paragraphe. On peut déjà dire que, comme pour la variance, on dispose d'une formule alternative pour la covariance :

$$\mathrm{Cov}(x,y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_iy_i\right) - \bar{x}\bar{y}\,,$$

qu'on utilise en pratique pour calculer une covariance. Cette formule s'obtient en développant le facteur à l'intérieur de la somme de (7).

Démonstration. On développe l'expression de l'erreur quadratique, ce qui donne une fonction quadratique en (a,b) :

$$\mathcal{E}(a,b) = c_{2,0}a^2 + 2c_{1,1}ab + c_{0,2}b^2 + 2c_{1,0}a + 2c_{0,1}b + c_{0,0}\,,$$

avec :

$$c_{2,0} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \mathrm{Var}(x) + \bar{x}^2\,, \qquad c_{1,1} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}\,,$$

$$c_{0,2} = 1\,, \qquad c_{1,0} = -\frac{1}{n}\sum_{i=1}^{n} x_iy_i = -\mathrm{Cov}(x,y) - \bar{x}\bar{y}\,,$$

$$c_{0,1} = -\frac{1}{n}\sum_{i=1}^{n} y_i = -\bar{y}\,, \qquad c_{0,0} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = \mathrm{Var}(y) + \bar{y}^2\,.$$

On veut montrer que cette fonction quadratique a un minimum global, c'est-à-dire qu'on veut identifier un point (a_0, b_0) tel que $\mathcal{E}(a,b) > \mathcal{E}(a_0,b_0)$ dès que $(a,b) \neq (a_0,b_0)$.

Si un tel point existe, alors il doit en particulier vérifier $\mathcal{E}(a, b_0) \geqslant \mathcal{E}(a_0, b_0)$ pour tout a et $\mathcal{E}(a_0, b) \geqslant \mathcal{E}(a_0, b_0)$ pour tout b. Or, si on fixe b, la fonction $a \mapsto \mathcal{E}(a, b)$ est un polynôme du second degré. Un minimum de ce polynôme est un extremum, donc un point a où la dérivée de la fonction $a \mapsto \mathcal{E}(a, b)$ s'annule. De même, si on fixe a, la fonction $b \mapsto \mathcal{E}(a, b)$ est un polynôme du second degré. Un minimum de ce polynôme est un extremum, donc un point b où la dérivée de la fonction $b \mapsto \mathcal{E}(a, b)$ s'annule. On cherche donc un point (a, b) vérifiant $\frac{\partial \mathcal{E}}{\partial a}(a, b) = 0$ et $\frac{\partial \mathcal{E}}{\partial b}(a, b) = 0$, ce qui se traduit par :

$$c_{2,0}a + c_{1,1}b + c_{1,0} = 0\,,$$

$$c_{1,1}a + c_{0,2}b + c_{0,1} = 0\,.$$

Le déterminant de ce système linéaire 2×2 est $c_{2,0}c_{0,2} - c_{1,1}^2 = \mathrm{Var}(x)$. Si $n \geqslant 2$ et les x_i ne sont pas tous égaux, alors $\mathrm{Var}(x) > 0$ ce qui montre que le déterminant est non nul. Par conséquent le système linéaire admet une unique solution, ce qui veut dire qu'il existe un unique point (a_0, b_0) pour lequel la fonction quadratique $\mathcal{E}$ vérifie $\frac{\partial \mathcal{E}}{\partial a}(a_0, b_0) = 0$ et $\frac{\partial \mathcal{E}}{\partial b}(a_0, b_0) = 0$ et ce point est situé en :

$$a_0 = \frac{c_{1,1}c_{0,1} - c_{0,2}c_{1,0}}{c_{2,0}c_{0,2} - c_{1,1}^2}\,, \qquad b_0 = \frac{c_{1,1}c_{1,0} - c_{2,0}c_{0,1}}{c_{2,0}c_{0,2} - c_{1,1}^2}\,.$$

On obtient ainsi le point (a_0, b_0) décrit dans l'énoncé de la proposition. Mais il faut encore montrer que ce point est un minimum. La fonction quadratique $\mathcal{E}$ peut se développer de manière exacte autour du point (a_0, b_0) :

$$\begin{aligned}
\mathcal{E}(a_0 + a, b_0 + b) &= \mathcal{E}(a_0, b_0) + c_{2,0}a^2 + c_{0,2}b^2 + 2c_{1,1}ab \\
&\quad + 2c_{2,0}a_0a + 2c_{1,1}(a_0b + b_0a) + 2c_{0,2}b_0b + 2c_{1,0}a + 2c_{0,1}b \\
&= \mathcal{E}(a_0, b_0) + c_{2,0}a^2 + c_{0,2}b^2 + 2c_{1,1}ab \\
&\quad + 2(c_{2,0}a_0 + c_{1,1}b_0 + c_{1,0})a + 2(c_{1,1}a_0 + c_{0,2}b_0 + c_{0,1})b \\
&= \mathcal{E}(a_0, b_0) + \mathrm{Var}(x)a^2 + (b + \bar{x}a)^2\,.
\end{aligned}$$

Les deux derniers termes sont positifs ou nuls, ce qui montre que $\mathcal{E}(a, b) \geqslant \mathcal{E}(a_0, b_0)$ pour tout (a, b). Pour avoir égalité, il faut que les deux termes soient nuls, ce qui se traduit par :

$$\sqrt{\mathrm{Var}(x)}a = 0 \qquad \text{et} \qquad \bar{x}a + b = 0\,.$$

C'est un système linéaire en (a, b), de déterminant $\sqrt{\mathrm{Var}(x)}$ qui est non-nul. Donc l'unique solution est $(a, b) = (0, 0)$. Ceci montre bien que (a_0, b_0) est un minimum global car $\mathcal{E}(a_0 + a, b_0 + b) > \mathcal{E}(a_0, b_0)$ dès que $(a, b) \neq (0, 0)$. ■

L'erreur quadratique minimale d'ajustement par une droite $y = ax + b$ obtenue avec les paramètres (a_0, b_0) obtenus ci-dessus est :

$$\begin{aligned}
\mathcal{E}(a_0, b_0) &= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \bar{y}) - a_0(x_i - \bar{x})\right)^2 \\
&= \mathrm{Var}(y) - 2a_0\mathrm{Cov}(x, y) + a_0^2\mathrm{Var}(x) = \frac{\mathrm{Var}(x)\mathrm{Var}(y) - \mathrm{Cov}(x, y)^2}{\mathrm{Var}(x)}\,. \qquad (8)
\end{aligned}$$

Qualitativement, plus cette erreur est grande, et moins bon est l'ajustement linéaire obtenu. Une analyse quantitative réclame une étude approfondie de la covariance de x et y.

3.2 Covariance et corrélation

Le paramètre d'ajustement linéaire principal de x et y est lié à la covariance de x et y.

Définition 4. La covariance de x et y est définie par :

$$\mathrm{Cov}(x,y) := \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\,. \tag{9}$$

La covariance joue un rôle analogue à la variance dans le cas de deux caractères : elle mesure la dispersion conjointe des deux caractères. Notez qu'en prenant $y = x$, on trouve que :

$$\mathrm{Cov}(x,x) = \mathrm{Var}(x)\,.$$

En supposant que $\mathrm{Var}(x) > 0$ et $\mathrm{Var}(y) > 0$, c'est-à-dire que $n \geqslant 2$ et les x_i (resp. les y_i) ne sont pas tous égaux, on peut définir la notion de corrélation.

Définition 5. Si $\mathrm{Var}(x) > 0$ et $\mathrm{Var}(y) > 0$, alors le coefficient de corrélation de x et y est défini par :

$$\mathrm{R}(x,y) := \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}\,. \tag{10}$$

Ce coefficient est en quelque sorte normalisé et ne dépend que des fluctuations relatives des caractères, et pas des fluctuations en valeur absolue. Pour tout $\lambda \in \mathbb{R}^*$, on a :

$$\mathrm{R}(\lambda x, \lambda y) = \mathrm{R}(x,y)\,.$$

Proposition 5. Le coefficient de corrélation a une valeur comprise entre -1 et 1.

Démonstration.

On considère la fonction f définie sur $\mathbb{R}$ et à valeurs dans $\mathbb{R}_+$ et donnée par :

$$f(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big((y_i-\bar{y}) - \lambda(x_i-\bar{x})\big)^2\,.$$

En développant les carrés, on obtient que f est un polynôme du second degré :

$$f(\lambda) = c_2\lambda^2 + 2c_1\lambda + c_0\,,$$

avec $c_2 = \mathrm{Var}(x)$, $c_1 = -\mathrm{Cov}(x,y)$, et $c_0 = \mathrm{Var}(y)$. Comme f garde un signe constant, son discriminant est négatif ou nul, ce qui veut dire que $c_1^2 - c_2c_0 \leqslant 0$, soit :

$$\mathrm{Cov}(x,y)^2 - \mathrm{Var}(x)\mathrm{Var}(y) \leqslant 0\,.$$

En divisant par $\mathrm{Var}(x)\mathrm{Var}(y)$:

$$\mathrm{R}(x,y)^2 = \frac{\mathrm{Cov}(x,y)^2}{\mathrm{Var}(x)\mathrm{Var}(y)} \leqslant 1\,.$$

■

FIGURE V.1.7 – Qualité des ajustements linéaires en fonction du coefficient de corrélation.

3.3 Corrélation et régression

Le coefficient de corrélation mesure la force et la direction de la relation entre x et y. Deux cas extrêmes peuvent être facilement analysés. Si $\mathrm{R}(x, y) = \pm 1$, alors il existe un $\lambda_0 \in \mathbb{R}$ tel que $\forall i \in [\![1, n]\!]$, $y_i - \bar{y} = \lambda_0(x_i - \bar{x})$. Cela montre que x et y sont des caractères parfaitement corrélés. Si $\mathrm{R}(x, y) = 0$, alors la meilleure droite d'ajustement linéaire est simplement la droite horizontale $y = \bar{y}$ ce qui tend à montrer que les deux caractères ne sont pas corrélés.

La figure V.1.7 donne plusieurs exemples pour différentes valeurs du coefficient de corrélation. Une valeur de $\mathrm{R}(x, y)$ proche de 1 indique que les caractères x et y sont positivement corrélés, et la meilleure droite d'ajustement linéaire obtenue par la méthode des moindres carrés a une pente positive. Une valeur de $\mathrm{R}(x, y)$ proche de -1 indique que les caractères x et y sont négativement corrélés, et la meilleure droite d'ajustement linéaire a une pente négative. Notez que le coefficient de corrélation mesure seulement la qualité d'une relation linéaire entre x et y. C'est visible sur la figure V.1.7(f) : les caractères ont l'air

corrélés, le nuage de points n'a pas l'air du tout aléatoire, mais le coefficient de corrélation est petit. En fait, dans ce cas, il faudrait mener une analyse de régression non linéaire, en essayant d'ajuster le nuage de points par une parabole par exemple, et non pas par une droite. C'est tout-à-fait possible, mais cela dépasse le cadre de ce cours.

L'erreur quadratique minimale d'ajustement par une droite $y = ax + b$ obtenue avec les paramètres (a_0, b_0) décrits dans la proposition 4 de la page 930 est donnée par (8). Cette quantité s'exprime aussi en termes du coefficient de corrélation comme :

$$\mathcal{E}(a_0, b_0) = \mathrm{Var}(y)\left[1 - \mathrm{R}(x, y)^2\right] .$$

Cette erreur quadratique doit être comparée à la variance de y. L'erreur quadratique $\mathcal{E}(a_0, b_0)$ mesure la part de la variance de y qui ne peut pas être expliquée et prédite par une fonction linéaire en x.

On voit donc que le coefficient de corrélation peut être utilisé pour mesurer la qualité d'une approximation de y par une fonction linéaire en x. Lorsque $\mathrm{R}(x, y)$ est en valeur absolue proche de 1 (en pratique strictement supérieur à $0{,}7$), la droite de régression linéaire $a_0x + b_0$ représente une bonne approximation de y.

Dans l'exemple 4, la variance de x (ancienneté) est $10{,}74$, la variance de y (salaire) est $1{,}22$ et la covariance de x et y vaut $2{,}57$. En conséquence, le coefficient de corrélation vaut $\mathrm{R}(x, y) \simeq 0{,}71$. La droite de régression linéaire obtenue par la méthode des moindres carrés et dessinée sur la figure V.1.6 est donc raisonnablement bonne, mais sans plus.

3.4 Régression linéaire facile avec des outils logiciels

Le logiciel Excel (ou ses versions libres) possède des outils de statistique élémentaire, et en particulier un module de régression linéaire. Les étapes sont simples et faciles à mettre en œuvre :

– Lancez Excel, ouvrez une feuille et saisissez dans deux colonnes les données d'une série statistique double.

– Sélectionnez les deux colonnes et tracez un nuage de points (pour cela, vous pouvez utiliser l'assistant graphique).

– Allez dans l'onglet « graphique » et sélectionnez « ajouter une courbe de tendance ». Dans le menu suivant sélectionner « régression linéaire » dans l'onglet Type, et cochez « afficher l'équation sur le graphique » et « afficher le coefficient de détermination sur le graphique » dans l'onglet Options. La droite de régression linéaire doit alors se superposer au nuage de points, ainsi que l'équation de cette droite $y = a_0x + b_0$ avec les valeurs numériques de a_0 et b_0, et le coefficient de corrélation au carré R^2. Beaucoup d'autres logiciels sont capables de réaliser le même travail.

Exercice type corrigé

V.1.0 Lors d'un contrôle noté sur 5, un groupe de $n_1 = 10$ étudiants a reçu les notes suivantes :

$$4,\ 1,\ 3,\ 3,\ 4,\ 2,\ 3,\ 5,\ 3,\ 4\,.$$

1. Calculer la moyenne m_1, le mode et la médiane des notes.
2. Un autre groupe de $n_2 = 4$ étudiants a reçu les notes 4, 4, 4 et 5.
   (a) Calculer la moyenne m_2, le mode et la médiane des notes de ce groupe de 4 étudiants.
   (b) Calculer la moyenne m_3, le mode et la médiane des notes du groupe global de 14 étudiants.
   (c) Vérifier la linéarité de la moyenne dans le sens où $m_3 = \dfrac{n_1 m_1 + n_2 m_2}{n_1 + n_2}$.
3. Le même groupe de 10 étudiants passe un nouveau contrôle où les étudiants reçoivent les notes (dans le même ordre que le premier contrôle) :

$$5,\ 2,\ 3,\ 4,\ 4,\ 3,\ 4,\ 4,\ 5,\ 4\,.$$

   (a) Calculer la moyenne m_4, le mode et la médiane des notes du second contrôle.
   (b) La note finale de chaque étudiant est la moyenne de ses deux notes aux deux contrôles (par exemple, la note finale de l'étudiant 1 est $\dfrac{4+5}{2} = 4{,}5$). Calculer les notes finales des 10 étudiants. Calculer la moyenne m_5, le mode et la médiane des notes finales du groupe de 10 étudiants.
   (c) Vérifier la linéarité de la moyenne, dans le sens où $m_5 = \dfrac{m_1 + m_4}{2}$.

Solution.

1. Le tableau des fréquences est

Note	Effectif (Nombre d'étudiants)	Fréquence (Proportion d'étudiants)
1	1	0,1
2	1	0,1
3	4	0,4
4	3	0,3
5	1	0,1
	10	1

La moyenne est :

$$m_1 = \frac{1}{10}\,(4+1+3+3+4+2+3+5+3+4) = 3{,}2\,.$$

Le mode est 3 (c'est la classe la plus importante).

La médiane est 3, car quand on met les notes dans l'ordre, les 5ème et 6ème notes sont toutes les deux égales à 3 :

$$1,\ 2,\ 3,\ 3,\ \underbrace{3,\ 3}_{},\ 4,\ 4,\ 4,\ 5\,.$$

2. (a) La moyenne est $m_2 = 4{,}25$, le mode est 4 et la médiane est 4.

   (b) La moyenne est $m_3 = 3{,}5$, le mode est 4 et la médiane est 4.

   (c) On vérifie ici la linéarité de la moyenne car :

$$m_3 = 3{,}5 = \frac{10 \times 3{,}2 + 4 \times 4{,}25}{10 + 4} = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}\,.$$

3. (a) La moyenne est $m_4 = 3{,}8$, le mode est 4 et la médiane est 4.

   (b) Les notes finales sont

$$4{,}5,\ 1{,}5,\ 3,\ 3{,}5,\ 4,\ 2{,}5,\ 3{,}5,\ 4{,}5,\ 4,\ 4$$

   La moyenne est $m_5 = 3{,}5$, le mode est 4 et la médiane est 4.

   (c) On vérifie ici la linéarité de la moyenne, car :

$$m_5 = 3{,}5 = \frac{3{,}2 + 3{,}8}{2} = \frac{m_1 + m_4}{2}\,.$$

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

V.1.1 On teste dix voitures en notant leur puissance (en chevaux) et leur consommation (en litres par 100 km). On obtient le tableau suivant :

Puissance	Consommation	Puissance	Consommation
85	6,0	90	7,0
110	5,9	115	9,0
75	5,7	90	4,5
145	11,2	85	4,4
85	5,1	70	5,2

1. Représenter graphiquement ces données par un nuage de points. Que peut-on déduire de ce graphique ?
2. Confirmer vos conclusions par un calcul de corrélation.
3. Tracer la droite de régression linéaire de la consommation en fonction de la puissance d'une voiture.
4. Si on veut acheter une voiture de 130 chevaux, quelle prévision de consommation peut-on faire ?

V.1.2 On considère le tableau (classement de la ligue 1 au 31 octobre 2017) :

Pos.	Equipe	Pts	G	N	P	Bp	Bc	Diff	J
1	Psg	29	9	2	0	34	8	26	11
2	Monaco	25	8	1	2	28	12	16	11
3	Lyon	22	6	4	1	27	15	12	11
4	Marseille	21	6	3	2	21	17	4	11
5	Nantes	20	6	2	3	9	8	1	11
6	St. Etienne	18	5	3	3	13	10	3	11
7	Caen	18	6	0	5	8	8	0	11
8	Bordeaux	16	4	4	3	16	17	-1	11
9	Montpellier	15	4	3	4	8	7	1	11
10	Toulouse Fc	15	4	3	4	11	15	-4	11
11	Guingamp	14	4	2	5	13	15	-2	11
12	Angers	13	2	7	2	16	14	2	11
13	Rennes	12	3	3	5	13	15	-2	11
14	Estac Troyes	12	3	3	5	9	15	-6	11
15	Dijon	12	3	3	5	14	20	-6	11
16	Sc Amiens	10	3	1	6	6	12	-6	10
17	Nice	10	3	1	7	13	19	-6	11
18	Strasbourg	10	2	4	5	13	20	-7	11
19	Lille	6	1	3	6	6	15	-9	10
20	Metz	3	1	0	10	5	21	-16	11

On considère la série double constituée du « nombre de buts encaissés » (colonne « Bc » du tableau) et du « nombre de parties gagnées » (colonne « G » du tableau).

1. Pour chaque série simple, calculer la moyenne et la variance.
2. Calculer la covariance et le coefficient de corrélation.
3. Quelle est l'équation de la droite d'ajustement linéaire d'explication du « nombre de parties gagnées » par le « nombre de buts encaissés » ? Tracer le nuage de points et la droite d'ajustement linéaire. L'ajustement est-il bon ?
4. Recommencer avec comme variable explicative la « différence de buts » (colonne « Diff » du tableau) au lieu du « nombre de buts encaissés ».

V.1.3 Une étude sur le budget consacré aux vacances d'été auprès de ménages a donné les résultats du tableau suivant :

Budget	Fréquence
$[800, 1000[$	0,08
$[1000, 1400[$	0,1
$[1400, 1600[$	0,16
$[1600, z[$	0,3
$[z, 2400[$	0,09
$[2400, 4000[$	0,27

1. Calculer la borne manquante z dans les deux cas suivants :
   - Le budget moyen est égal à $1\,995$ euros.
   - Le budget médian est égal à $1\,920$ euros.
2. Considérons dorénavant que la borne manquante est égale à $2\,000$ euros.
   Donner une représentation graphique (histogramme) de la distribution des budgets « vacances ».
3. Calculer et interpréter le budget moyen et médian. Que pouvez-vous conclure de la comparaison entre ces deux valeurs ?

V.1.4 Soit n un entier naturel et $(x_i)_{i\in[\![1,n]\!]}$ un n-uplet de réels. On souhaite trouver un réel x minimisant la somme des valeurs absolues des écarts ou la somme des écarts au carré. On définit donc sur $\mathbb{R}$ les deux fonctions G et L par :

$$G(x) = \sum_{i=1}^{n}(x - x_i)^2, \qquad L(x) = \sum_{i=1}^{n}|x - x_i|.$$

1. Minimisation de G.
   (a) Démontrer que la fonction G admet un minimum sur $\mathbb{R}$ et indiquer en quelle valeur de x il est atteint.
   (b) Que représente d'un point de vue statistique la valeur de x trouvée à la question précédente ?
2. Minimisation de L.
   On suppose désormais que la série est ordonnée, c'est-à-dire que $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$.
   (a) Représenter graphiquement la fonction L dans le cas où :
   $$n = 3 \qquad \text{et} \qquad x_1 = -2, \qquad x_2 = 3, \qquad x_3 = 4.$$
   (b) Représenter graphiquement la fonction L dans le cas où :
   $$n = 4 \qquad \text{et} \qquad x_1 = -2, \qquad x_2 = 3, \qquad x_3 = 4, \qquad x_4 = 7.$$
   (c) Démontrer que la fonction L admet un minimum sur $\mathbb{R}$ et indiquer pour quelle(s) valeur(s) de x il est atteint (on distinguera les cas n pair et n impair).
   (d) Que représentent, d'un point de vue statistique, les valeurs de x trouvées à la question précédente ?

V.1.5 Soit $(x_i)_{i\in[\![1,n]\!]}$ un échantillon constitué de n nombres réels (non nécessairement rangés par ordre croissant). On note $\bar{x}$ la moyenne arithmétique de la série et σ son écart-type.

1. Soit p le nombre d'éléments de l'échantillon compris entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$.
   Montrer que $\sum\limits_{k=1}^{n}(x_k - \bar{x})^2 \geqslant 4(n-p)\sigma^2$. En déduire qu'au moins les trois quarts des éléments de l'échantillon sont compris entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$.
2. Plus généralement, montrer que pour tout réel $t > 1$, l'intervalle $[\bar{x} - t\sigma, \bar{x} + t\sigma]$ contient au moins une proportion $1 - 1/t^2$ des éléments de l'échantillon.

V.1.6

On considère des données bivariées $(x, y) = (x_i, y_i)_{i \in [\![1,n]\!]}$ avec $\mathrm{Var}(x) \neq 0$ et $\mathrm{Var}(y) \neq 0$. On note D_1 la droite de régression de y par rapport à x et D_2 la droite de régression de x par rapport à y. Démontrer que $D_1 = D_2$ si, et seulement si, tous les points (x_i, y_i) sont alignés.

V.1.7 Soit $x = (x_i)_{i \in [\![1,n]\!]}$ un échantillon. On suppose ici que les observations $x_1, \ldots, x_n$ sont toutes des nombres réels positifs. On peut alors introduire la moyenne géométrique de ces observations, notée M et définie par :

$$M = \sqrt[n]{x_1 \cdots x_n}.$$

Exemple : Supposons que pendant une décennie, les salaires aient été multipliés par 2 et que pendant la décennie suivante ils aient été multipliés par 4. Pour la période de l'ensemble de ces deux décennies le coefficient multiplicateur est $2 \times 4 = 8$. Le coefficient multiplicateur moyen par décennie pour cette période de vingt ans est, par définition, le coefficient μ qui ne change pas d'une décennie à l'autre, et qui permet une multiplication par 8 des salaires entre le début et la fin de la période. On a donc $\mu^2 = 8$, d'où $\mu = \sqrt{2 \times 4} \simeq 2{,}83$. Ainsi, il apparaît que μ est la moyenne géométrique des deux coefficients multiplicateurs correspondant aux deux décennies ; aussi, il est important de noter que μ est différent (strictement plus petit) de la moyenne arithmétique de ces deux coefficients, en effet cette dernière moyenne vaut $(2 + 4)/2 = 3$.
On veut montrer que :

$$\min_{i \in [\![1,n]\!]} x_i \leqslant M \leqslant \bar{x} \leqslant \max_{i \in [\![1,n]\!]} x_i.$$

1. Montrer que $\min\limits_{i \in [\![1,n]\!]} x_i \leqslant M$ et $\bar{x} \leqslant \max\limits_{i \in [\![1,n]\!]} x_i$.
2. Montrer que $\sqrt{x_1 x_2} \leqslant \frac{x_1 + x_2}{2}$ pour tous $x_1, x_2 \geqslant 0$.
3. Montrer que $\sqrt[n]{x_1 \cdots x_n} \leqslant \frac{x_1 + \cdots + x_n}{n}$.

Probabilités finies V.2

1 Introduction aux probabilités

On peut rencontrer dans la vie courante, lors d'une expérience de physique ou en économie, des situations où le résultat n'est pas prévisible précisément. C'est le cas en particulier des jeux dits de hasard (dés, cartes, roulette, loto, ...). Le résultat d'une partie de cartes est impossible à prédire avec exactitude, car il dépend du tirage des cartes, i.e. des jeux de chacun des joueurs. Cependant, nous savons tous que la chance n'est pas tout, le talent des joueurs compte. Ceci montre qu'une analyse rigoureuse d'une situation où le hasard intervient est possible. Un bon joueur est capable d'évaluer ses chances de remporter une partie, c'est-à-dire d'évaluer la vraisemblance de cet événement. Le calcul des probabilités d'événements est l'objet de la seconde partie de ce module.

1.1 Expériences aléatoires, événements

Comme le montre la discussion précédente, la réalisation de l'événement considéré (remporter une partie de cartes) n'est pas certaine, elle dépend du résultat d'une expérience aléatoire (le tirage des cartes). Lors d'une expérience aléatoire, le résultat n'est pas prévisible exactement, mais l'ensemble des résultats possibles est connu.

1.1.1 Ensemble fondamental

Lors d'une expérience aléatoire, on commence par recenser l'ensemble de tous les résultats possibles de l'expérience. Cet ensemble non vide noté Ω est l'ensemble fondamental ou l'univers des possibles ; ses éléments ω sont appelés épreuves ou réalisations. L'ensemble fondamental Ω peut-être fini, dénombrable ou infini non dénombrable. On va se limiter ici à des ensembles finis, on étudiera les autres cas dans le cours L3.

Expérience 1. On lance une pièce et on regarde sur quel côté elle retombe. Deux résultats sont possibles, Pile et Face, de sorte que $\Omega = \{P, F\}$.

Expérience 2. On lance un dé et on lit le numéro de la face supérieure. L'ensemble fondamental est :

$$\Omega = [\![1, 6]\!] = \{1, 2, 3, 4, 5, 6\}.$$

Expérience 3. On lance deux fois un dé et on note les numéros obtenus. L'ensemble fondamental est :

$$\Omega = [\![1, 6]\!]^2 = \{(\omega_1, \omega_2) \mid \omega_1, \omega_2 \text{ entiers compris entre } 1 \text{ et } 6\}.$$

Il s'agit ici de couples ordonnés (de 2-uplets) parce qu'on distingue le premier lancer du second lancer du dé (ou, de façon équivalente, parce qu'on tire deux dés de couleurs différentes, qui permettent de les individualiser).

Expérience 4. On bat un jeu de 32 cartes numérotées de 1 à 32. On note $\omega(i)$ le numéro porté par la $i^{ème}$ carte du paquet après battage. ω est une permutation des 32 premiers entiers, c'est-à-dire une bijection de $[\![1,32]\!]$ dans lui-même. L'ensemble fondamental est donc :

$$\Omega = \{\omega \text{ permutation de } [\![1,32]\!]\} \ .$$

Expérience 5. Une usine produit $10\,000$ ampoules qui ont chacune leur durée de vie propre. L'ensemble fondamental est :

$$\Omega = [0,+\infty[^{10\,000} = \left\{(\omega_i)_{i\in[\![1,10\,000]\!]} \mid \omega_i \in [0,\infty[\right\} \ ,$$

où ω_i est la durée de vie de la $i^{ème}$ ampoule.

On va se limiter dans ce module à des ensembles finis, on étudiera le cas des ensembles infinis (expérience 5) dans le cours L3.

1.1.2 La notion d'événement

Un événement A est en général défini intuitivement comme une assertion qui peut être vraie ou fausse, un fait qui peut arriver ou non, selon le résultat de l'expérience.

Exemples. Dans l'expérience 2, on peut considérer les événements suivants :
a) A_1 est l'événement « le numéro obtenu est 6 »,
b) A_2 est l'événement « le numéro obtenu est pair »,
c) A_3 est l'événement « le numéro obtenu est supérieur ou égal à 5 ».
Dans l'expérience 3, on peut considérer l'événement B défini par « la somme des deux numéros obtenus est égale à 8 ».

Il existe deux événements particuliers importants. L'événement impossible I n'est jamais réalisé (dans l'expérience 2, on pourrait le définir par « le numéro obtenu est égal à 0 » ou de bien d'autres façons). L'événement certain C est toujours réalisé (dans l'expérience 2, on pourrait le définir par « le numéro obtenu est un entier compris entre 1 et 6 »).

Un événement A est ici défini par une assertion, mais il correspond aussi à un sous-ensemble, également noté A, de Ω. C'est l'ensemble des réalisations pour lesquelles l'événement est vrai.

Exemples. Pour l'expérience 2, on a $A_1 = \{6\}$, $A_2 = \{2,4,6\}$, $A_3 = \{5,6\}$, $I = \varnothing$, et $C = \Omega$.
Pour l'expérience 3, $B = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$.

Plus généralement, à chaque expérience aléatoire, chaque événement observable peut être décrit comme une partie de l'ensemble fondamental Ω. L'ensemble de tous les événements est $\mathcal{P}(\Omega)$; en particulier, l'événement impossible est $\varnothing$, l'événement certain est Ω.

1.1.3 Lien avec la théorie des ensembles

• L'événement contraire de A (celui qui se produit exactement quand A ne se produit pas) correspond au complémentaire A^c.

Exemples.
1) Le contraire de l'événement certain est l'événement impossible, $\Omega^c = \varnothing$.
2) Dans l'expérience 2, « le numéro obtenu est impair » correspond à l'ensemble $\{1, 3, 5\}$ qui est le complémentaire dans Ω de $A_2 = \{2, 4, 6\}$.

• Si A et B sont deux événements, l'événement « A et B se produisent tous les deux » correspond à $A \cap B$.

Exemple. Dans l'expérience 2, les événements A_2 et A_3 se produisent tous les deux si, et seulement si, le résultat est 6, c'est-à-dire si l'événement $A_2 \cap A_3 = \{6\} = A_1$ se réalise.

• Si A et B sont deux événements tels que $A \cap B = \varnothing$ (c'est-à-dire qu'ils ne se produisent jamais simultanément, comme par exemple dans le cas d'un événement et de son contraire), alors on dit que ces deux événements sont incompatibles (on dit aussi disjoints).

• Si A et B sont deux événements, l'événement « A ou B se produit (l'un ou l'autre, ou tous les deux) » correspond à $A \cup B$.

• Si A et B sont deux événements, on dira que A entraîne B si dans tous les cas où A se produit, alors B se produit, c'est-à-dire que $\omega \in A \implies \omega \in B$, c'est-à-dire $A \subset B$.

Exemple. Dans l'expérience 2, si le nombre obtenu est 6, alors il est pair : $A_1 \subset A_2$.

• Pour tout $\omega \in \Omega$, on appelle le singleton $\{\omega\}$ un événement élémentaire.

Exemple. Dans l'expérience 2, l'événement $A_1 = \{6\}$ est un événement élémentaire.

1.2 Espace de probabilité fini

1.2.1 Approche intuitive des probabilités

Nous voulons à présent quantifier la vraisemblance d'un événement. Supposons que l'on répète un grand nombre N de fois une expérience aléatoire (par exemple, le lancer de dé de l'expérience 2). Notons $N(A)$ le nombre de fois où un certain événement A s'est produit (par exemple, l'événement A_1 : le 6 est sorti). La fréquence empirique d'apparition de A, $N(A)/N$, est elle-même aléatoire et fluctue d'une suite de N répétitions à une autre suite. Mais notre expérience nous apprend que $N(A)/N$ se stabilise, quand le nombre N de répétitions tend vers l'infini, autour d'une valeur limite indépendante de la suite de répétitions. Les observations expérimentales suggèrent que c'est le cas pour beaucoup de situations, comme les jeux de dés. La limite de la fréquence empirique d'apparition de A est notre idée intuitive de la probabilité de l'événement A, soit $\mathbb{P}(A)$.

Exemple. Dans le cas de l'expérience 2, si le dé est bien équilibré, on peut penser qu'il y a une chance sur six de faire un six, c'est-à-dire que $\mathbb{P}(A_1) = 1/6$.

Clairement, si A est l'événement certain $A = \Omega$, alors $N(\Omega) = N$ et donc $\mathbb{P}(\Omega) = 1$. De plus, si les événements A et B sont incompatibles, c'est-à-dire si $A \cap B = \varnothing$, on a $N(A \cup B) = N(A) + N(B)$ et donc $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

1.2.2 Définition axiomatique

La théorie des probabilités est une théorie mathématique, fondée sur des bases axiomatiques. La définition axiomatique des probabilités reprend les propriétés intuitives évoquées ci-dessus, en les posant comme axiomes.

Définition 1. Soient Ω un ensemble fini non-vide et $\mathcal{P}(\Omega)$ l'ensemble des parties de Ω. On appelle probabilité finie sur $\big(\Omega, \mathcal{P}(\Omega)\big)$ une application $\mathbb{P}$ de $\mathcal{P}(\Omega)$ dans $[0,1]$ vérifiant :

1. $\mathbb{P}(\Omega) = 1$.
2. Pour tous $A, B \in \mathcal{P}(\Omega)$ tels que $A \cap B = \varnothing$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Le triplet $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ est appelé espace de probabilité fini ; pour tout événement A (c'est-à-dire pour tout sous-ensemble A de Ω), le réel positif ou nul $\mathbb{P}(A)$ est appelé probabilité de l'événement A.

Proposition 1 (Additivité finie). Soit p un entier supérieur ou égal à 2.
Si A_1, A_2, $\ldots$, A_p sont des événements deux à deux disjoints, c'est-à-dire qui vérifient $A_i \cap A_j = \varnothing$ pour tout $i \neq j$, alors :

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_p) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_p)\,. \tag{1}$$

Démonstration. On montre (1) par récurrence sur p.

- $p = 2$: ce cas est vérifié par définition d'une probabilité.
- Supposons le résultat vrai au rang $p-1$. Soient A_1, A_2, $\ldots$, A_p des événements deux à deux disjoints. On note $A = \bigcup\limits_{i=1}^{p-1} A_i$. On vérifie que A et A_p sont disjoints :

$$A \cap A_p = \bigcup_{i=1}^{p-1} \big(A_i \cap A_p\big) = \bigcup_{i=1}^{p-1} \varnothing = \varnothing\,.$$

Donc on peut appliquer le second axiome de la définition d'une probabilité :

$$\mathbb{P}\left(\bigcup_{i=1}^{p} A_i\right) = \mathbb{P}(A \cup A_p) = \mathbb{P}(A) + \mathbb{P}(A_p)\,. \tag{2}$$

En appliquant l'hypothèse de récurrence à la famille des $p-1$ événements A_1, A_2, $\ldots$, A_{p-1} deux à deux disjoints :

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i=1}^{p-1} A_i\right) = \sum_{i=1}^{p-1} \mathbb{P}(A_i)\,.$$

En injectant dans (2) on obtient (1). ∎

Proposition 2. Soient $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini, et $A, B \in \mathcal{P}(\Omega)$. On a alors :

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. Si $A \subset B$, alors $\mathbb{P}(A) \leqslant \mathbb{P}(B)$.
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Démonstration. 1. On a $\Omega = A^c \cup A$ et $A^c \cap A = \varnothing$, donc $1 = \mathbb{P}(\Omega) = \mathbb{P}(A^c) + \mathbb{P}(A)$, d'où $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

2. Soient A et B deux événements quelconques. Comme $B \subset \Omega$, on a $B = B \cap \Omega$. Or $\Omega = A \cup A^c$, donc par distributivité on obtient :

$$B = (B \cap A^c) \cup (B \cap A)\,. \tag{3}$$

Mais $(B \cap A^c) \cap (B \cap A) = B \cap (A \cap A^c) = \varnothing$, donc on peut appliquer le second axiome de la définition 1 de la page précédente :

$$\mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(B \cap A) \geqslant \mathbb{P}(B \cap A)\,, \tag{4}$$

puisque $\mathbb{P}(B \cap A^c) \geqslant 0$ (l'application $\mathbb{P}$ est à valeurs dans $[0, 1]$).

Si maintenant $A \subset B$, alors $A \cap B = A$ et (4) implique $\mathbb{P}(B) \geqslant \mathbb{P}(A)$.

3. D'après la décomposition (3) de B :

$$A \cup B = A \cup (B \cap A) \cup (B \cap A^c) = (A \cup (B \cap A)) \cup (B \cap A^c)\,.$$

Comme $B \cap A \subset A$, on a $A \cup (B \cap A) = A$ et donc $A \cup B = A \cup (B \cap A^c)$. Les deux ensembles du membre de droite étant disjoints, c'est-à-dire :

$$A \cap (B \cap A^c) = (A \cap A^c) \cap B = \varnothing\,,$$

on peut appliquer le second axiome de la définition 1 de la page ci-contre :

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)\,.$$

Il reste à calculer $\mathbb{P}(B \cap A^c)$. Or on déduit de la première égalité de (4) que :

$$\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(B \cap A)\,,$$

et finalement :

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(B \cap A)\,.$$

Le résultat n'est pas étonnant : on peut en effet décomposer l'événement $A \cup B$ en une union disjointe de trois événements $A \cap B^c$, $A \cap B$ et $A^c \cap B$ (voir Figure V.2.1), et donc $\mathbb{P}(A \cup B)$ est la somme des probabilités de ces trois événements d'après la proposition 1. ■

FIGURE V.2.1 – Décomposition de $A \cup B$ en une union disjointe de trois événements $A \cap B^c$, $A \cap B$ et $A^c \cap B$.

1.2.3 Caractérisation d'une probabilité

La proposition suivante affirme qu'une probabilité finie $\mathbb{P}$ sur $(\Omega, \mathcal{P}(\Omega))$ est entièrement caractérisée par les valeurs des probabilités des singletons $\mathbb{P}(\{\omega\})$, $\omega \in \Omega$.

Proposition 3. Soit Ω un ensemble fini non-vide comprenant n éléments. On numérote ses éléments $\Omega = \{\omega_i \mid i \in [\![1, n]\!]\}$.
(i) Toute suite de n réels positifs $(p_i)_{i \in [\![1,n]\!]}$, telle que $\sum_{i=1}^{n} p_i = 1$, définit une probabilité $\mathbb{P}$ sur $\big(\Omega, \mathcal{P}(\Omega)\big)$ par la formule :

$$\forall A \in \mathcal{P}(\Omega), \quad \mathbb{P}(A) := \sum_{i \,|\, \omega_i \in A} p_i \,. \tag{5}$$

(ii) Réciproquement, toute probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{P}(\Omega))$ est du type précédent, avec $p_i := \mathbb{P}(\{\omega_i\})$.

Démonstration. Montrons d'abord (i). Soit $\mathbb{P}$ définie par (5). Il faut montrer que $\mathbb{P}$ vérifie les axiomes de la définition 1 de la page 944 d'une probabilité. Le premier axiome est facile car $\mathbb{P}(\Omega) = \sum_i p_i = 1$. Vérifions maintenant le second axiome. Soient A et B deux événements disjoints. On a alors :

$$\mathbb{P}\left(A \cup B\right) = \sum_{\omega \in A \cup B} \mathbb{P}(\{\omega\}) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) + \sum_{\omega \in B} \mathbb{P}(\{\omega\}) = \mathbb{P}(A) + \mathbb{P}(B)\,.$$

Montrons la réciproque (ii). Soit $\mathbb{P}$ une probabilité sur $(\Omega, \mathcal{P}(\Omega))$. Soit $A \in \mathcal{P}(\Omega)$. Les événements $(\{\omega\})_{\omega \in A}$ sont disjoints deux à deux, donc la propriété d'additivité finie de $\mathbb{P}$ (proposition 1) implique que :

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) = \sum_{i \,|\, \omega_i \in A} \mathbb{P}(\{\omega_i\})\,. \qquad \blacksquare$$

1.2.4 Équiprobabilité

Il arrive que les différentes réalisations de l'ensemble fondamental aient toutes autant de chances d'être obtenues. On parle alors d'équiprobabilité.

Définition 2. On dit qu'il y a équiprobabilité lorsque les probabilités de tous les singletons (ou événements élémentaires) sont égales.

La propriété d'équiprobabilité est très forte, car il n'existe qu'une probabilité qui la satisfasse.

Proposition 4. Soit Ω un ensemble fini non-vide. Il existe une unique probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{P}(\Omega))$ qui satisfasse la propriété d'équiprobabilité. On l'appelle probabilité uniforme. Pour tout événement A, on a :

$$\mathbb{P}(A) = \frac{\operatorname{card}(A)}{\operatorname{card}(\Omega)} = \frac{\text{« nombre de cas favorables »}}{\text{« nombre de cas possibles »}}\,. \tag{6}$$

Démonstration. On note $n = \mathrm{card}\,(\Omega)$. On numérote les éléments de Ω :

$$\Omega = \{\omega_1, \ldots, \omega_n\} = \{\omega_1\} \cup \cdots \cup \{\omega_n\}\,.$$

i) On vérifie immédiatement que (6) est bien une probabilité, c'est-à-dire qu'elle satisfait les deux axiomes de la définition 1 de la page 944. De plus, pour tout $i \in [\![1, n]\!]$, on a $\mathbb{P}(\{\omega_i\}) = 1/n$, donc cette probabilité satisfait la propriété d'équiprobabilité.

ii) Réciproquement, supposons que $\mathbb{P}$ soit une probabilité qui vérifie la propriété d'équiprobabilité. On a alors :

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\{\omega_1\}) + \cdots + \mathbb{P}(\{\omega_n\}) = n\mathbb{P}(\{\omega_i\})\,,$$

pour n'importe quel indice i variant entre 1 et n, donc $\mathbb{P}(\{\omega_i\}) = 1/n$ pour tout i. D'après la proposition 3 de la page précédente, la probabilité $\mathbb{P}$ est entièrement caractérisée. ■

Exemples.

1. Dans l'expérience 2 de la page 941 : on suppose le dé équilibré, et donc les événements élémentaires sont équiprobables. Pour tout $i \in \Omega = [\![1, 6]\!]$, nous avons $\mathbb{P}(\{i\}) = 1/6$. Si nous considérons l'événement $A_2 = \{2, 4, 6\}$, alors :

$$\mathbb{P}(A_2) = \frac{\mathrm{card}\,(A_2)}{\mathrm{card}\,(\Omega)} = \frac{3}{6} = \frac{1}{2}\cdot$$

2. Expérience du tirage d'une chaussette dans un tiroir contenant 6 chaussettes noires et 4 chaussettes blanches ; on a deux manières de voir les choses :

(i) On peut prendre $\Omega = \{\mathrm{noir}, \mathrm{blanc}\}$, qui est un ensemble à deux éléments. Dans ce cas il est clair que les singletons ne sont pas équiprobables. On a en fait $\mathbb{P}(\{\mathrm{noir}\}) = 6/10 = 3/5$ et $\mathbb{P}(\{\mathrm{blanc}\}) = 4/10 = 2/5$.

(ii) On peut aussi choisir de définir Ω comme l'ensemble des tirages de chacune des 10 chaussettes : $\Omega = \{n_1, \ldots, n_6, b_1, \ldots, b_4\}$. On a alors équiprobabilité, et pour tout $i \in [\![1, 6]\!]$ et tout $j \in [\![1, 4]\!]$, on a $\mathbb{P}(\{n_i\}) = \mathbb{P}(\{b_j\}) = 1/10$. Si on note A l'événement « on tire une chaussette blanche », $A = \{b_1, \ldots, b_4\}$, alors on a :

$$\mathbb{P}(A) = \frac{\mathrm{card}\,(A)}{\mathrm{card}\,(\Omega)} = \frac{4}{10} = \frac{2}{5}\cdot$$

1.3 Probabilités de réunions d'ensembles : règle d'inclusion-exclusion

On sait calculer la probabilité d'une union d'événements deux à deux disjoints, c'est simplement la somme des probabilités des événements. On s'intéresse ici au calcul de la probabilité d'une union d'événements quelconques. La règle d'inclusion-exclusion s'applique à une suite $(A_i)_{i \in [\![1,n]\!]}$ d'événements pour lesquels on connaît a priori les probabilités des conjonctions d'événements. On peut alors calculer la probabilité de $A = A_1 \cup A_2 \cup \cdots \cup A_n$, i.e. la probabilité que « au moins l'un des A_i est réalisé ».

Proposition 5 (Formule d'inclusion-exclusion). Soient $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini et $(A_i)_{i\in[\![1,n]\!]}$ une suite d'événements. La formule d'inclusion-exclusion s'écrit :
1) Dans le cas de deux événements :

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)\,.$$

2) Dans le cas de trois événements :

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3) = \mathbb{P}(A_1) &+ \mathbb{P}(A_2) + \mathbb{P}(A_3)\\ &- \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3)\\ &+ \mathbb{P}(A_1 \cap A_2 \cap A_3)\,.\end{aligned}$$

3) Dans le cas général de n événements :

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right)\,. \tag{7}$$

La relation (7) s'appelle formule d'inclusion-exclusion car c'est exactement ce qu'on fait. Pour calculer $\mathbb{P}(\cup_i A_i)$, on commence par calculer la somme $\sum\limits_i \mathbb{P}(A_i)$. Mais on sur-estime ainsi le résultat, car on a compté plusieurs fois les intersections. On enlève alors la somme des probabilités des intersections d'ordre deux : $\sum\limits_{i<j} \mathbb{P}(A_i \cap A_j)$. Cette fois-ci, on sous-estime le résultat, car on a enlevé plusieurs fois les intersections d'ordre trois. Il faut alors les remettre, etc.

Démonstration. Le 1) a déjà été montré dans la proposition 2. On va montrer le cas général 3) par récurrence sur n. Supposons la récurrence montrée au rang n. Soit une suite d'événements $(A_i)_{i\in[\![1,n+1]\!]}$. En appliquant le 1) au couple $\bigcup\limits_{i=1}^{n} A_i$ et A_{n+1}, on obtient :

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathbb{P}\left(A_{n+1}\right) - \mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \bigcap A_{n+1}\right)\,.$$

En appliquant l'hypothèse de récurrence au premier terme du membre de droite :

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right)\,.$$

En appliquant l'hypothèse de récurrence au troisième terme du membre de droite :

$$\begin{aligned}\mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \bigcap A_{n+1}\right) &= \mathbb{P}\left(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right) \qquad \text{par distributivité}\\ &= \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k} \cap A_{n+1}\right)\,.\end{aligned}$$

En regroupant :

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k}\right)$$
$$+\sum_{k=0}^{n}(-1)^{k} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k}\cap A_{n+1}\right) . \quad (8)$$

Or en cassant le membre de droite de l'Eq. (7) en deux, avec d'un côté les ensembles d'indices qui contiennent $n+1$, et de l'autre côté ceux qui ne le contiennent pas, on obtient :

$$\sum_{k=1}^{n+1}(-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n+1} \mathbb{P}\left(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k}\right)$$
$$=\sum_{k=1}^{n+1}(-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k}\right)$$
$$+\sum_{k=1}^{n+1}(-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_{k-1}\leqslant n} \mathbb{P}\left(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_{k-1}}\cap A_{n+1}\right) .$$

Dans le premier terme du membre de droite la somme s'arrête en fait à $k=n$, car il y a au plus n entiers distincts entre 1 et n. Dans le second terme, on opère le changement de variables $k'=k-1$. On retrouve alors l'expression (8).
On verra dans le module V.3, page 978, une démonstration plus élégante et directe de cette proposition. ■

2 Combinatoire

On vient de voir (proposition 4) que le calcul des probabilités sur un espace de probabilité fini avec une probabilité uniforme se ramenait en fait à du dénombrement. Pour calculer la probabilité d'un certain événement A, il faut compter le nombre de cas favorables pour lesquels A se réalise, et diviser par le nombre total de cas possibles. On va dans cette partie donner quelques outils fondamentaux de la théorie du dénombrement ou combinatoire. Cette section reprend une partie des résultats introduits dans le module II.1 et elle donne des exemples adaptés à notre problématique. Le problème est plus simple ici que dans le module II.1, car on ne considère que des ensembles finis.

2.1 Généralités sur le dénombrement

La notion de cardinal a été introduite dans la définition 12 de la page 70 du module II.1, qu'on peut résumer ainsi :

Définition 3. Soit E un ensemble fini. Le cardinal de E, noté $\operatorname{card}(E)$, est le nombre d'éléments de E.

Dénombrer un ensemble E consiste à déterminer le cardinal de E. Certains ensembles sont faciles à dénombrer, mais d'autres le sont moins. Il est donc nécessaire de donner des

outils qui permettent de dénombrer systématiquement les éléments de E, sans en oublier, ni compter un élément plusieurs fois.
La proposition suivante est une reformulation des théorèmes 13-14 du module II.1. Ses assertions sont complètement intuitives et triviales, mais leur importance tient au fait qu'on les applique très fréquemment.

Proposition 6.

1. Soient E et F des ensembles finis tels qu'il existe une injection de E dans F. Alors $\operatorname{card}(E) \leqslant \operatorname{card}(F)$.
2. Soient E et F des ensembles finis tels qu'il existe une surjection de E dans F. Alors $\operatorname{card}(E) \geqslant \operatorname{card}(F)$.
3. Soient E et F des ensembles finis tels qu'il existe une bijection de E dans F. Alors $\operatorname{card}(E) = \operatorname{card}(F)$.
4. Soient E un ensemble fini et A une partie de E. Alors $\operatorname{card}(A) \leqslant \operatorname{card}(E)$.

Il arrive fréquemment en probabilités, lorsqu'on répète une certaine expérience plusieurs fois (lancer de dé, tirage avec remise, ...), que l'ensemble fondamental soit un ensemble produit. Le cardinal d'un tel ensemble est facile à calculer, comme le montre la proposition suivante.

Proposition 7. Si E est un ensemble fini de cardinal n et si on note E^p l'ensemble des p-uplets d'éléments de E :

$$E^p := \{(x_1, \ldots, x_p) \mid x_i \in E, i \in [\![1, p]\!]\} ,$$

alors $\operatorname{card}(E^p) = n^p$.

Démonstration. On peut démontrer ce résultat en appliquant le théorème 18 de la page 73 du module II.1. On peut aussi donner une preuve élémentaire. Pour construire un p-uplet, on a n choix pour x_1. Puisqu'on admet les répétitions, pour chaque choix de x_1, on a encore n choix pour x_2, donc au total $n \times n$ choix pour (x_1, x_2). Pour chacun de ces choix, on a à nouveau n choix pour x_3, et ainsi de suite : on a au total $n \times n \times \cdots \times n = n^p$ choix possibles. ∎

Exemple. Dans l'expérience 3 de la page 941 (double lancer d'un dé) : l'ensemble fondamental est $\Omega = E^p$ avec $E = [\![1, 6]\!]$ et $p = 2$. Le cardinal de l'ensemble fondamental est $\operatorname{card}(\Omega) = 6^2 = 36$. Les résultats des lancers étant équiprobables, la probabilité $\mathbb{P}$ est uniforme, c'est-à-dire que, pour tout couple $(i, j) \in \Omega$, on a $\mathbb{P}(\{(i, j)\}) = 1/36$. Pour l'événement $B =$« la somme des lancers est 8 », on a $\operatorname{card}(B) = 5$.
Par conséquent $\mathbb{P}(B) = \operatorname{card}(B)/\operatorname{card}(\Omega) = 5/36$.

La proposition suivante reprend des résultats extraits des théorèmes 24-26 du module II.1 et qui sont utiles en probabilités.

Proposition 8.
Soient E et F des ensembles finis, $n := \operatorname{card}(E)$ et $p := \operatorname{card}(F)$.

1. L'ensemble $\mathcal{F}(E, F)$ des applications de E dans F est de cardinal p^n.
2. L'ensemble $\mathcal{P}(E)$ des parties de E est de cardinal 2^n.

Démonstration. On donne juste des preuves élémentaires de ces résultats déjà démontrés dans le module II.1.
1. Une application de E dans F est spécifiée par la donnée de l'image de chaque élément de E. Or un élément de E a exactement p images possibles. Il y a n éléments dans E, donc le nombre total de choix est p^n.
2. On numérote les éléments de E si bien que $E = \{x_1, \ldots, x_n\}$. On considère l'application :

$$\begin{aligned} \Psi : \ & \{0,1\}^n \to \mathcal{P}(E) \\ & (y_i)_{i \in [\![1,n]\!]} \mapsto A = \bigcup_{i \,|\, y_i = 1} \{x_i\} \end{aligned}$$

L'application Ψ est une bijection, sa bijection réciproque est :

$$\begin{aligned} \Psi^{-1} : \ & \mathcal{P}(E) \to \{0,1\}^n \\ & A \mapsto (y_i)_{i \in [\![1,n]\!]} \text{ avec } y_i = 1 \text{ si } x_i \in A,\ y_i = 0 \text{ si } x_i \notin A \end{aligned}$$

Par la proposition 6 de la page ci-contre, on a $\operatorname{card}[\mathcal{P}(E)] = \operatorname{card}[\{0,1\}^n]$. Par la proposition 7 de la page précédente, $\operatorname{card}[\{0,1\}^n] = 2^n$, ce qui établit le résultat désiré. ∎

2.2 Dénombrements classiques

Proposition 9. Soient E un ensemble fini de cardinal n, et p un entier $\leqslant n$. On appelle arrangement à p éléments de E une liste ordonnée $(x_1, \ldots, x_p)$ d'éléments de E tous distincts. Le nombre d'arrangements à p éléments de E est noté A_n^p et vaut :

$$A_n^p = \frac{n!}{(n-p)!} = n(n-1)(n-2)\cdots(n-p+1)\,.$$

La seconde formule $A_n^p = n(n-1)(n-2)\cdots(n-p+1)$ peut être utile en particulier lorsqu'on cherche à évaluer numériquement la valeur de A_n^p avec un n grand et un p petit. En effet, dans ce cas l'évaluation de $n!$ et de $(n-p)!$ conduit à des nombres très grands, et la précision numérique du rapport peut être mauvaise.

Démonstration. Ce résultat est une reformulation du théorème 27 de la page 77 du module II.1. On en donne ici une preuve élémentaire. On a n choix pour x_1. Pour chaque choix de x_1, on n'a plus que $n-1$ choix pour x_2, puis $n-2$ choix pour x_3 une fois que x_1 et x_2 ont été choisis, et ainsi de suite jusqu'à x_p pour lequel on a $n-(p-1) = n-p+1$ choix. On a donc $n \times (n-1)$ choix pour (x_1, x_2), $n \times (n-1) \times (n-2)$ choix pour (x_1, x_2, x_3), et finalement $A_n^p = n(n-1)(n-2)\cdots(n-p+1)$ choix pour l'arrangement. Et comme $n! = n \times (n-1) \times (n-2) \times \cdots \times (n-p+1) \times (n-p) \times \cdots \times 2 \times 1$, et $(n-p)! = (n-p) \times \cdots \times 2 \times 1$, on obtient finalement que $A_n^p = n!/(n-p)!$. ∎

Remarques.

1. Comme on l'a vu dans le théorème 27 de la page 77 du module II.1, si E et F sont des ensembles finis, de cardinal $\operatorname{card}(E) = p$ et $\operatorname{card}(F) = n$, alors le nombre d'injections de E dans F est égal à A_n^p si $p \leqslant n$ et 0 si $p > n$.

2. Le nombre d'arrangements de n éléments d'un ensemble de cardinal n est $A_n^n = n!$. On appelle un tel arrangement une permutation. Le nombre $n!$ est donc le nombre de manières d'attribuer des numéros aux éléments de E. C'est le nombre de bijections entre l'ensemble $[\![1, n]\!]$ et E, ou, plus généralement, le nombre de bijections entre deux ensembles E et F finis de même cardinal n.

3. L'ordre est important pour les arrangements. Par exemple, si on considère l'ensemble $E = [\![1, 6]\!]$ et $p = 2$, alors les deux arrangements à deux éléments $(1, 2)$ et $(2, 1)$ sont différents et doivent être comptés pour 1 chacun.

Proposition 10. Soient E un ensemble fini de cardinal n, et p un entier $\leqslant n$. Le nombre de sous-ensembles de E qui ont p éléments, c'est-à-dire des parties de E de cardinal p, est noté $\binom{n}{p}$ ou C_n^p. On parle aussi du nombre de combinaisons à p éléments dans un ensemble à n éléments. Alors :

$$\binom{n}{p} = \frac{A_n^p}{p!} = \frac{n!}{p!(n-p)!} = \frac{n(n-1)(n-2)\cdots(n-p+1)}{p!}.$$

Remarque. En tenant compte du fait que $0! = 1$, on a $\binom{n}{n} = n!/(0!n!) = 1$, le seul sous-ensemble de E de cardinal égal à celui de E est E lui-même. De même $\binom{n}{0} = 1$ car le seul ensemble de cardinal 0 inclus dans E est l'ensemble vide.

Démonstration. Pour les sous-ensembles, l'ordre ne compte pas. Si l'on reprend l'exemple avec $E = [\![1, 6]\!]$ et $p = 2$, les notations $\{1, 2\}$ et $\{2, 1\}$ représentent le même sous-ensemble. Donc il y a plusieurs arrangements possibles correspondant à un même sous-ensemble. Plus précisément, étant donné un sous-ensemble $\{x_1, \ldots, x_p\}$, tous les arrangements mettant en jeu ces mêmes p éléments s'obtiennent en changeant l'ordre, en renumérotant l'ensemble $\{x_1, \ldots, x_p\}$. Nous venons de voir qu'il y avait $p!$ manières de faire cela. Donc à chaque combinaison (sous-ensemble) de p éléments de E correspondent $p!$ arrangements, et le nombre de sous-ensembles est celui d'arrangements, divisé par $p!$. Cela prouve au passage que la formule de $\binom{n}{p}$ donne toujours un nombre entier, ce qui n'était pas évident. ■

Exemple. Si l'on interprète l'expérience 3 de la page 941 avec des dés indistinguables (on les lance en même temps et ils ont la même couleur), alors le résultat où on obtient un 2 et un 6 est noté $(2, 6)$ par exemple, et l'ensemble fondamental est :

$$\begin{aligned}\Omega' &= \{(i, j) \mid i, j \in [\![1, 6]\!], i \leqslant j\} \\ &= \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6), (1,2), (1,3), (1,4), (1,5), (1,6)\} \\ &\qquad \cup \{(2,3), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6), (4,5), (4,6), (5,6)\}\ .\end{aligned}$$

L'ensemble Ω' contient 6 paires et toutes les combinaisons à deux éléments pris dans $[\![1, 6]\!]$, donc $\operatorname{card}(\Omega') = 6 + \binom{6}{2} = 21$. Dans ce modèle, l'événement $B =$« la somme des dés est 8 » a pour éléments $(2, 6)$, $(3, 5)$ et $(4, 4)$.

Supposons qu'on adopte un modèle de probabilité uniforme sur Ω'. La probabilité de B serait alors $3/21 = 1/7 > 5/36$. Il y a clairement un problème. En effet, les résultats $(4,4)$ et $(3,5)$ ne sont pas équiprobables, car, pour sortir un double 4, il faut que les deux dés soient égaux à 4, alors que pour sortir un 3 et un 5, il faut qu'un des dés donne 3 et l'autre donne 5, donc on a deux « possibilités ». On peut obtenir un modèle valide en ne considérant pas les singletons de notre Ω' comme équiprobables, mais que $\mathbb{P}(\{(i,j)\}) = 2/36 = 1/18$ si $i < j$, $\mathbb{P}(\{(i,j)\}) = 1/36$ si $i = j$. On obtient alors les mêmes probabilités que le modèle précédent pour les événements qui peuvent être définis dans les deux modèles (c'est-à-dire ceux qui ne dépendent pas de l'ordre des dés). Ainsi :

$$\mathbb{P}(B) = \mathbb{P}(\{(2,6)\}) + \mathbb{P}(\{(3,5)\}) + \mathbb{P}(\{(4,4)\}) = \frac{1}{18} + \frac{1}{18} + \frac{1}{36} = \frac{5}{36}\cdot$$

2.3 Dénombrement appliqué au loto

On considère tout d'abord la forme classique du loto, qui a été utilisée de 1976 à 2008. Un joueur choisit 6 numéros parmi 49 (sans répétition). Les organisateurs du jeu tirent eux aussi de leur côté 6 numéros. Si les deux choix sont identiques, le joueur est gagnant. Il peut y avoir des « petits » gagnants (lorsque le joueur trouve 3, 4 ou 5 numéros sur les six). La formule comporte aussi le tirage d'un numéro complémentaire en plus des six numéros. Un joueur qui trouve cinq bons numéros plus le numéro complémentaire gagne un lot de consolation.

Le calcul de la probabilité de gain par un joueur n'est pas compliqué. Il y a $\binom{49}{6}$ possibilités de choisir six numéros distincts parmi 49, et toutes les combinaisons sont équiprobables. Donc la probabilité qu'un joueur trouve les six bons numéros est :

$$p_6 = \frac{1}{\binom{49}{6}} = \frac{1}{13\,983\,816} \simeq 0{,}00000715\ \%\,.$$

La probabilité qu'il en trouve 5 est :

$$p_5 = \frac{\binom{43}{1}\binom{6}{5}}{\binom{49}{6}} = \frac{43}{2\,330\,636} \simeq 0{,}00184\ \%\,,$$

car il y a $\binom{6}{5}$ choix possibles de cinq bons numéros, et $\binom{43}{1}$ choix possibles d'un mauvais numéro. De même, les probabilités d'avoir quatre ou trois bons numéros sont :

$$p_4 = \frac{\binom{43}{2}\binom{6}{4}}{\binom{49}{6}} = \frac{645}{665\,896} \simeq 0{,}0969\ \%\,,$$

$$p_3 = \frac{\binom{43}{3}\binom{6}{3}}{\binom{49}{6}} = \frac{8\,815}{499\,422} \simeq 1{,}765\ \%\,.$$

Enfin, la probabilité d'avoir cinq bons numéros plus le complémentaire est :

$$p_c = \frac{\binom{6}{5}}{\binom{49}{6}} = \frac{1}{2\,330\,636} \simeq 0{,}0000429\ \%\,.$$

L'intuition peut être mise en défaut dans ce genre de situations. Par exemple, il paraît intuitivement peu probable que la « bonne » combinaison de six numéros comporte deux numéros qui se suivent. Et pourtant, l'exercice qui suit montre que cette probabilité est exactement $1 - \binom{44}{6}/\binom{49}{6}$, c'est-à-dire à peu près $50\ \%$.

La forme moderne du loto, mis en place à partir de 2008, est légèrement différente : elle comporte le tirage de 5 boules parmi 49, et le tirage d'un « numéro chance » parmi 10. Donc la probabilité qu'un joueur trouve la bonne combinaison est :

$$p = \frac{1}{\binom{49}{5}} \frac{1}{10} = \frac{1}{19\ 068\,840} \simeq 0{,}00000524\ \%\,,$$

car il y a $\binom{49}{5}$ choix possibles des cinq premiers numéros, et 10 choix possibles du « numéro chance ». Notez qu'on a moins de chance de gagner le gros lot avec la nouvelle version qu'avec la version classique.

Exercice 1.

On choisit au hasard (i.e. de manière équiprobable) r nombres distincts parmi les n premiers entiers strictement positifs. Calculer la probabilité que cette suite de r nombres ne possède pas de numéros qui se suivent.

Solution. Il faut dénombrer le nombre de choix possibles de telles suites. Cela revient à dénombrer le nombre de « manières » de choisir r objets parmi n objets alignés, en laissant au moins un objet entre deux objets sélectionnés. Or l'ensemble E de ces « manières » est en bijection avec l'ensemble F des « manières » de choisir r objets parmi $n-(r-1)$ objets alignés sans prendre garde aux espacements. En effet, cette bijection se construit en effaçant les $r-1$ objets qui suivent les $r-1$ premiers objets sélectionnés. La fonction réciproque se construit en insérant $r-1$ objets après chacun des $r-1$ premiers objets sélectionnés. Dans le cas $n=8$ et $r=3$, cela donne :

$$\begin{array}{ccc} E & \to & F \\ \times\times\otimes\times\otimes\times\times\otimes & \mapsto & \times\times\otimes\otimes\times\otimes \\ \otimes\times\otimes\times\times\otimes\times\times & \mapsto & \otimes\otimes\times\otimes\times\times \end{array}$$

où on a juste représenté la transformation de deux objets particuliers. On a donc :

$$\operatorname{card}(E) = \operatorname{card}(F) = \binom{n-r+1}{r},$$

et la probabilité cherchée est :

$$p = \frac{\binom{n-r+1}{r}}{\binom{n}{r}}\,.$$

3 Conditionnement et indépendance

3.1 Probabilité conditionnelle

3.1.1 Approche intuitive des probabilités conditionnelles

La notion de probabilité conditionnelle intervient chaque fois que l'on dispose d'une information supplémentaire qui restreint le champ du possible, et qui permet (par exemple) de réviser ses pronostics sur la probabilité de réalisation d'un événement. Considérons l'exemple du loto. On assiste en direct au tirage, et l'on voit la première boule sortir : le 13. La probabilité de toutes les combinaisons de six numéros ne contenant pas le 13 devient instantanément nulle, alors que les combinaisons de numéros qui contiennent le 13 voient leur probabilité augmenter.

Prenons un exemple un peu moins calculatoire, l'expérience 3 de la page 941 : on lance deux fois un dé, et on considère l'événement B = « le résultat du second lancer est plus grand (strictement) que le résultat du premier ». La probabilité de cet événement est $\mathbb{P}(B) = 15/36 = 5/12$, car il y a 15 combinaisons sur 36 pour lesquelles cet événement est réalisé. Supposons qu'on observe le résultat du premier lancer. Si l'on obtient un 6, alors la probabilité d'obtenir un plus grand chiffre au second lancer est clairement nulle. Si on obtient un 5 (événement noté A), alors la probabilité d'obtenir un plus grand chiffre au second lancer n'est pas nulle, mais on voit bien qu'elle est réduite par rapport à ce qu'elle était sans cette information supplémentaire. Pour que B se réalise, il faut maintenant que le résultat du second dé soit égal à 6, donc la probabilité de l'événement est maintenant $1/6$. On dit que la probabilité de B sachant A est $1/6$, et on note $\mathbb{P}(B|A) = 1/6$. Notez que $A \cap B$ se réduit au singleton $\{(5,6)\}$, donc $\mathbb{P}(A \cap B) = 1/36$, et $\mathbb{P}(A) = 1/6$, si bien que l'égalité $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$ est satisfaite.

On peut deviner que cette égalité devrait être largement vérifiée, en raisonnant avec notre représentation intuitive d'une probabilité comme une fréquence empirique (voir le paragraphe 1.2.1). Considérons une certaine expérience aléatoire (par exemple, le lancer de deux dés). On cherche à évaluer la probabilité qu'un certain événement B se réalise (par exemple, le résultat du second dé est plus grand que le premier) sachant qu'un événement A s'est réalisé (par exemple, la premier dé donne un 5). Pour avoir une estimée intuitive de cette probabilité conditionnelle, on va mener N répétitions de l'expérience aléatoire dans des « conditions identiques et indépendantes », puis on ne va considérer que les $N(A)$ répétitions de l'expérience qui ont réalisé A. Nous pouvons penser que la fréquence $N(A \cap B)/N(A)$ de réalisation de l'événement B parmi les $N(A)$ répétitions de l'expérience qui ont réalisé A fournit une estimation de la probabilité conditionnelle de B sachant A. Puisque $N(A \cap B)/N(A)$ est le quotient de $N(A \cap B)/N$ par $N(A)/N$, la relation suivante paraît naturelle :

$$\mathbb{P}(B|A) \simeq \frac{N(A \cap B)}{N(A)} = \frac{N(A \cap B)}{N} \times \frac{N}{N(A)} \simeq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \cdot$$

3.1.2 Définition axiomatique

Les considérations précédentes motivent la définition suivante.

Définition 4. Soient $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini, et A un événement de probabilité non nulle. Pour tout événement $B \in \mathcal{P}(\Omega)$, on appelle probabilité conditionnelle de B sachant A, le nombre :

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \cdot$$

Notez bien qu'il s'agit d'une définition axiomatique. Le raisonnement avec les mains fait dans le paragraphe précédent ne nous sert qu'à justifier l'intérêt pratique de la définition et son lien avec l'intuition.

On a ainsi défini une nouvelle application $B \mapsto \mathbb{P}(B|A)$ qui va de $\mathcal{P}(\Omega)$ dans $\mathbb{R}$. Comme le montre la proposition suivante, cette application est une probabilité.

Proposition 11. Soient $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini, et A un événement de probabilité non nulle. Le triplet $\big(\Omega, \mathcal{P}(\Omega), \mathbb{P}(\cdot|A)\big)$ définit un espace de probabilité fini.

Démonstration. Nous devons vérifier les deux axiomes d'une probabilité. Le premier axiome de la définition 1 de la page 944 explique pourquoi nous avons divisé par $\mathbb{P}(A)$. En effet :

$$\mathbb{P}(\Omega|A) = \frac{\mathbb{P}(A \cap \Omega)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A)} = 1\,.$$

Si maintenant nous avons deux événements $B_1, B_2 \in \mathcal{P}(\Omega)$ tels que $B_1 \cap B_2 = \varnothing$, alors $(A \cap B_1) \cap (A \cap B_2) = A \cap B_1 \cap B_2 = \varnothing$, donc :

$$\begin{aligned}
\mathbb{P}(B_1 \cup B_2|A) &= \frac{\mathbb{P}(A \cap (B_1 \cup B_2))}{\mathbb{P}(A)} = \frac{\mathbb{P}((A \cap B_1) \cup (A \cap B_2))}{\mathbb{P}(A)} \\
&= \frac{\mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B_1)}{\mathbb{P}(A)} + \frac{\mathbb{P}(A \cap B_2)}{\mathbb{P}(A)} \\
&= \mathbb{P}(B_1|A) + \mathbb{P}(B_2|A)\,,
\end{aligned}$$

ce qui montre que $\mathbb{P}(\cdot|A)$ vérifie le second axiome de la définition 1 de la page 944. ■

3.2 Probabilités composées, probabilités totales

3.2.1 Probabilités composées

En multipliant la formule définissant la probabilité conditionnelle $\mathbb{P}(B|A)$ par $\mathbb{P}(A)$ (donnée par la définition 4 de la page précédente), on obtient la formule des probabilités composées :

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B|A)\,, \tag{9}$$

valable pour tout couple d'événements A et B, pourvu que $\mathbb{P}(A) > 0$. C'est une formule très intuitive : pour que A et B se produisent, il faut que A se produise, et, sachant que A se produit, il faut que B se produise.

Exercice 2.

Soit un tiroir contenant 6 chaussettes noires et 4 chaussettes blanches. Si on tire au hasard (sans remise) deux chaussettes de ce tiroir, quelle est la probabilité de tirer une paire de chaussettes blanches ?

Solution. On considère l'événement $A =$ « la première chaussette tirée est blanche » et l'événement $B =$ « la seconde chaussette tirée est blanche ». On cherche $\mathbb{P}(A \cap B)$. La probabilité que la première chaussette soit blanche est $\mathbb{P}(A) = 4/10 = 2/5$. Une fois que l'on a tiré une chaussette blanche, c'est-à-dire que A est réalisé, le tiroir ne contient plus que 6 chaussettes noires et 3 blanches. La probabilité que la seconde chaussette tirée soit blanche est alors $\mathbb{P}(B|A) = 3/9 = 1/3$. On obtient en appliquant la formule des probabilités composées :

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) = \frac{2}{5} \times \frac{1}{3} = \frac{2}{15}\,.$$

3.2.2 Formule des probabilités totales

Considérons deux événements A et B, avec $\mathbb{P}(A) \in]0,1[$. La formule des probabilités composées appliquée au couple (A,B) nous donne la relation (9). On peut aussi appliquer cette formule au couple (A^c,B), ce qui donne :

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(A^c)\mathbb{P}(B|A^c)\,. \tag{10}$$

Or, $A \cap B$ et $A^c \cap B$ sont disjoints, et leur union donne B. On a donc :

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B)\,.$$

En substituant (9) et (10) dans cette équation, on obtient la formule des probabilités totales :

Proposition 12. Soient A et B deux événements, avec $\mathbb{P}(A) \in]0,1[$. On a :

$$\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c)\,.$$

Exercice 3.

Reprendre l'exercice précédent et calculer la probabilité de l'événement $B =$« la seconde chaussette tirée est blanche ».

Solution. On va faire ce calcul en faisant intervenir l'événement $A =$ « la première chaussette tirée est blanche ». On a $\mathbb{P}(A) = 4/10 = 2/5$ et $\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 3/5$. D'une part, une fois que l'événement A est réalisé, il reste dans le tiroir, après ce premier tirage, 6 chaussettes noires et 3 chaussettes blanches. Donc $\mathbb{P}(B|A) = 3/9 = 1/3$. D'autre part, une fois que l'événement A^c est réalisé, il reste dans le tiroir 5 chaussettes noires et 4 chaussettes blanches. Donc $\mathbb{P}(B|A^c) = 4/9$. En appliquant la formule des probabilités totales, on obtient que la probabilité de B est donnée par :

$$\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c) = \frac{2}{5} \times \frac{1}{3} + \frac{3}{5} \times \frac{4}{9} = \frac{2}{15} + \frac{4}{15} = \frac{2}{5}\cdot$$

On trouve donc ici que $\mathbb{P}(B) = \mathbb{P}(A)$, ce qui n'était pas évident *a priori*. En guise de nouvel exercice, on peut considérer le cas d'un tiroir contenant p chaussettes blanches et n chaussettes noires, avec $n \geqslant 1$, $p \geqslant 1$, et vérifier que $\mathbb{P}(A) = \mathbb{P}(B) = p/(n+p)$ en suivant la démonstration précédente.

On peut écrire une version plus générale de la proposition 12, après avoir posé la définition suivante :

Définition 5. On appelle système complet d'événements une partition de Ω, c'est-à-dire une famille d'événements $A_1, \ldots, A_n$ non-vides et deux à deux disjoints (incompatibles), tels que $\Omega = A_1 \cup \cdots \cup A_n$.

Par exemple, si A est un événement quelconque distinct de $\varnothing$ et Ω, alors (A, A^c) est un système complet d'événements.

Proposition 13. Soit $A_1, \ldots, A_n$ un système complet d'événements, avec, pour tout $i = 1, \ldots, n$, $\mathbb{P}(A_i) > 0$. Alors pour tout événement B, on a :

$$\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(B|A_1) + \mathbb{P}(A_2)\mathbb{P}(B|A_2) + \cdots + \mathbb{P}(A_n)\mathbb{P}(B|A_n)\,.$$

Démonstration. Comme $A_1 \cup \cdots \cup A_n = \Omega$, on a :

$$B = B \cap \Omega = B \cap (A_1 \cup \cdots \cup A_n) = (B \cap A_1) \cup \cdots \cup (B \cap A_n)\,,$$

et comme les événements $B \cap A_i$ sont deux à deux disjoints :

$$\mathbb{P}(B) = \mathbb{P}(B \cap A_1) + \cdots + \mathbb{P}(B \cap A_n)\,.$$

On applique enfin la formule des probabilités composées à chaque terme de la somme, ce qui donne la formule espérée. ■

Dans le cas $n = 2$, on retrouve exactement l'énoncé de la proposition 12 de la page précédente.

Exercice 4.

Une célèbre marque de montres produit des montres de très bonne qualité, avec une probabilité de panne de 10^{-2}. Malheureusement, des contrefaçons de mauvaise qualité ont été mises en circulation, et celles-ci occupent $20\ \%$ du marché. Ces contrefaçons ont une probabilité de panne de $0{,}1$. Si on achète une montre, quelle est la probabilité qu'elle ne marche pas ?

Solution. On introduit les événements $A =$ « la montre est une contrefaçon » et $B =$ « la montre ne marche pas ». D'après l'énoncé, on a $\mathbb{P}(A) = 0{,}2$ et $\mathbb{P}(B|A) = 0{,}1$. Si la montre est un article authentique, alors elle a une chance sur 100 de ne pas marcher, donc $\mathbb{P}(B|A^c) = 0{,}01$. D'après la formule des probabilités totales, la probabilité que la montre ne marche pas est :

$$\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c) = 0,2 \times 0,1 + (1 - 0{,}2) \times 0{,}01 = 2{,}8\ \%\,.$$

3.3 Formule de Bayes

Dans les exercices précédents, on cherche la probabilité d'un événement, l'« effet », en fonction de celles d'événements qui ont pu l'influer, les « causes », et on détermine la probabilité de l'effet en fonction des probabilités des causes. Dans certaines situations, on peut être amené à adopter le point de vue inverse, dans le sens où on observe un certain effet, et on cherche les causes qui ont pu conduire à cet effet. Supposons que, dans l'exercice précédent, on achète une montre qui s'avère défectueuse. Quelle est la probabilité que ce soit une contrefaçon ? Notons A l'événement « la montre est une contrefaçon », et B l'événement « la montre ne marche pas ». D'après les calculs faits dans l'exercice précédent, nous voyons que $\mathbb{P}(A) = 0{,}2$, $\mathbb{P}(B|A) = 0{,}1$, donc $\mathbb{P}(A \cap B) = 0{,}02$, et la probabilité de B est $0{,}028$. La probabilité de A sachant B est donc :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{0{,}02}{0{,}028} = \frac{10}{14} \simeq 71\ \%\,.$$

Ce type de raisonnements se généralise et conduit à la formule de Bayes :

Proposition 14 (Formule de Bayes). Soit $A_1, \ldots, A_n$ un système complet d'événements de probabilités non nulles. Alors pour tout événement de probabilité non nulle B, pour tout $i \in [\![1, n]\!]$, on a :

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(A_1)\mathbb{P}(B|A_1) + \cdots + \mathbb{P}(A_n)\mathbb{P}(B|A_n)}\,.$$

En particulier, si A est un événement tel que $0 < \mathbb{P}(A) < 1$ (et donc n'est ni impossible, ni certain), alors :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c)}\,.$$

Démonstration. On commence par écrire la définition d'une probabilité conditionnelle :

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)}\,.$$

Le numérateur peut se réécrire $\mathbb{P}(A_i \cap B) = \mathbb{P}(A_i|B)\mathbb{P}(B)$. On applique enfin la formule des probabilités totales au dénominateur. ∎

Exercice 5.

Reprenons encore une fois l'exercice 2 de la page 956. On tire une paire de chaussettes dans un tiroir qui contient initialement 6 chaussettes noires et 4 chaussettes blanches. On ne regarde pas la première chaussette tirée, on regarde la seconde et on constate qu'elle est blanche : quelle est la probabilité que la première chaussette tirée soit blanche aussi ?

Solution. On note $A =$ « la première chaussette est blanche » et $B =$ « la seconde chaussette est blanche ». Sans l'information sur la couleur de la seconde chaussette tirée, la probabilité que la première chaussette soit blanche est $\mathbb{P}(A) = 4/10 = 2/5$. Avec l'information supplémentaire, on cherche la probabilité conditionnelle $\mathbb{P}(A|B)$. En appliquant la formule de Bayes, et en utilisant toutes les valeurs des probabilités conditionnelles qu'on a déjà obtenues, on obtient :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c)} = \frac{2/5 \times 1/3}{2/5 \times 1/3 + 3/5 \times 4/9} = \frac{1}{3}\,.$$

3.4 Indépendance de deux événements

Il arrive que la réalisation d'un événement A n'ait aucune influence sur la probabilité de réalisation d'un autre événement B. Prenons l'exemple du lancer de deux dés. Si A est l'événement « le premier dé donne un 6 », et B est l'événement « le second dé donne un 6 », alors il est clair que, comme le second dé n'a pas de relation avec le premier, on a $\mathbb{P}(B|A) = \mathbb{P}(B|A^c) = \mathbb{P}(B)$. Cette relation est vérifiée si, et seulement si, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Cette dernière formule, qui a un sens même quand les probabilités s'annulent, est celle que nous adopterons comme définition.

Définition 6. Deux événements A et B sont dits indépendants si :

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\,.$$

Remarque. Il ne faut pas confondre les notions d'événements indépendants et d'événements incompatibles. Si deux événements sont incompatibles, alors la réalisation de l'un d'entre eux influe beaucoup sur la probabilité de réalisation de l'autre, puisqu'en fait elle l'annule. L'exercice suivant montre que deux événements ne peuvent être à la fois indépendants et incompatibles que si l'un des deux est de probabilité nulle.

Exercice 6.
Montrer que, si A et B sont deux événements incompatibles et indépendants, alors $\mathbb{P}(A) = 0$ ou $\mathbb{P}(B) = 0$.
Solution. D'une part, A et B sont indépendants, donc $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. D'autre part, A et B sont incompatibles, donc $A \cap B = \varnothing$ et $\mathbb{P}(A \cap B) = 0$. Ainsi $\mathbb{P}(A)\mathbb{P}(B) = 0$ ce qui donne le résultat énoncé.

Exemple. On lance deux dés, et on considère l'événement A =« le premier dé donne un 6 », et l'événement B =« le second dé donne un 6 ». Alors $\mathbb{P}(A) = 1/6$, $\mathbb{P}(B) = 1/6$ et $\mathbb{P}(A \cap B) = 1/36$: on a bien $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, les deux événements sont indépendants. De façon générale, tout événement qui peut se décrire uniquement en termes du résultat du premier dé est indépendant de tout événement qui peut se décrire uniquement en termes du résultat du second dé.

Il arrive fréquemment que l'intuition soit mise en défaut, et que deux événements paraissent dépendants alors qu'ils sont indépendants, ou vice-versa.

Exemple. On tire à pile ou face avec une pièce équilibrée trois fois. On modélise cette expérience par l'espace fondamental $\Omega = \{P, F\}^3$ et par la probabilité $\mathbb{P}$ uniforme sur Ω. On considère l'événement A =« les trois lancers donnent le même résultat », et l'événement B =« on obtient au plus un pile P ». Ces événements ne paraissent pas indépendants, du moins intuitivement, et pourtant on va montrer qu'ils le sont. Décrivons les ensembles A et B :

$$A = \{(FFF), (PPP)\}, \quad B = \{(FFF), (PFF), (FPF), (FFP)\},$$

ce qui donne $\mathbb{P}(A) = 2/8 = 1/4$ et $\mathbb{P}(B) = 4/8 = 1/2$, car $\operatorname{card}(\Omega) = 2^3 = 8$. On obtient aussi $A \cap B = \{(FFF)\}$, et donc $\mathbb{P}(A \cap B) = 1/8$. On a :

$$\mathbb{P}(A \cap B) = \frac{1}{8} = \frac{1}{4} \times \frac{1}{2} = \mathbb{P}(A)\mathbb{P}(B),$$

ce qui montre que les événements A et B sont indépendants.

Remarque. La notion d'indépendance est liée à la probabilité en vigueur $\mathbb{P}$. Deux événements indépendants sous une certaine probabilité $\mathbb{P}$ n'ont pas de raison de l'être si on change la probabilité, comme le montre l'exemple suivant.

Exemple. On considère l'ensemble fondamental $\Omega = \{0, 1\}^2 = \{(i, j) \mid i, j = 0 \text{ ou } 1\}$. On introduit les événements :

$$A = \{(i, j) \in \Omega \mid i = 1\} = \{(1, 0)\} \cup \{(1, 1)\},$$
$$B = \{(i, j) \in \Omega \mid i = j\} = \{(0, 0)\} \cup \{(1, 1)\}.$$

On a bien sûr $A \cap B = \{(1,1)\}$.
1. Prenons comme probabilité $\mathbb{P}_1 =$ la probabilité uniforme sur Ω. Alors $\mathbb{P}_1(A) = 1/2$, $\mathbb{P}_1(B) = 1/2$ et $\mathbb{P}_1(A \cap B) = 1/4$, donc $\mathbb{P}_1(A \cap B) = \mathbb{P}_1(A)\mathbb{P}_1(B)$, ce qui montre que A et B sont indépendants sous $\mathbb{P}_1$.
2. Changeons de probabilité et prenons $\mathbb{P}_2$ définie par :

$$\mathbb{P}_2((0,0)) = \mathbb{P}_2((1,1)) = \frac{1}{3}, \qquad \mathbb{P}_2((0,1)) = \mathbb{P}_2((1,0)) = \frac{1}{6}\cdot$$

Alors $\mathbb{P}_2(A) = 1/2$, $\mathbb{P}_2(B) = 2/3$ et $\mathbb{P}_2(A \cap B) = 1/3$, donc $\mathbb{P}_2(A \cap B) = \mathbb{P}_2(A)\mathbb{P}_2(B)$, ce qui montre que A et B sont aussi indépendants sous $\mathbb{P}_2$.
3. Considérons enfin la probabilité $\mathbb{P}_3$ définie par :

$$\mathbb{P}_3((0,0)) = \frac{1}{2}, \qquad \mathbb{P}_3((0,1)) = \mathbb{P}_3((1,0)) = \mathbb{P}_3((1,1)) = \frac{1}{6}\cdot$$

Alors $\mathbb{P}_3(A) = 1/3$, $\mathbb{P}_3(B) = 2/3$ et $\mathbb{P}_3(A \cap B) = 1/6$, et donc $\mathbb{P}_3(A \cap B) \neq \mathbb{P}_3(A)\mathbb{P}_3(B)$ ce qui montre que A et B ne sont pas indépendants sous $\mathbb{P}_3$.

Il faut donc faire attention et ne pas faire d'extrapolation abusive. En particulier, si deux événements A et B sont indépendants, et C est un événement de probabilité non nulle, alors on n'a pas forcément $\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$.

Exemple. Reprenons l'exemple précédent avec la probabilité uniforme $\mathbb{P}_1$ et considérons $C = \{(0,1)\} \cup \{(1,1)\}$. On vérifie que :

$$\mathbb{P}(A|C) = 1/2,\ \mathbb{P}(B|C) = 1/2 \quad \text{et} \quad \mathbb{P}(A \cap B|C) = 1/2 \neq 1/4 = \mathbb{P}(A|C)\mathbb{P}(B|C).$$

La proposition suivante permet d'éviter dans une situation donnée de faire plusieurs fois le même calcul inutilement. Elle offre aussi la possibilité de démontrer que A et B sont indépendants en considérant A^c et B par exemple.

Proposition 15. Si A et B sont indépendants, alors A et B^c sont indépendants, A^c et B sont indépendants, et A^c et B^c sont indépendants.

Démonstration. Il suffit de montrer la première conclusion, les autres s'en déduisent en échangeant les rôles de A et B. Comme $A = (A \cap B^c) \cup (A \cap B)$ avec union disjointe, on peut écrire que $\mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B)$ et donc :

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^c)\,,$$

ce qui montre que A et B^c sont indépendants. ■

Exercice 7.

On lance deux dés, et on considère l'événement $A =$ « le premier dé donne un 6 », l'événement $B =$ « le second dé donne un 6 » et l'événement $C =$ « les résultats des deux dés sont différents ». Montrer que ces trois événements sont indépendants deux à deux.
Solution. On a déjà vu que les événements A et B étaient indépendants et qu'on avait :

$$\mathbb{P}(A) = \mathbb{P}(B) = 1/6.$$

Le complémentaire de C est l'événement $C^c =$ « les résultats des deux dés sont identiques », qui est un ensemble plus facile à dénombrer, et on trouve $\mathbb{P}(C^c) = 6/36 = 1/6$. On a aussi $A \cap C^c = \{(6,6)\}$, donc :

$$\mathbb{P}(A \cap C^c) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = \mathbb{P}(A)\mathbb{P}(C^c)\,,$$

ce qui montre que A et C^c sont indépendants. D'après la proposition 15 de la page précédente, A et C sont donc indépendants.
Enfin, comme $B \cap C^c = \{(6,6)\}$ on a de même $\mathbb{P}(B \cap C^c) = 1/36 = \mathbb{P}(B)\mathbb{P}(C^c)$: B et C^c sont indépendants, donc B et C sont indépendants.

3.5 Indépendance d'une famille d'événements

Dans l'exercice précédent, les trois événements A, B et C sont indépendants deux à deux. Cependant, ces trois événements ne sont pas indépendants, dans le sens où la réalisation de deux d'entre eux influe fortement sur la réalisation du troisième. En fait, dans cet exercice, si A et B se réalisent, alors C devient impossible : $\mathbb{P}(C|A \cap B) = 0 \neq 1/6 = \mathbb{P}(C)$. Ceci montre que la notion d'indépendance d'une famille d'événements ne se réduit pas à une indépendance deux à deux. On adoptera la définition suivante.

Définition 7. Une famille d'événements $A_1, \ldots, A_n$ est dite indépendante ou mutuellement indépendante si, quel que soit $p \in [\![2, n]\!]$, et quelle que soit la sous-famille de p événements $A_{i_1}, \ldots, A_{i_p}$, on a :

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_p}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_p})\,.$$

Dorénavant, quand on parle d'événements indépendants, on comprend l'indépendance comme dans la définition 7.
Dans le cas d'une famille de deux événements $n = 2$, on retrouve bien sûr la définition 6 de la page 959, car les seuls choix possibles sont $i_1 = 1$ et $i_2 = 2$.
Dans le cas d'une famille de trois événements $n = 3$, les conditions à vérifier sont au nombre de quatre. On a trois conditions pour $p = 2$:

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)\,,$$
$$\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_3)\,,$$
$$\mathbb{P}(A_2 \cap A_3) = \mathbb{P}(A_2)\mathbb{P}(A_3)\,,$$

et une condition pour $p = 3$:

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)\,.$$

Exercice 8.
Reprendre l'exercice précédent du lancer de deux dés (avec $A_1 = A$, $A_2 = B$, $A_3 = C$). Montrer que les événements A_1, A_2 et A_3 ne sont pas indépendants.
Solution. D'après l'exercice précédent, les événements sont deux à deux indépendants, c'est-à-dire qu'ils vérifient toutes les conditions pour $p = 2$, mais $A_1 \cap A_2 \cap A_3 = \varnothing$,

d'où $\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0$, tandis que $\mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = 1/8$. La dernière condition n'est donc pas vérifiée, ce qui montre que les événements A_1, A_2 et A_3 ne sont pas indépendants.

Proposition 16. Soit $A_1, \ldots, A_n$ une famille d'événements indépendants. Pour chaque i, on prend $A'_i = A_i$ ou A_i^c. Alors les événements $A'_1, \ldots, A'_n$ sont indépendants.

Démonstration. On fixe n. On note p le nombre d'indices i pour lesquels $A'_i = A_i^c$. On va démontrer la propriété par récurrence sur p. Pour $p \in [\![0, n]\!]$ on note l'assertion (H_p) : Pour tout sous-ensemble d'indices $J' \subset [\![1, n]\!]$, pour tout sous-ensemble d'indices $J \subset [\![1, n]\!]$ avec $J' \cap J = \varnothing$ et $\operatorname{card}(J) = p$, on a :

$$\mathbb{P}\left(\bigcap_{j\in J} A_j^c \bigcap \bigcap_{j\in J'} A_j\right) = \prod_{j\in J} \mathbb{P}(A_j^c) \prod_{j\in J'} \mathbb{P}(A_j)\,.$$

(H_0) est immédiat car $(A_j)_{j\in[\![1,n]\!]}$ est une famille indépendante.

Supposons (H_p) pour un $p \in [\![0, n-1]\!]$. Soient deux sous-ensembles d'indices J et J' avec $J' \cap J = \varnothing$ et $\operatorname{card}(J) = p+1$. On découpe J en deux : $J = J_p \cup \{j_{p+1}\}$, avec $\operatorname{card}(J_p) = p$. On pose $C_p = \bigcap_{j\in J_p} A_j^c \cap \bigcap_{j\in J'} A_j$. Comme $C_p \cap A_{j_{p+1}}^c$ et $C_p \cap A_{j_{p+1}}$ sont disjoints et que leur union forme C_p, on a :

$$\mathbb{P}(C_p \cap A_{j_{p+1}}^c) = \mathbb{P}(C_p) - \mathbb{P}(C_p \cap A_{j_{p+1}})\,. \tag{11}$$

L'événement C_p est une intersection d'un certain nombre d'événements A_i et de p événements A_i^c. De même pour $C_p \cap A_{j_{p+1}}$. On peut donc appliquer l'hypothèse de récurrence aux deux termes du membre de droite de (11) :

$$\mathbb{P}(C_p \cap A_{j_{p+1}}^c) = \prod_{j\in J_p} \mathbb{P}(A_j^c) \prod_{j\in J'} \mathbb{P}(A_j) - \prod_{j\in J_p} \mathbb{P}(A_j^c) \prod_{j\in J'} \mathbb{P}(A_j)\mathbb{P}(A_{j_{p+1}})\,. \tag{12}$$

Or on a :

$$\bigcap_{j\in J} A_j^c \bigcap \bigcap_{j\in J'} A_j = C_p \cap A_{j_{p+1}}^c\,.$$

En regroupant les termes du membre de droite de (12) :

$$\mathbb{P}\left(\bigcap_{j\in J} A_j^c \bigcap \bigcap_{j\in J'} A_j\right) = \prod_{j\in J_p} \mathbb{P}(A_j^c)(1 - \mathbb{P}(A_{j_{p+1}})) \prod_{j\in J'} \mathbb{P}(A_j)\,,$$

et comme $1 - \mathbb{P}(A_{j_{p+1}}) = \mathbb{P}(A_{j_{p+1}}^c)$ et $J = J_p \cup \{j_{p+1}\}$ la récurrence est démontrée. ■

Exercice 9.

Un avion quadriréacteur a besoin d'au moins deux moteurs pour pouvoir voler. Un avion biréacteur a besoin d'au moins un moteur pour pouvoir voler. La probabilité de panne d'un réacteur lors d'un vol transatlantique est $p \in [0, 1]$. On suppose que les pannes des réacteurs sont indépendantes les unes des autres. Quel avion vous semble le plus sûr ?

Solution. On note A_j l'événement « le réacteur numéro j tombe en panne » et A l'événement « l'avion est forcé d'amerrir ».
1) On considère le cas d'un avion biréacteur. L'avion est forcé d'amerrir si les deux réacteurs tombent en panne, c'est-à-dire $A = A_1 \cap A_2$. Comme les événements A_1 et A_2 sont indépendants, on a donc :

$$\mathbb{P}(A) = \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2) = p^2 .$$

2) On considère le cas d'un avion quadriréacteur. L'avion est forcé d'amerrir si trois des quatre réacteurs tombent en panne, ou les quatre, donc :

$$A = \bigcup_{j=1}^{4} \left(\bigcap_{i \neq j} A_i \cap A_j^c \right) \bigcup \left(\bigcap_{i=1}^{4} A_i \right) .$$

Les cinq événements $\bigcap_{i \neq j} A_i \cap A_j^c$, $j \in [\![1,4]\!]$ et $\bigcap_{i=1}^{4} A_i$ étant disjoints, la probabilité de A est la somme des probabilités de ces événements :

$$\mathbb{P}(A) = \sum_{j=1}^{4} \mathbb{P}\left(\bigcap_{i \neq j} A_i \cap A_j^c \right) + \mathbb{P}\left(\bigcap_{i=1}^{4} A_i \right) .$$

On peut alors utiliser l'indépendance des événements A_j et la proposition 16 de la page précédente pour obtenir :

$$\mathbb{P}(A) = 4p^3(1-p) + p^4 = 4p^3 - 3p^4 .$$

En général, $p \ll 1$ (du moins, on peut l'espérer !), et alors $4p^3 - 3p^4 \ll p^2$ ce qui montre que les quadriréacteurs sont plus sûrs. En fait, on vérifie que $4p^3 - 3p^4 - p^2 < 0$ si, et seulement si, $4p - 3p^2 - 1 < 0$ si, et seulement si, $-3(p-1)(p-1/3) < 0$ si, et seulement si, $p < 1/3$ (on ne considère que les $p \in [0,1]$). Pour compenser cet effet, on réclame des critères de qualité très serrés pour les moteurs des biréacteurs long-courrier.

Exercice type corrigé

V.2.0 Problème de météorologie

Dans un certain pays, le temps est soit sec (S) soit humide (H). La règle d'évolution du temps est la suivante : si le temps est sec aujourd'hui, il sera sec demain avec la probabilité $4/5$ (et donc humide avec la probabilité $1/5$). Si le temps est humide aujourd'hui, il sera humide demain avec la probabilité $3/5$.

Appelons S_n (resp. H_n) l'événement « le temps est sec (resp. humide) le $n^{\text{ème}}$ jour ».
On note s_n et h_n les probabilités de ces événements.

On note également x_n le vecteur colonne : $x_n = \begin{pmatrix} s_n \\ h_n \end{pmatrix}$.

1. En utilisant la règle d'évolution, exprimer s_{n+1} et h_{n+1} en fonction de s_n et h_n.
2. En déduire que l'on a $x_{n+1} = A\,x_n$, où A est une matrice à déterminer.

3. Nous sommes dimanche et il fait sec. Quelle est la probabilité que le temps soit sec mardi ? soit humide mercredi ?

Solution.
1. Pour calculer $\mathbb{P}(S_{n+1})$, on applique la formule des probabilités totales en conditionnant par la famille (S_n, H_n) qui est un système complet d'événements (i.e. une partition de Ω) :

$$s_{n+1} = \mathbb{P}(S_{n+1}) = \mathbb{P}(S_{n+1}|S_n)\mathbb{P}(S_n) + \mathbb{P}(S_{n+1}|H_n)\mathbb{P}(H_n) = \frac{4}{5}s_n + \frac{2}{5}h_n\,.$$

On calcule de même :

$$h_{n+1} = \mathbb{P}(H_{n+1}) = \mathbb{P}(H_{n+1}|S_n)\mathbb{P}(S_n) + \mathbb{P}(H_{n+1}|H_n)\mathbb{P}(H_n) = \frac{1}{5}s_n + \frac{3}{5}h_n\,.$$

2. De manière algébrique, en posant $x_n = \begin{pmatrix} s_n \\ h_n \end{pmatrix}$ et $A = \begin{pmatrix} 4/5 & 2/5 \\ 1/5 & 3/5 \end{pmatrix}$ les deux dernières relations s'écrivent :

$$x_{n+1} = Ax_n\,.$$

3. Le temps de dimanche est sec, donc $x_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Le temps de mardi est décrit par x_2, et celui de mercredi par x_3. Le calcul donne :

$$x_2 = A^2x_0 = \begin{pmatrix} 18/25 \\ 7/25 \end{pmatrix}\ ,\ \text{car } A^2 = \begin{pmatrix} 18/25 & 14/25 \\ 7/25 & 11/25 \end{pmatrix}\ ,$$

donc la probabilité qu'il fasse sec mardi est $18/25$. De même,

$$x_3 = A^3x_0 = \begin{pmatrix} 86/125 \\ 39/125 \end{pmatrix}\ ,\ \text{car } A^3 = \begin{pmatrix} 86/125 & 78/125 \\ 39/125 & 47/125 \end{pmatrix}\ ,$$

donc la probabilité qu'il fasse humide mercredi est $39/125$.

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

V.2.1 Un peu de combinatoire

Le poker consiste en une main de cinq cartes tirées parmi un jeu de 32 cartes. Combien y a-t-il de mains possibles ? Combien y a-t-il de mains permettant d'annoncer

1. un carré (quatre cartes de même hauteur).
2. un full (trois cartes de même hauteur plus deux cartes de même hauteur).
3. un brelan (trois cartes de même hauteur, les deux autres cartes ne formant rien de particulier).

V.2.2

Soient E et F des ensembles finis, avec $p = \operatorname{card}(E)$ et $n = \operatorname{card}(F)$.

1. On suppose que $n = 2 \leqslant p$. Calculer le nombre de surjections de E dans F.
2. On suppose que $p \leqslant n$ et que E et F sont tous les deux munis d'un ordre total. Calculer le nombre d'applications strictement croissantes de E dans F.

V.2.3 Soient A et B des événements avec $\mathbb{P}(A) = 1/2$ et $\mathbb{P}(B) = 1/3$.

1. Donner un encadrement de $\mathbb{P}(A \cup B)$ et $\mathbb{P}(A \cap B)$.
2. Déterminer $\mathbb{P}(A \cup B)$ lorsque A et B sont incompatibles puis lorsque $\mathbb{P}(A \cap B) = 1/4$.

V.2.4 Un joueur achète k billets à une loterie. Chaque billet a une chance sur n d'être gagnant.

1. Calculer le nombre k_n de billets que le joueur doit acheter pour avoir plus d'une chance sur deux de gagner.
   Application numérique : $n = 10$, $n = 100$.
2. En supposant n grand, montrer que $k_n/n \simeq \alpha$ pour un α qu'on identifiera.

V.2.5 Un bibliothécaire fou permute au hasard les n livres de sa bibliothèque. Quelle est la probabilité que les volumes 1 et 2 de « Guerre et Paix » de Tolstoï se retrouvent côte à côte dans le bon ordre ?

V.2.6 Mon voisin a deux enfants dont une fille. Quelle est la probabilité que l'autre enfant soit un garçon ?
Ma voisine a deux enfants. Le plus jeune est une fille. Quelle est la probabilité pour que le plus âgé soit un garçon ?

V.2.7 Montrer que lorsque l'on tire une carte au hasard dans un jeu de 52 cartes, la couleur et la valeur sont indépendantes.

V.2.8 On lance deux dés non pipés, un noir et un blanc. Montrer que les événements suivants sont deux à deux indépendants, mais ne sont pas mutuellement indépendants :

- « le chiffre du dé noir est pair »,
- « le chiffre du dé blanc est pair »,
- « les chiffres des deux dés ont même parité ».

V.2.9 En Belgique, on mange deux types de frites : les frites traditionnelles à section rectangulaire et les frites new-look à section hexagonale. Parmi les frites que consomment les Flamands, il y a $65\,\%$ de frites traditionnelles alors que les Wallons en mangent $75\,\%$. L'équipe de Belgique de football (les fameux Diables rouges) est composée de sept Flamands et quatre Wallons. Un joueur est surpris à la mi-temps avec un cornet de frites hexagonales. Calculer la probabilité pour qu'il soit Flamand.

V.2.10 Monsieur le curé fait sa visite au château de la baronne, comme chaque mercredi, après les Vêpres. Sa 2CV est équipée d'essuie-glaces dont le moteur tombe en panne un voyage sur dix. La commande manuelle fonctionne une fois sur trois. À cette époque de l'année, il pleut trois fois sur quatre.

1. Calculer la probabilité pour que la baronne le voit arriver dans la cour du manoir en actionnant ses essuie-glaces d'une main et tenant son volant de l'autre.
2. Sachant qu'il pleut, calculer la probabilité pour qu'il soit obligé de terminer son chemin à pied, sous son parapluie.

V.2.11 On cherche un parapluie dans un meuble constitué de quatre tiroirs. La probabilité qu'il soit effectivement dans ce meuble est p, et s'il y est, il peut se trouver dans n'importe quel tiroir. Sachant qu'on a examiné les trois premiers tiroirs sans succès, quelle est la probabilité qu'il soit dans le quatrième ?

Variables aléatoires V.3

1 Lois et variables aléatoires

1.1 Définitions

Dans une expérience aléatoire, on s'intéresse souvent à des quantités particulières dépendant du résultat ω de l'expérience. Ainsi on peut être intéressé par le nombre de Pile dans une suite de lancers de pièces sans être préoccupé par le détail ou l'ordre de la suite. Dans cette situation, la quantité en question est a priori inconnue, mais on peut recenser toutes les valeurs possibles que peut prendre cette quantité et déterminer, une fois l'expérience faite, quelle valeur exacte elle prend. Une telle quantité est appelée variable aléatoire (en abrégé, v.a.), c'est donc une fonction $X : \omega \mapsto X(\omega)$ définie sur l'espace fondamental Ω.

Définition 1. Une variable aléatoire sur un espace de probabilité fini $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ est une application $X : \Omega \to E$ à valeurs dans un espace fini E.
La loi de X est la fonction $P_X : \mathcal{P}(E) \to [0,1]$ définie par :

$$P_X(A) := \mathbb{P}(X \in A), \text{ pour tout } A \in \mathcal{P}(E). \tag{1}$$

Il est d'usage de noter par des majuscules $(X, Y, \ldots)$ les v.a., réservant $x, y, \ldots$ pour désigner des valeurs déterministes. Dans (1) on a utilisé la notation abrégée :

$$\mathbb{P}(X \in A) = \mathbb{P}(\{X \in A\})$$

et

$$\{X \in A\} = \{\omega \in \Omega \mid X(\omega) \in A\},$$

utilisée couramment en probabilité.

Exemple. Considérons un lancer de dé. L'espace fondamental est $\Omega = [\![1,6]\!]$. Le résultat du lancer d'un dé est ici la variable aléatoire $X(\omega) = \omega$, à valeurs dans $E = [\![1,6]\!]$.

Exemple. Considérons un double lancer de dé. L'espace fondamental est $\Omega = [\![1,6]\!]^2$. On s'intéresse au résultat final qui est la somme des deux lancers, il est donné par la v.a. S définie par :

$$S : \left| \begin{array}{lcl} \Omega & \to & [\![2,12]\!] \\ \omega = (\omega_1, \omega_2) & \mapsto & S(\omega) = \omega_1 + \omega_2 \end{array} \right. .$$

Cette v.a. prend ses valeurs dans l'ensemble des entiers $E = [\![2,12]\!]$.

La proposition suivante justifie l'appellation loi de probabilité pour la fonction P_X.

Proposition 1.
1) La loi P_X d'une v.a. X est caractérisée par la donnée des nombres $\{p_X(x),\ x \in E\}$ définis par :

$$p_X(x) := P_X(\{x\}) \text{ pour tout } x \in E\,. \tag{2}$$

Pour retrouver la loi on somme comme suit :

$$P_X(A) = \sum_{x \in A} p_X(x) \text{ pour tout } A \in \mathcal{P}(E)\,.$$

2) La fonction P_X est une probabilité sur $(E, \mathcal{P}(E))$.

Démonstration. Soit $A \in \mathcal{P}(E)$. On peut décomposer $\{X \in A\}$ comme une union finie d'événements disjoints $\{X \in A\} = \bigcup_{x \in A} \{X = x\}$. La propriété d'additivité vérifiée par la probabilité $\mathbb{P}$ donne :

$$P_X(A) = \mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} p_X(x)\,,$$

ce qui montre le 1).
Pour montrer le 2) il faut montrer que P_X vérifie les axiomes d'une probabilité (définition 1 de la page 944).
(i) $P_X(E) = \sum_{x \in E} p_X(x) = \sum_{x \in E} \mathbb{P}(X = x)$. Or les événements $(\{X = x\} \mid x \in E)$ sont deux à deux disjoints et leur union forme Ω, donc $P_X(E) = 1$.
(ii) Soient A et B deux événements incompatibles (ou sous-ensembles deux à deux disjoints) de $\mathcal{P}(E)$. Alors :

$$P_X\big(A \cup B\big) = \sum_{x \in A \cup B} p_X(x) = \sum_{x \in A} p_X(x) + \sum_{x \in B} p_X(x) = P_X(A) + P_X(B)\,.$$

Ainsi P_X vérifie tous les axiomes de la définition d'une probabilité sur $\big(E, \mathcal{P}(E)\big)$. ∎

1.2 Histogrammes

Une loi de probabilité sur un espace E fini est caractérisée par la donnée des nombres $p_X(x)$, pour $x \in E$. Dans le cas courant où E est un sous-ensemble de l'ensemble des entiers $\mathbb{Z}$, la loi peut être représentée par un histogramme, c'est-à-dire un ensemble de bâtons de largeurs unité, centrés en $x \in E$ et de hauteurs $p_X(x)$.

Exemple. On reprend l'exemple du lancer de dé. L'espace fondamental est $\Omega = [\![1, 6]\!]$ muni de la probabilité uniforme. Le résultat du lancer d'un dé est la variable aléatoire $X(\omega) = \omega$, à valeurs dans $[\![1, 6]\!]$, et on a donc :

$$p_X(1) = \frac{1}{6}\,, \quad p_X(2) = \frac{1}{6}\,, \quad \ldots, \quad p_X(6) = \frac{1}{6}\,.$$

On représente cette loi par un histogramme sur la figure V.3.1(a).

(a) **(b)**

FIGURE V.3.1 – Histogramme de la loi d'un lancer de dé (a) et de la loi de la somme de deux lancers de dé (b).

Exemple. Considérons un double lancer de dé. L'espace fondamental est $\Omega = [\![1,6]\!]^2$ muni de la probabilité uniforme. Le résultat total est la somme des deux lancers, $S(\omega) = \omega_1 + \omega_2$, où $\omega = (\omega_1, \omega_2)$. Cette v.a. prend ses valeurs dans l'ensemble des entiers $[\![2,12]\!]$, et on obtient par des opérations de dénombrement élémentaires :

$$p_S(2) = \frac{1}{36}\,, \quad p_S(3) = \frac{2}{36}\,, \quad \ldots, \quad p_S(7) = \frac{6}{36}\,,$$
$$p_S(8) = \frac{5}{36}\,, \quad \ldots, \quad p_S(11) = \frac{2}{36}\,, \quad p_S(12) = \frac{1}{36}\,.$$

On représente cette loi par un histogramme sur la figure V.3.1(b).

L'aire de chaque bâton d'un histogramme situé au dessus d'un certain entier x est donc égale à la probabilité de l'événement $\{X = x\}$. Si on cherche la probabilité de l'événement $X \in [a,b]$, avec $a, b \in \mathbb{Z}$, $a < b$, alors il suffit de mesurer l'aire des bâtons compris entre a et b, puisque :

$$\mathbb{P}(X \in [a,b]) = \sum_{x=a}^{b} p_X(x)\,.$$

2 Quelques lois usuelles

2.1 Loi de Bernoulli

Définition 2 (Loi de Bernoulli). Soit $p \in [0,1]$. La loi de Bernoulli de paramètre p est la probabilité P sur $\big(\{0,1\}, \mathcal{P}(\{0,1\})\big)$ déterminée par :

$$P(\{0\}) = 1 - p\,, \quad P(\{1\}) = p\,.$$

On la note $\mathcal{B}(p)$.

On représente les histogrammes de la loi de Bernoulli pour différentes valeurs du paramètre p sur la figure V.3.2.

Toute v.a. ne prenant pour valeurs que 0 ou 1 suit une loi de Bernoulli, comme le montre la proposition suivante.

$p = 0{,}8$ $p = 0{,}5$ $p = 0{,}2$

FIGURE V.3.2 – Histogramme de la loi de Bernoulli de paramètre p

Proposition 2. Soit $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini quelconque.
Soit $X : \Omega \to \{0, 1\}$ une v.a. ne prenant pour valeurs que 0 ou 1.
Si l'on note $p = \mathbb{P}(X = 1)$, alors la loi de X est la loi de Bernoulli de paramètre p.

Démonstration.
La loi de X est caractérisée par les nombres $p_X(x) = \mathbb{P}(X = x)$, $x \in \{0, 1\}$. On a $p_X(1) = p$ et, comme il s'agit d'une loi de probabilité, on a $p_X(0) = 1 - p_X(1) = 1 - p$. La loi de X est donc bien une loi de Bernoulli de paramètre p. ■

On dit que X est une v.a. de Bernoulli de paramètre p. En abrégé, on écrit $X \sim \mathcal{B}(p)$ pour « X suit la loi $\mathcal{B}(p)$ ».

Un premier exemple de v.a. de Bernoulli est le suivant (jeu de Pile ou Face). Soit $\Omega := \{0, 1\}$ et $\mathbb{P}$ la probabilité $\mathbb{P}(\{0\}) = 1 - p$, $\mathbb{P}(\{1\}) = p$, où $p \in [0, 1]$ est un paramètre fixé. La fonction $X(\omega) := \omega$ est une v.a. de Bernoulli de paramètre p.

Les v.a. de Bernoulli apparaissent donc dans le jeu de Pile ou Face (« Pile » $:= \{X = 1\}$, « Face » $:= \{X = 0\}$), mais aussi de façon naturelle dans les espaces de probabilité finis $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ les plus généraux. En effet, considérons un espace de probabilité fini $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ arbitraire. Pour tout événement $A \in \mathcal{P}(\Omega)$, on définit la fonction indicatrice $\mathbf{1}_A : \Omega \to \mathbb{R}$ par :

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{si} \quad \omega \in A \\ 0 & \text{si} \quad \omega \notin A \end{cases} .$$

Il s'agit d'une v.a. car c'est une fonction de Ω dans $\{0, 1\}$. Comme la v.a. $\mathbf{1}_A$ ne prend pour valeurs que 0 ou 1, et comme $\mathbb{P}(\mathbf{1}_A = 1) = \mathbb{P}(A)$, la fonction indicatrice $\mathbf{1}_A$ est une v.a. de Bernoulli de paramètre $p = \mathbb{P}(A)$.

2.2 Loi binomiale et nombre de succès

Il existe de nombreuses situations où la question posée tourne autour du nombre de succès dans une suite d'expériences de Bernoulli. Considérons l'exemple de n lancers d'une pièce, avec une pièce qui tombe sur Pile [resp. Face] avec probabilité p [resp. $1 - p$] ($p \in [0, 1]$). On considère le nombre total S de succès (Pile) lors du jeu. Ici l'ensemble fondamental est $\Omega := \{F, P\}^n$, et la probabilité sur $\big(\Omega, \mathcal{P}(\Omega)\big)$ est $\mathbb{P}(\{\omega\}) = p^k(1 - p)^{n-k}$ si ω comporte exactement k Piles et $n - k$ Faces. On note A_k le sous-ensemble de Ω constitué des éléments qui contiennent exactement k Piles. On introduit la variable aléatoire :

$$S = \sum_{k=0}^{n} k \mathbf{1}_{A_k} ,$$

c'est-à-dire que $S(\omega) =$ nombre de Piles contenus dans ω. S prend ses valeurs dans $[\![0, n]\!]$, et on remarque que pour $k \in [\![0, n]\!]$, $\binom{n}{k}$ éléments ω de Ω comportent un total de k Piles. Puisque chacun de ces ω a même probabilité, on obtient :

$$\mathbb{P}(S = k) = \sum_{\omega \in A_k} \mathbb{P}(\{\omega\}) = \binom{n}{k} p^k (1-p)^{n-k} . \tag{3}$$

La loi de S est donc la loi du nombre de succès (Pile) dans un jeu de Pile ou Face, mais elle intervient couramment dans des modèles plus complexes, qui peuvent se reformuler comme la détermination du nombre de succès lors d'une suite de n expériences.

Définition 3 (Loi binomiale). Soient $n \in \mathbb{N}^*$ et $p \in [0, 1]$. On appelle loi binomiale de paramètres (n, p) la probabilité sur l'espace $[\![0, n]\!]$ déterminée par :

$$P(\{k\}) = \binom{n}{k} p^k (1-p)^{n-k} , \qquad k \in [\![0, n]\!] . \tag{4}$$

En appliquant la formule du binôme on retrouve bien que la somme des probabilités élémentaires de la loi binomiale vaut 1 :

$$\sum_{k=0}^{n} P(\{k\}) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1 .$$

Le nom de la loi vient de cette formule. On note cette loi $\mathcal{B}(n, p)$.

V.a. binomiale. Soit $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini. Soit $X : \Omega \to [\![0, n]\!]$ une v.a.. On dit que X est une v.a. binomiale de paramètres (n, p) si elle a pour loi la loi binomiale de paramètres (n, p). En abrégé, on écrit $X \sim \mathcal{B}(n, p)$ pour « X suit la loi $\mathcal{B}(n, p)$ ». Notez que, pour $n = 1$, on retrouve la loi de Bernoulli : $\mathcal{B}(1, p) = \mathcal{B}(p)$.

Proposition 3. Soit un espace de probabilité fini $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, sur lequel est définie une suite de n événements indépendants, notés $(A_i)_{i \in [\![1,n]\!]}$, qui possèdent la même probabilité de réalisation $p \in [0, 1]$. On note S le nombre d'événements A_i réalisés :

$$S := \text{card } \left\{ i \in [\![1, n]\!] \, , \, A_i \text{ est réalisé} \right\} = \sum_{i=1}^{n} \mathbf{1}_{A_i} .$$

Alors S suit la loi binomiale de paramètres (n, p).

Cette proposition recouvre l'exemple précédent du lancer de n pièces, où $\Omega = \{F, P\}^n$ et $A_i = \{\omega \mid \omega_i = P\}$.

Démonstration. La démonstration précédente avec la suite de tirages de pile ou face peut être directement appliquée à ce cadre plus général :

$$\mathbb{P}(S = k) = \sum_{\{i_1, \ldots, i_k\} \subset [\![1,n]\!]} \mathbb{P}\left(\left(\bigcap_{i \in \{i_1, \ldots, i_k\}} A_i \right) \bigcap \left(\bigcap_{i \in [\![1,n]\!] \setminus \{i_1, \ldots, i_k\}} A_i^c \right) \right) ,$$

où la somme est prise sur tous les sous-ensembles $\{i_1, \ldots, i_k\}$ à k éléments pris dans $[\![1, n]\!]$. Par indépendance des A_i, et en utilisant la proposition 16 de la page 963 on trouve que, pour tout sous-ensemble $\{i_1, \ldots, i_k\}$:

$$\begin{aligned}\mathbb{P}\left(\left(\bigcap_{i\in\{i_1,\ldots,i_k\}} A_i\right)\bigcap\left(\bigcap_{i\in[\![1,n]\!]\setminus\{i_1,\ldots,i_k\}} A_i^c\right)\right) &= \prod_{i\in\{i_1,\ldots,i_k\}} \mathbb{P}(A_i) \prod_{i\in[\![1,n]\!]\setminus\{i_1,\ldots,i_k\}} \mathbb{P}(A_i^c)\\ &= p^k(1-p)^{n-k}\,.\end{aligned}$$

Comme il y a $\binom{n}{k}$ sous-ensembles $\{i_1, \ldots, i_k\}$ à k éléments pris dans $[\![1, n]\!]$, on obtient finalement le résultat cherché. ■

On peut donner une autre démonstration de la formule (3), moins combinatoire, et faisant appel à des probabilités conditionnelles. Reprenons l'exemple de n lancers de pièce, où on cherche le nombre total de Piles obtenus. On note $A_{n,k}$ l'événement « on obtient exactement k Piles sur les n lancers », et $p_{n,k} := \mathbb{P}(A_{n,k})$. On introduit aussi l'événement $B =$ « le premier lancer donne un Pile ». D'après la formule des probabilités totales :

$$p_{n,k} = \mathbb{P}(A_{n,k}|B)\mathbb{P}(B) + \mathbb{P}(A_{n,k}|B^c)\mathbb{P}(B^c)\,.$$

Sachant B, l'événement $A_{n,k}$ se produit s'il y a exactement $k-1$ Piles sur les $n-1$ derniers lancers, donc $\mathbb{P}(A_{n,k}|B) = p_{n-1,k-1}$. De même, sachant B^c, l'événement $A_{n,k}$ se produit s'il y a exactement k Piles sur les $n-1$ derniers lancers, donc $\mathbb{P}(A_{n,k}|B^c) = p_{n-1,k}$. On obtient donc :

$$p_{n,k} = p_{n-1,k-1}p + p_{n-1,k}(1-p)\,.$$

En tenant compte du fait que $\mathbb{P}(A_{n,0}) = (1-p)^n$ et $\mathbb{P}(A_{n,n}) = p^n$, c'est-à-dire $p_{n,0} = (1-p)^n$ et $p_{n,n} = p^n$, on obtient une formule de récurrence en n donnant de proche en proche les valeurs de $(p_{n,k})_{0\leqslant k\leqslant n}$, qui est très similaire à la formule définissant le remplissage du triangle de Pascal. Pour compléter cette analogie, dans le cas $p \in]0, 1[$, notons :

$$q_{n,k} := \frac{p_{n,k}}{p^k(1-p)^{n-k}}\,.$$

La suite de nombres $(q_{n,k})_{0\leqslant k\leqslant n<\infty}$ satisfait alors :

$$q_{n,k} = q_{n-1,k-1} + q_{n-1,k}\,,$$

avec les conditions $q_{n,0} = q_{n,n} = 1$. Il s'agit cette fois-ci exactement de la relation du triangle de Pascal, et cela montre donc que $q_{n,k} = \binom{n}{k}$.

On peut donner quelques autres propriétés importantes de la loi binomiale. En particulier, il existe une relation simple entre $P(\{k+1\})$ et $P(\{k\})$, valable pour tout $k \in [\![0, n-1]\!]$:

$$\begin{aligned}P(\{k+1\}) &= \binom{n}{k+1}p^{k+1}(1-p)^{n-k-1} = \frac{n!}{(k+1)!(n-k-1)!}p^{k+1}(1-p)^{n-k-1}\\ &= \frac{n-k}{k+1}\,\frac{n!}{k!(n-k)!}\,\frac{p}{1-p}p^k(1-p)^{n-k} = \frac{n-k}{k+1}\,\frac{p}{1-p}\binom{n}{k}p^k(1-p)^{n-k}\\ &= \frac{n-k}{k+1}\,\frac{p}{1-p}P(\{k\})\,.\end{aligned}$$

FIGURE V.3.3 – Histogramme de la loi binomiale de paramètres (n, p)

Cette relation de récurrence partant de $P(\{0\}) = (1-p)^n$ est utile d'un point de vue pratique pour calculer les valeurs de $P(\{k\})$ dans le cas où n est grand, car le calcul de la formule donnée par (4) fait intervenir des facteurs factoriels qui peuvent être extrêmement grands. D'autre part, cette relation est utile pour déterminer la forme de la loi. On a :

$$\frac{P(\{k+1\})}{P(\{k\})} = \frac{n-k}{k+1}\,\frac{p}{1-p}\,,$$

qui est supérieur à 1 si $k < (n+1)p - 1$, ce qui montre que les valeurs des probabilités $P(\{k\})$ sont croissantes jusqu'à ce point. Le rapport est inférieur à 1 si $k > (n+1)p-1$, ce qui montre que les valeurs des probabilités $P(\{k\})$ sont décroissantes à partir de ce point. Dans le détail, notons $k_{max} = [(n+1)p]$. Le maximum de l'application $k \mapsto P(\{k\})$ est atteint uniquement en k_{max} si $(n+1)p$ n'est pas entier, et simultanément en $k_{max} - 1$ et en k_{max} si $(n+1)p$ est entier. La forme de la loi est visible lorsqu'on trace son histogramme. On représente les histogrammes de la loi binomiale pour $n = 20$ et pour différentes valeurs du paramètre p sur la figure V.3.3.

2.3 Echantillonnage et loi hypergéométrique

Le problème de l'échantillonnage est central en statistique. Le problème typique est le suivant. On souhaite connaître les intentions de vote dans une certaine population de m individus. Cette population se divise en m_a personnes qui souhaitent voter pour le candidat A, et m_b personnes qui souhaitent voter pour le candidat B. On suppose que $m_a + m_b = m$ (il

n'y a pas d'indécis). On effectue un sondage sur n personnes tirées au hasard, par exemple en tirant au hasard des numéros dans l'annuaire téléphonique (en supposant que tout le monde a un numéro de téléphone et un seul). Quelle est la loi du nombre de personnes déclarant vouloir voter pour A ?
De manière plus générale, on peut supposer qu'on a affaire à une population de m individus divisée en deux classes complémentaires, qu'on tire n individus au hasard dans cette population, et qu'on souhaite savoir combien d'individus tirés appartiennent à une classe donnée. La formulation la plus neutre et classique de ce problème est celle de l'urne : on dispose d'une urne contenant m_a boules noires et m_b boules blanches. On tire n boules de cette urne, et on compte le nombre de boules noires. Il y a en fait deux manières de tirer les n boules (ou individus) :

1. Après avoir tiré un certain individu, et déterminé sa classe, on peut décider de ne plus le tirer, et donc de choisir au hasard parmi les individus restants de la population. On parle alors de tirage sans remise. Dans l'exemple du sondage sur les intentions de vote, après avoir tiré un numéro de téléphone et sondé une personne, on décide de ne plus l'appeler pour ce sondage, et de tirer un nouveau numéro au hasard dans l'annuaire, hormis les numéros déjà tirés. Dans l'exemple de l'urne avec les boules, après avoir tiré une boule et noté sa couleur, on ne remet pas cette boule dans l'urne, et on continue le tirage.

2. On peut décider, après avoir tiré un certain individu, et déterminé sa classe, de continuer le sondage comme si de rien n'était, et de retirer un nouvel individu dans la population globale. On peut donc retirer le même individu. On parle alors de tirage avec remise. Dans l'exemple de l'urne avec les boules, après avoir tiré une boule et noté sa couleur, on remet cette boule dans l'urne, et on continue le tirage.

On peut modéliser le problème de l'urne comme suit.
L'espace fondamental est $\Omega = \{A, B\}^n$, où $\omega_i = A$ (resp. B) signifie que la $i^{ème}$ boule tirée est noire (resp. blanche). Les expériences du tirage avec ou sans remise correspondent à deux probabilités différentes sur $(\Omega, \mathcal{P}(\Omega))$.

- Pour un tirage avec remise, le problème se ramène en fait à une suite de n expériences de Bernoulli avec une probabilité de succès (i.e. tirer une boule noire) égale à $p = m_a/(m_a + m_b)$. Par conséquent, le nombre total S de boules noires tirées suit une loi binomiale de paramètres (n, p) :

$$\mathbb{P}(S = k) = \binom{n}{k} \left(\frac{m_a}{m_a + m_b}\right)^k \left(\frac{m_b}{m_a + m_b}\right)^{n-k} .$$

- Pour un tirage sans remise, il faut faire un peu de combinatoire. Pour éviter de se retrouver avec des sous-cas qui compliqueraient inutilement la situation, on va supposer que n est plus petit que m_a et m_b, si bien qu'il existe effectivement des configurations où l'on tire k boules noires, avec k compris entre 0 et n (si $m_a < n$, alors on ne peut clairement pas tirer n boules noires de l'urne, et si $m_b < n$, alors on ne peut clairement pas tirer n boules blanches, soit 0 boule noire).

D'une part, il y a en tout $\binom{m_a+m_b}{n}$ manières possibles de tirer n boules de l'urne. D'autre part, il y a $\binom{m_a}{k}\binom{m_b}{n-k}$ manières possibles de tirer k boules noires parmi les m_a boules noires de l'urne et $n-k$ boules blanches parmi les m_b boules blanches de l'urne. Comme tous ces tirages sont équiprobables, on obtient, pour tout $k \in [\![0, n]\!]$:

$$\mathbb{P}(S = k) = \frac{\binom{m_a}{k}\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} \cdot \tag{5}$$

En écrivant le second facteur combinatoire du numérateur de manière différente :

$$\binom{m_b}{n-k} = \frac{m_b!}{(n-k)!(m_b-n+k)!} = \frac{\frac{m_b!}{n!(m_b-n)!}\frac{n!}{k!(n-k)!}}{\frac{(m_b-n+k)!}{(m_b-n)!k!}} = \frac{\binom{m_b}{n}\binom{n}{k}}{\binom{m_b-n+k}{k}},$$

on obtient :

$$\mathbb{P}(S = k) = \frac{\binom{m_b}{n}}{\binom{m_a+m_b}{n}} \times \frac{\binom{n}{k}\binom{m_a}{k}}{\binom{m_b-n+k}{k}} \cdot$$

Le second facteur apparaît dans l'expression d'une fonction tabulée célèbre, la fonction hypergéométrique (plus exactement, ce facteur apparaît dans le développement en série entière de cette fonction, mais on abordera cette notion seulement dans le cours de L2). La fonction hypergéométrique est une solution d'une équation différentielle qu'on rencontre notamment en physique. Le nom de la loi définie par (5) vient de cette remarque.

Proposition 4. Lors du tirage sans remise de n boules dans une urne contenant m_a boules noires et m_b boules blanches, avec $m_a \geqslant n$ et $m_b \geqslant n$, le nombre total S de boules noires tirées suit la loi hypergéométrique de paramètres (m_a, m_b, n) :

$$\mathbb{P}(S = k) = \frac{\binom{m_a}{k}\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} = \frac{\binom{m_b}{n}}{\binom{m_a+m_b}{n}} \times \frac{\binom{n}{k}\binom{m_a}{k}}{\binom{m_b-n+k}{k}}, \quad k \in [\![0, n]\!].$$

3 Espérance de variables aléatoires réelles

Dans ce paragraphe on considère des variables aléatoires à valeurs dans un sous-espace fini E de $\mathbb{R}$ qu'on appelle des v.a. réelles. Bien qu'une v.a. réelle fluctue d'une expérience à l'autre, on peut lui attacher plusieurs valeurs « typiques » : sa valeur moyenne, sa valeur la plus probable, ... Toutes ces quantités sont d'intérêt, mais la plus importante, et souvent la plus utile, est la moyenne. Le vocabulaire utilise la dénomination d'espérance pour cette quantité, faisant ainsi référence au calcul de l'espérance de vie si important dans le domaine de l'assurance et qui a motivé au XIXe siècle de nombreux développements de la théorie des probabilités.

3.1 Définition de l'espérance

L'espérance d'une v.a. réelle est la moyenne pondérée des valeurs que peut prendre cette variable, la pondération étant donnée par les probabilités d'apparition de chaque valeur.

Définition 4 (Espérance).
Soient $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini et $X : \Omega \to E \subset \mathbb{R}$ une v.a. réelle. L'espérance de X est définie par :

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\})\,. \tag{6}$$

Elle est aussi donnée par :

$$\mathbb{E}[X] = \sum_{x \in E} x p_X(x)\,, \tag{7}$$

où $\{p_X(x), x \in E\}$ est la loi de X.

Le passage de (6) à (7) est une conséquence de l'additivité de la probabilité $\mathbb{P}$ (voir la preuve ci-dessous). La deuxième formule (7) est la plus importante car elle montre que l'espérance d'une v.a. réelle X ne dépend que de sa loi. Ainsi, une fois qu'on a calculé l'espérance d'une v.a. réelle de loi donnée, l'espérance de toute v.a. réelle qui suit cette loi sera égale à l'espérance en question, quel que soit l'espace de probabilité sur lequel la v.a. est définie.

Démonstration. Comme

$$\{X = x\} = \bigcup_{\omega \in \Omega | X(\omega) = x} \{\omega\}\,,$$

et que l'union est disjointe, la propriété d'additivité de la probabilité $\mathbb{P}$ donne :

$$\begin{aligned}\sum_{x \in E} x p_X(x) &= \sum_{x \in E} x \mathbb{P}\Big(\bigcup_{\omega \in \Omega | X(\omega) = x} \{\omega\}\Big) = \sum_{x \in E} x \sum_{\omega \in \Omega | X(\omega) = x} \mathbb{P}(\{\omega\}) \\ &= \sum_{x \in E} \sum_{\omega \in \Omega | X(\omega) = x} X(\omega)\mathbb{P}(\{\omega\})\,.\end{aligned}$$

La somme double peut se réécrire comme :

$$\sum_{x \in E} x p_X(x) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\})\,,$$

ce qui donne le résultat. ∎

Exemples. 1) Si A est un événement, l'espérance de la v.a. de Bernoulli $\mathbf{1}_A$ est :

$$\mathbb{E}[\mathbf{1}_A] = 0 \times \mathbb{P}(\mathbf{1}_A = 0) + 1 \times \mathbb{P}(\mathbf{1}_A = 1) = 0 \times \mathbb{P}(A^c) + 1 \times \mathbb{P}(A)\,,$$

soit :

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)\,. \tag{8}$$

Ceci montre que la notion d'espérance prolonge celle de probabilité, et elle en héritera des propriétés.
2) Si X est une v.a. constante, soit $X := a$, alors son espérance est $\mathbb{E}[X] = a$.

3.2 Propriétés élémentaires de l'espérance

Proposition 5 (Linéarité et positivité de l'espérance). Soient X et Y deux v.a. réelles définies sur le même espace de probabilité fini $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. Soient λ, μ deux réels.
1) La v.a. $Z := \lambda X + \mu Y$ est une v.a. réelle et $\mathbb{E}[Z] = \lambda \mathbb{E}[X] + \mu \mathbb{E}[Y]$.
2) Si $X \geqslant Y$ (dans le sens où $X(\omega) \geqslant Y(\omega)$ pour tout $\omega \in \Omega$), alors $\mathbb{E}[X] \geqslant \mathbb{E}[Y]$.

Démonstration. Pour montrer la linéarité de l'espérance, on utilise la définition (6) de l'espérance :

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{\omega\in\Omega} Z(\omega)\mathbb{P}(\{\omega\}) = \sum_{\omega\in\Omega} \big(\lambda X(\omega) + \mu Y(\omega)\big)\mathbb{P}(\{\omega\}) \\ &= \lambda \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\{\omega\}) + \mu \sum_{\omega\in\Omega} Y(\omega)\mathbb{P}(\{\omega\}) = \lambda\mathbb{E}[X] + \mu\mathbb{E}[Y]\,.\end{aligned}$$

Pour montrer la positivité, il suffit d'observer que $\mathbb{E}[X-Y] \geqslant 0$ car $X-Y$ est une v.a. qui ne prend que des valeurs positives, donc son espérance est une somme de termes positifs et est donc elle même positive. En appliquant la linéarité de l'espérance qui vient d'être démontrée on obtient le résultat voulu ∎

Le corollaire suivant ne fait que reformuler la proposition en des termes d'algèbre linéaire.

Corollaire 6. L'ensemble $L(\Omega, \mathbb{P})$ des v.a. réelles définies sur un même espace de probabilité fini $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ est un espace vectoriel, et $\mathbb{E} : \begin{array}{rcl} L(\Omega,\mathbb{P}) & \longrightarrow & \mathbb{R} \\ X & \longmapsto & \mathbb{E}[X] \end{array}$ est une forme linéaire positive.

3.3 Propriétés de transport

Considérons une v.a. $X : \Omega \to E$. Étant donnée une fonction f définie sur E (à valeurs dans l'ensemble fini $E' = f(E)$), la loi de la v.a. $Y = f(X)$ se calcule aisément à l'aide de la loi de X. Notons $p_X(x) = \mathbb{P}(X = x)$, $x \in E$, et $p_Y(y) = \mathbb{P}(Y = y)$, $y \in E'$. Puisque $\{Y = y\} = \bigcup_{x\in E \,|\, f(x)=y} \{X = x\}$ on a :

$$p_Y(y) = \sum_{x\in E} \mathbf{1}_y\big(f(x)\big)\, p_X(x)\,, \quad y \in E'\,, \tag{9}$$

où $\mathbf{1}_y\big(f(x)\big) = 1$ si $f(x) = y$ et 0 sinon.
Il arrive souvent qu'on ait besoin de calculer, non l'espérance de la variable aléatoire X, mais l'espérance d'une fonction $Y = f(X)$ de celle-ci, où f est une fonction de E dans $\mathbb{R}$. Si l'on applique la définition (6) de l'espérance, cela suppose qu'on calcule une somme sur Ω, qui peut être très grand. Si l'on applique la définition (7) de l'espérance (avec Y au lieu de X dans (7)), cela suppose qu'on doit d'abord calculer la loi de Y, ce qui peut être incommode. La proposition suivante montre qu'on n'a pas forcément besoin de calculer explicitement la loi d'une v.a. $Y = f(X)$ pour calculer son espérance.

Proposition 7. Soient $X : \Omega \to E$ une v.a. et $f : E \to \mathbb{R}$ une fonction à valeurs réelles. Alors $Y = f(X)$ est une v.a. réelle et l'on a :

$$\mathbb{E}[f(X)] = \sum_{x \in E} f(x) p_X(x) \,. \tag{10}$$

Démonstration. En utilisant (9) :

$$\begin{aligned}\sum_{y \in E'} y p_Y(y) = \sum_{y \in E'} \sum_{x \in E} y \mathbf{1}_y\big(f(x)\big)\, p_X(x) &= \sum_{x \in E} \Big[\sum_{y \in E'} y \mathbf{1}_{f(x)}(y)\Big] p_X(x) \\ &= \sum_{x \in E} f(x) p_X(x) \,.\end{aligned}$$

On obtient ainsi (10). ∎

3.4 Une application de la linéarité de l'espérance : formule d'inclusion-exclusion

Cette formule a déjà été démontrée dans le chapitre précédent, page 948. On propose ici une démonstration plus simple. La règle d'inclusion-exclusion s'applique à une suite $(A_i)_{i \in [\![1,n]\!]}$ d'événements pour lesquels on connaît a priori les probabilités des conjonctions d'événements. On peut alors calculer la probabilité de $A = A_1 \cup A_2 \cup \cdots \cup A_n$, i.e. la probabilité que « au moins l'un des A_i est réalisé ». Plus précisément, soit $(A_i)_{i \in [\![1,n]\!]}$ une suite d'événements. Alors :

$$\mathbb{P}\Big(\bigcup_{i=1}^n A_i\Big) = \sum_{k=1}^n (-1)^{k+1} \sum_{1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) . \tag{11}$$

Démonstration. D'une part on vérifie que :

$$1 - \mathbf{1}_{\bigcup_{i=1}^n A_i} = \prod_{i=1}^n (1 - \mathbf{1}_{A_i}) \,. \tag{12}$$

En effet,

$$\begin{aligned}\Big(1 - \mathbf{1}_{\bigcup_{i=1}^n A_i}\Big)(\omega) = 1 &\iff \omega \in \Big(\bigcup_{i=1}^n A_i\Big)^c = \bigcap_{i=1}^n A_i^c \iff \omega \in A_i^c \ \forall i \in [\![1,n]\!] \\ &\iff \big(1 - \mathbf{1}_{A_i}\big)(\omega) = 1 \ \forall i \in [\![1,n]\!] \iff \prod_{i=1}^n (1 - \mathbf{1}_{A_i})(\omega) = 1 \,.\end{aligned}$$

D'autre part, on vérifie que, pour tout $k \in [\![1,n]\!]$ et pour tous indices $1 \leqslant i_1 < \cdots < i_k \leqslant n$:

$$\prod_{j=1}^k \mathbf{1}_{A_{i_j}} = \mathbf{1}_{\bigcap_{j=1}^k A_{i_j}} \,.$$

En effet,

$$\prod_{j=1}^k \mathbf{1}_{A_{i_j}}(\omega) = 1 \iff \omega \in A_{i_j} \ \forall j \in [\![1,k]\!] \iff \omega \in \bigcap_{j=1}^k A_{i_j} \iff \mathbf{1}_{\bigcap_{j=1}^k A_{i_j}}(\omega) = 1 \,.$$

En prenant l'espérance de l'égalité (12) et en développant le membre de droite, on obtient directement la formule d'inclusion-exclusion (11) avec la relation (8) de la page 976. ■

3.5 Variance

Pour $k \in \mathbb{N}^*$, le $k^{ème}$ moment de la v.a. réelle X est par définition l'espérance $\mathbb{E}[X^k]$. La proposition 7 de la page précédente permet de calculer ces moments à l'aide de la loi de X. Les moments les plus utilisés sont le moment d'ordre un (l'espérance), et celui d'ordre deux sous forme centrée (la variance).

Définition 5. On appelle variance d'une variable aléatoire réelle X le nombre positif :

$$\mathrm{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] .$$

Ce nombre positif est l'écart quadratique moyen de X à son espérance $\mathbb{E}[X]$, il mesure donc l'ampleur de la dispersion de la loi de la v.a. X. En fait, puisque la variance a la dimension d'un carré, c'est plutôt l'écart-type de la v.a. X :

$$\sigma(X) := \sqrt{\mathrm{Var}(X)} ,$$

qui donne l'ordre de grandeur des fluctuations autour de la moyenne.

Proposition 8. Soit X une v. a. réelle.
1) Une expression équivalente pour la variance est (formule de Huygens) :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 . \tag{13}$$

2) Si $a, b \in \mathbb{R}$, alors la variance de la v.a. $aX + b$ est :

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X) . \tag{14}$$

3) Si $\mathrm{Var}(X) = 0$, alors $\mathbb{P}(X = \mathbb{E}[X]) = 1$.

La formule (13) est utile en pratique pour calculer la variance.

Démonstration. Pour le 1), on a part de la définition de $\mathrm{Var}(X)$, on développe le carré, et par linéarité, on obtient :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}\left[2X\mathbb{E}[X]\right] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 .$$

On vérifie le 2) en utilisant la linéarité de l'espérance :

$$\begin{aligned}
\mathrm{Var}(aX + b) &= \mathbb{E}\big[(aX + b)^2\big] - \mathbb{E}[aX + b]^2 = \mathbb{E}\big[(aX + b)^2\big] - \big(a\mathbb{E}[X] + b\big)^2 \\
&= \mathbb{E}\big[a^2X^2 + 2abX + b^2\big] - \big(a^2\mathbb{E}[X]^2 + 2ab\mathbb{E}[X] + b^2\big) \\
&= a^2\big(\mathbb{E}[X^2] - \mathbb{E}[X]^2\big) = a^2\mathrm{Var}(X) .
\end{aligned}$$

Pour démontrer le 3), on introduit la v.a. réelle $Y := X - \mathbb{E}[X]$. On a donc :

$$0 = \mathbb{E}[Y^2] = \sum_y y^2 \mathbb{P}(Y = y) .$$

La somme dans le membre de droite ne comporte que des termes positifs ou nuls. Comme la somme est nulle on doit avoir $\mathbb{P}(Y = y) = 0$ pour tout $y \neq 0$. Donc :

$$\mathbb{P}(X = \mathbb{E}[X]) = \mathbb{P}(Y = 0) = 1 - \sum_{y \neq 0} \mathbb{P}(Y = y) = 1 - 0 = 1\,,$$

ce qui donne le résultat. ■

Exemple.
Si X est une v.a. de Bernoulli de paramètre p, alors on a facilement $\mathbb{E}[X] = \mathbb{E}[X^2] = p$, et par (13) :

$$\mathrm{Var}(X) = p(1-p)\,.$$

En étudiant la fonction $p \mapsto p(1-p)$, on trouve que la dispersion maximale correspond à $p = 1/2$, elle est nulle pour $p = 0$ ou 1.

Proposition 9 (Inégalité de Bienaymé-Tchebychef).
Pour toute v.a. réelle X, on a pour tout $a > 0$:

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geqslant a\right) \leqslant \frac{1}{a^2}\mathrm{Var}(X)\,. \tag{15}$$

Démonstration.
En examinant séparément les cas $|y| \geqslant a$ et $|y| < a$, on trouve que, pour tout $y \in \mathbb{R}$, on a :

$$\mathbf{1}_{[a,\infty[}(|y|) \leqslant \frac{1}{a^2}y^2\,.$$

Par conséquent, si Y est une v.a. réelle, alors :

$$\mathbf{1}_{\{|Y| \geqslant a\}} \leqslant \frac{1}{a^2}Y^2\,.$$

En prenant l'espérance et en utilisant (8) (page 976), on trouve :

$$\mathbb{P}(|Y| \geqslant a) \leqslant \frac{1}{a^2}\mathbb{E}[Y^2]\,.$$

L'inégalité cherchée s'obtient en prenant $Y = X - \mathbb{E}[X]$. ■

Cette proposition montre que la variance mesure effectivement la dispersion d'une v.a. réelle autour de sa moyenne. En effet, en appliquant l'inégalité de Bienaymé-Tchebychef avec $a = 3{,}16\sigma$ par exemple, on trouve qu'il est improbable (probabilité inférieure à 10%) qu'une v.a. X dévie de $\mathbb{E}[X]$ de plus de $3{,}16$ fois son écart-type $\sigma = \sqrt{\mathrm{Var}(X)}$. L'étude des (grandes) déviations d'une v.a. est une branche importante de la théorie des probabilités, car elle concerne les probabilités d'événements rares tels que des pannes de systèmes ou des accidents. Il existe en fait des théorèmes donnant des inégalités plus fines que l'inégalité de Bienaymé-Tchebychef, mais qui dépassent le cadre de ce cours.

3.6 Espérances et variances pour des lois usuelles

Proposition 10. Si X suit la loi binomiale $\mathcal{B}(n,p)$, alors :

$$\mathbb{E}[X] = np\,, \qquad \mathrm{Var}(X) = np(1-p)\,.$$

Démonstration. On verra une démonstration très rapide de ce résultat dans la suite. On donne ici une preuve complète calculatoire. Par définition de l'espérance et de la loi binomiale :

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \times \binom{n}{k} p^k (1-p)^{n-k} = (1-p)^n \left[\sum_{k=0}^{n} k \binom{n}{k} x^k\right]_{x=p/(1-p)}\,. \tag{16}$$

Si on dérive par rapport à x l'identité du binôme :

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k\,,$$

alors on obtient après avoir multiplié par x :

$$nx(1+x)^{n-1} = \sum_{k=0}^{n} k \binom{n}{k} x^k\,. \tag{17}$$

En substituant dans l'expression (16) de $\mathbb{E}[X]$ cette égalité évaluée en $x = p/(1-p)$, on trouve :

$$\mathbb{E}[X] = (1-p)^n \left[nx(1+x)^{n-1}\right]_{x=p/(1-p)} = np\,.$$

Un argument similaire permet de calculer $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = (1-p)^n \left[\sum_{k=0}^{n} k^2 \binom{n}{k} x^k\right]_{x=p/(1-p)}\,.$$

On dérive (17) et on multiplie par x à nouveau :

$$n(1+x)^{n-2}(x+nx^2) = \sum_{k=0}^{n} k^2 \binom{n}{k} x^k\,,$$

ce qui donne :

$$\mathbb{E}[X^2] = (1-p)^n \left[n(1+x)^{n-2}(x+nx^2)\right]_{x=p/(1-p)} = np + n(n-1)p^2\,,$$

et finalement :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = np + n(n-1)p^2 - n^2p^2 = np(1-p)\,.$$

■

Proposition 11. Si X suit la loi hypergéométrique de paramètres (m_a, m_b, n), alors :

$$\mathbb{E}[X] = \frac{m_a n}{m_a + m_b}, \qquad \mathrm{Var}(X) = \frac{n(m_a + m_b - n)m_a m_b}{(m_a + m_b - 1)(m_a + m_b)^2}.$$

Démonstration. Par définition de l'espérance et de la loi hypergéométrique :

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} k \frac{\binom{m_a}{k}\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} = \sum_{k=0}^{n} k \frac{m_a!}{k!(m_a-k)!} \frac{\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} \\
&= \sum_{k=1}^{n} \frac{m_a!}{(k-1)!(m_a-k)!} \frac{\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} = \sum_{j=0}^{n-1} \frac{m_a!}{j!(m_a-1-j)!} \frac{\binom{m_b}{n-1-j}}{\binom{m_a+m_b}{n}} \\
&= m_a \sum_{j=0}^{n-1} \frac{\binom{m_a-1}{j}\binom{m_b}{n-1-j}}{\binom{m_a+m_b}{n}}.
\end{aligned}$$

Or la somme des probabilités élémentaires de la loi hypergéométrique de paramètres $(m_a - 1, m_b, n-1)$ doit être égale à 1 :

$$\sum_{j=0}^{n-1} \frac{\binom{m_a-1}{j}\binom{m_b}{n-1-j}}{\binom{m_a+m_b-1}{n-1}} = 1,$$

ce qui implique que $\mathbb{E}[X] = m_a \dfrac{\binom{m_a+m_b-1}{n-1}}{\binom{m_a+m_b}{n}} = \dfrac{m_a n}{m_a + m_b}$.

Le calcul de la variance s'effectue de manière similaire (laissé en exercice). ■

Le loto. On considère la forme classique du loto, qui a été utilisée de 1976 à 2008. Le joueur coche 6 numéros sur une grille en comportant 49, dans le cas d'un bulletin simple. Les 6 numéros gagnants sont déterminés par tirage au sort. L'espace fondamental est ici l'ensemble des parties à 6 éléments (« tirage ») de $[\![1, 49]\!]$ muni de la probabilité uniforme. Son cardinal est $\mathrm{card}\,(\Omega) = \binom{49}{6} = 13\,983\,816$. Notant N le nombre de numéros gagnants figurant parmi les numéros cochés par le joueur (on rappelle que la grille du joueur comporte 6 numéros cochés et 43 non-cochés), l'événement $\{N = n\}$ ($n \in [\![0, 6]\!]$) est réalisé si le tirage produit n numéros cochés et $6 - n$ numéros non-cochés. La v.a. N a donc pour loi la loi hypergéométrique de paramètres $(6, 43, 6)$ (voir page 975) :

$$\mathbb{P}(N = n) = \frac{\binom{6}{n}\binom{43}{6-n}}{\binom{49}{6}}, \quad n \in [\![0, 6]\!],$$

et certaines valeurs numériques sont données dans le tableau ci-dessous. Au premier tirage du 10 mai 2006, on recevait pour une mise de 0,3 euros (soit une grille) le gain $G = g(N)$ suivant :

n numéros gagnants	gain $g(n)$	probabilité $\mathbb{P}(N = n)$
6	789 177,00 euros	$7{,}2\ 10^{-8}$
5	1 813,80 euros	$1{,}8\ 10^{-5}$
4	30,70 euros	$9{,}7\ 10^{-4}$
3	2,70 euros	$1{,}8\ 10^{-2}$

Le gain moyen est $\mathbb{E}[G] = 2{,}70 \times 1{,}8\ 10^{-2} + 30{,}70 \times 9{,}7\ 10^{-4} + \cdots \simeq 0{,}168$ euros, tandis que l'écart-type $\sigma(G) = (\mathbb{E}[G^2] - \mathbb{E}[G]^2)^{1/2} \simeq 212$ euros. On voit que le jeu est défavorable au joueur, dont le bénéfice moyen est $\mathbb{E}[G] - 0{,}3 = -0{,}132$ euros, et la grande valeur de σ vient de ce que parfois (mais très rarement) ça rapporte gros.

Poursuivons la discussion engagée page 973 au sujet des deux méthodes possibles d'échantillonnage (avec ou sans remise). Si l'on tire n individus au hasard dans une population se divisant en une classe de m_a individus de type A et m_b individus de type B, alors le nombre d'individus tirés dans la classe A suit une loi binomiale de paramètres (n, p), avec $p = m_a/(m_a + m_b)$, dans le cas d'un tirage avec remise, et une loi hypergéométrique de paramètres (m_a, m_b, n) dans le cas sans remise (sous réserve que $n \leqslant \min(m_a, m_b)$). Dans les deux cas, l'espérance de la variable aléatoire est $nm_a/(m_a + m_b)$, comme le montrent les propositions précédentes. De plus, si m_a et m_b sont nettement plus grands que n, alors la variance de la variable aléatoire hypergéométrique vérifie :

$$\mathrm{Var}(X) = \frac{n(m_a + m_b - n)m_a m_b}{(m_a + m_b - 1)(m_a + m_b)^2} \simeq n\frac{m_a m_b}{(m_a + m_b)^2} = np(1-p)\,,$$

c'est-à-dire qu'elle est à peu près égale à la variance de la variable aléatoire binomiale du cas avec remise. La comparaison entre loi binomiale et loi hypergéométrique (lorsque $m_a, m_b \gg n$) peut être menée encore plus loin, mais cela nécessite d'introduire la notion de convergence en loi qu'on ne verra pas dans le cours L1.

4 Familles de variables aléatoires

Dans la plupart des expériences aléatoires, il apparaît naturellement non pas une seule variable aléatoire, mais plusieurs.

Définition 6. Soit $n \geqslant 2$. Si $X_1, \ldots, X_n$ sont des v.a. définies sur le même espace de probabilité fini $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, alors $(X_1, \ldots, X_n)$ est appelé vecteur aléatoire, et si $n = 2$, couple aléatoire.

4.1 Loi d'un vecteur aléatoire

Proposition 12. Soient $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini et $n \geqslant 2$.
Soit $(X_i)_{i \in [\![1,n]\!]}$ une suite de v.a. définies sur $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, avec X_i à valeurs dans E_i pour tout $i \in [\![1, n]\!]$.
Le vecteur aléatoire $X = (X_1, \ldots, X_n)$ est une v.a. au sens de la définition 1, à valeurs dans l'espace $E = E_1 \times \cdots \times E_n$. Sa loi est donnée par les nombres :

$$p_X(x) = \mathbb{P}(X = x), \quad x = (x_1, \ldots, x_n) \in E\,,$$

$$\text{où} \qquad \{X = x\} = \{X_1 = x_1, \ldots, X_n = x_n\} = \bigcap_{i=1}^{n} \{X_i = x_i\}\,.$$

On appelle souvent la loi du vecteur X, loi jointe, pour la distinguer des lois p_{X_i} des composantes X_i ($i \in [\![1, n]\!]$) que l'on appelle lois marginales. On obtient ces lois marginales en sommant la loi jointe sur les v.a. que l'on veut éliminer.

Proposition 13 (Formule des marginales). Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire. La loi de X_1 se retrouve à partir de la loi de X par la formule :

$$p_{X_1}(x_1) = \sum_{x_2 \in E_2, \ldots, x_n \in E_n} p_X(x_1, x_2, \ldots, x_n)\,. \tag{18}$$

Démonstration. Il suffit de noter que :

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \mathbb{P}\Bigg(\bigcup_{(x_2,\ldots,x_n) \in E_2 \times \cdots \times E_n} \{X = (x_1, x_2, \ldots, x_n)\}\Bigg)\,,$$

et d'appliquer ensuite l'additivité de la probabilité $\mathbb{P}$. ∎

On obtient de la même manière la loi d'un sous-vecteur de X. Par exemple la loi de (X_1, X_2) est :

$$p_{X_1,X_2}(x_1, x_2) = \sum_{x_3 \in E_3, \ldots, x_n \in E_n} p_X(x_1, x_2, x_3, \ldots, x_n)\,.$$

Pour toute fonction $f : E \to \mathbb{R}$, l'espérance de la v.a. réelle $Y = f(X)$ est encore définie par (7), $\mathbb{E}[f(X)] = \sum_y y p_Y(y)$, la formule (9) s'applique encore, et on calculera $\mathbb{E}[f(X)]$ à l'aide de (10), c'est-à-dire ici :

$$\mathbb{E}[f(X_1, \ldots, X_n)] = \sum_{x_1 \in E_1, \ldots, x_n \in E_n} f(x_1, \ldots, x_n) p_X(x_1, \ldots, x_n)\,.$$

La proposition 5 de la page 977 donnant les propriétés de $\mathbb{E}$ s'applique toujours.

Loi multinomiale. Dans une population de taille m divisée en 3 groupes de tailles respectives m_1, m_2, m_3 ($m_1 + m_2 + m_3 = m$), on choisit successivement n individus au hasard (tirage avec remise : on peut choisir plusieurs fois le même individu). On s'intéresse au triplet (N_1, N_2, N_3) avec $N_i :=$ nombre d'individus du $i^{ème}$ groupe tirés, $i = 1, 2, 3$. Les variables aléatoires N_i sont liées par la relation $N_1 + N_2 + N_3 = n$. Le résultat de chaque tirage avec remise (de n individus) est représenté par une application :

$$\omega : [\![1, n]\!] \to [\![1, 3]\!]\,,$$

avec $\omega(j)$ le numéro du groupe où est tiré le $j^{ème}$ individu. D'après l'équiprobabilité :

$$\mathbb{P}(\{\omega\}) = \frac{m_1^{n_1} m_2^{n_2} m_3^{n_3}}{m^n} = p_1^{n_1} p_2^{n_2} p_3^{n_3}\,,$$

avec $n_i = N_i(\omega)$ et $p_i = m_i/m$, $i \in [\![1, 3]\!]$. Il faut ensuite faire un peu de combinatoire pour calculer, pour tout triplet (n_1, n_2, n_3) d'entiers de somme égale à n, le nombre C_{n_1,n_2,n_3} d'applications ω telles que $N_i(\omega) = n_i$ $\forall i \in [\![1, 3]\!]$. On choisit d'abord le sous-ensemble $\omega^{-1}(\{1\})$ de $[\![1, n]\!]$ des indices des tirages dans le groupe 1 (soit $\binom{n}{n_1}$ possibilités), puis le sous-ensemble $\omega^{-1}(\{2\})$ dans $[\![1, n]\!] \backslash \omega^{-1}(\{1\})$ des indices des tirages dans le groupe 2 (soit $\binom{n-n_1}{n_2}$ possibilités). Les autres tirages sont alors forcément dans le groupe 3.

Donc le nombre C_{n_1,n_2,n_3} est donné par le coefficient multinomial $n!/(n_1!n_2!n_3!)$ et :

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \frac{n!}{n_1!n_2!n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}\,, \tag{19}$$

pour n_1, n_2, n_3 entiers tels que $n_1 + n_2 + n_3 = n$. Cette loi s'appelle la loi multinomiale de paramètres (n, p_1, p_2, p_3).

Exercice 1.

On reprend l'exemple précédent (loi multinomiale).

1. Généraliser la formule (19) de la loi multinomiale, au cas où la population est divisée en d classes.
2. Vérifier que les lois marginales des N_i sont binomiales.
3. Vérifier que, pour $d = 2$, on a $(N_1, N_2) = (S, n - S)$ où S suit une loi binomiale de paramètres (n, p) avec $p = p_1$ (*cf.* page 970).

Solution.

1. Dans une population de taille m divisée en d groupes de tailles respectives $m_1, \ldots, m_d$ $(m_1 + \cdots + m_d = m)$, on choisit successivement n individus au hasard (avec remise). On s'intéresse au d-uplet $(N_1, \ldots, N_d)$ avec $N_i :=$ nombre d'individus du $i^{ème}$ groupe tirés, $i = 1, \ldots, d$. En notant $p_i = m_i/m$, et en refaisant exactement la même raisonnement que dans le cas $d = 3$, on trouve :
$$\mathbb{P}(N_1 = n_1, \ldots, N_d = n_d) = \frac{n!}{n_1! \cdots n_d!} p_1^{n_1} \cdots p_d^{n_d}\,,$$
pour $n_1, \ldots, n_d$ entiers tels que $n_1 + \cdots + n_d = n$. C'est la loi multinomiale de paramètres $(n, p_1, \ldots, p_d)$.
2. On utilise la formule des marginales : Pour tout $n_1 \in [\![0, n]\!]$,
$$\begin{aligned}\mathbb{P}(N_1 = n_1) &= \sum_{n_2,\ldots,n_d} \mathbb{P}(N_1 = n_1, N_2 = n_2, \ldots, N_d = n_d)\\ &= \frac{n!}{n_1!} p_1^{n_1} \sum_{n_2,\ldots,n_d\,,\ \sum_{j=2}^{d} n_j = n - n_1} \frac{1}{n_2! \cdots n_d!} p_2^{n_2} \cdots p_d^{n_d}\\ &= \binom{n}{n_1} p_1^{n_1} (1 - p_1)^{n-n_1} \sum_{n_2,\ldots,n_d\,,\ \sum_{j=2}^{d} n_j = n - n_1} \frac{(n - n_1)!}{n_2! \cdots n_d!} {p'_2}^{n_2} \cdots {p'_d}^{n_d}\,,\end{aligned}$$
avec $p'_j = p_j/(1 - p_1)$, $j \in [\![2, d]\!]$, qui sont tels que $\sum_{j=2}^{d} p'_j = 1$.

   La somme sur $n_2, \ldots, n_d$ est égale à 1 (car on reconnaît les termes d'une loi multinomiale de paramètres $(n - n_1, p'_2, \ldots, p'_d)$) et l'on obtient :
$$\mathbb{P}(N_1 = n_1) = \binom{n}{n_1} p_1^{n_1} (1 - p_1)^{n-n_1}\,,$$
ce qui montre que $N_1 \sim \mathcal{B}(n, p_1)$.
3. C'est immédiat : on a $N_2 = n - N_1$ et $N_1 \sim \mathcal{B}(n, p_1)$.

4.2 Covariance et corrélation

Dans le cas d'un couple de v.a., les notions d'espérance et de variance jouent un rôle fondamental pour caractériser les coordonnées séparément. Afin de les étudier simultanément, on introduit la notion de covariance et de corrélation de deux v.a..

Définition 7. Soient X et Y deux v.a. réelles définies sur le même espace de probabilité fini. On appelle covariance de X et de Y, le nombre :

$$\mathrm{Cov}(X,Y) := \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] . \tag{20}$$

Si $\mathrm{Var}(X) > 0$ et $\mathrm{Var}(Y) > 0$, le coefficient de corrélation $\mathrm{Cor}(X,Y)$ est le nombre défini par :

$$\mathrm{Cov}(X,Y) = \mathrm{Cor}(X,Y)\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} . \tag{21}$$

Notez qu'en développant le membre de droite de (20) on obtient que la covariance peut aussi s'écrire :

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[XY\right] - \mathbb{E}[X]\mathbb{E}[Y] .$$

Ces quantités décrivent les variations simultanées des v.a. X et Y. Si X et Y « varient dans le même sens », leur covariance et donc leur corrélation est positive. Par exemple, $\mathrm{Cov}(X,X) = \mathrm{Var}(X) \geqslant 0$, tandis que $\mathrm{Cov}(X,-X) = -\mathrm{Var}(X) \leqslant 0$.

Proposition 14 (Inégalité de Cauchy-Schwarz). Soient X et Y deux v.a. réelles définies sur le même espace de probabilité fini. Alors XY est une v.a. réelle et l'on a :

$$\mathbb{E}[XY]^2 \leqslant \mathbb{E}[X^2]\mathbb{E}[Y^2] .$$

Démonstration. Si $\mathbb{E}[Y^2] = 0$, cela implique que $Y = 0$ avec probabilité un, et donc $XY = 0$ avec probabilité 1 et l'inégalité de la proposition est triviale. Supposons que $\mathbb{E}[Y^2] > 0$. Par positivité et linéarité de l'espérance, on a pour tout $\lambda \in \mathbb{R}$:

$$0 \leqslant \mathbb{E}[(X-\lambda Y)^2] = \mathbb{E}[X^2] - 2\lambda\mathbb{E}[XY] + \lambda^2\mathbb{E}[Y^2] .$$

En prenant $\lambda = \mathbb{E}[XY]/\mathbb{E}[Y^2]$, on obtient :

$$0 \leqslant \mathbb{E}[X^2] - \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} ,$$

ce qui donne l'inégalité recherchée. ∎

L'inégalité de Cauchy-Schwarz montre que la corrélation est un nombre réel compris entre -1 et 1.

Corollaire 15. Soient X et Y deux v.a. réelles définies sur le même espace de probabilité fini telles que $\mathrm{Var}(X) > 0$ et $\mathrm{Var}(Y) > 0$. Alors $\mathrm{Cor}(X,Y) \in [-1,1]$.

Démonstration.
Il suffit d'appliquer l'inégalité de Cauchy-Schwarz aux variables centrées $X' = X - \mathbb{E}[X]$ et $Y' = Y - \mathbb{E}[Y]$. ∎

La variance, à la différence de l'espérance, n'est pas additive.

Cependant l'importance de la notion de covariance vient de la relation :

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)\,,$$

dont la forme générale est donnée dans la proposition suivante :

Proposition 16. La variance de la somme $S_n = \sum_{i=1}^{n} X_i$ de n v.a. réelles définies sur le même espace de probabilité fini est donnée par la formule :

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \leqslant l < m \leqslant n} \mathrm{Cov}(X_l, X_m)\,. \tag{22}$$

En termes d'algèbre, la variance définit une forme quadratique sur l'espace vectoriel des v.a. réelles définies sur un même espace de probabilité fini. La covariance est la forme bilinéaire symétrique associée à cette forme quadratique.

Démonstration. Puisque $\mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[X_i]$, on a :

$$S_n - \mathbb{E}[S_n] = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\,,$$

et donc

$$(S_n - \mathbb{E}[S_n])^2 = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])^2 + 2 \sum_{1 \leqslant l < m \leqslant n} (X_l - \mathbb{E}[X_l])(X_m - \mathbb{E}[X_m])\,.$$

La formule de la proposition s'obtient alors en prenant l'espérance des deux membres. ∎

4.3 Indépendance

Définition 8. Soit $(X_i)_{i \in [\![1,n]\!]}$ une suite de v.a. définies sur le même espace de probabilité fini à valeurs dans des espaces $(E_i)_{i \in [\![1,n]\!]}$. On dit que les v.a. sont indépendantes si, pour tout $x_1 \in E_1, \ldots, x_n \in E_n$, on a :

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right)\,. \tag{23}$$

Exemple. Dans le cas du jeu de Pile ou Face de durée n (décrit page 970), nous avons défini $\Omega = \{F, P\}^n$ et $\mathbb{P}(\{\omega\}) = p^k(1-p)^{n-k}$ pour tout résultat ω comportant exactement k Piles. Pour $i \in [\![1,n]\!]$, on note $X_i : \Omega \to \{0,1\}$ la v.a. de Bernoulli :

$$X_i = \mathbf{1}_{A_i}\,,$$

où $A_i = \{\omega \in \Omega \mid \omega(i) = P\}$ est l'événement « le $i^{\text{ème}}$ lancer est un Pile ».

Soient $x_i \in \{0,1\}$, $i \in [\![1,n]\!]$. On se rend compte que la conjonction d'événements :

$$\{X_1 = x_1, \ldots, X_n = x_n\} = \bigcap_{i=1}^{n} \{X_i = x_i\}$$

n'est réalisée que pour un unique $\tilde{\omega} \in \Omega$ qui est tel que la $i^{ème}$ coordonnée $\tilde{\omega}(i)$ est P si $x_i = 1$ et F sinon. Donc :

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(\{\tilde{\omega}\}) = p^{\bar{x}}(1-p)^{n-\bar{x}} ,$$

où $\bar{x} = \sum_{i=1}^{n} x_i$. D'autre part, pour tout $i \in [\![1, n]\!]$:

$$\mathbb{P}(X_i = x_i) = p^{x_i}(1-p)^{1-x_i} .$$

On en déduit que les X_i sont indépendantes. Il s'agit donc d'une suite de v.a. indépendantes et identiquement distribuées selon la loi de Bernoulli de paramètre p.

Remarque. Si $(X_i)_{i\in[\![1,n]\!]}$ est une suite de v.a. indépendantes, alors pour toute sous-famille d'indices $I \subset [\![1, n]\!]$, $(X_i)_{i\in I}$ est une suite de v.a. indépendantes. Il suffit d'appliquer la formule des marginales.

Proposition 17. Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite de v.a. définies sur le même espace de probabilité fini. Les v.a. sont indépendantes si, et seulement si, pour toutes fonctions à valeurs réelles $f_1, \ldots, f_n$, on a :

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}\left[f_i(X_i)\right] . \tag{24}$$

Démonstration. On note E_i l'espace des valeurs prises par X_i, pour $i \in [\![1, n]\!]$.
Montrons le sens direct. Soient f_i, $i \in [\![1, n]\!]$, des fonctions de E_i dans $\mathbb{R}$. D'après la définition de l'espérance :

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \sum_{x_1\in E_1,\ldots,x_n\in E_n} \Big(\prod_{i=1}^{n} f_i(x_i)\Big)\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) .$$

Comme les v.a. sont indépendantes :

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \sum_{x_1\in E_1,\ldots,x_n\in E_n} \prod_{i=1}^{n} \big(f_i(x_i)\mathbb{P}(X_i = x_i)\big) ,$$

et en factorisant la somme :

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \prod_{i=1}^{n} \sum_{x_i\in E_i} f_i(x_i)\mathbb{P}(X_i = x_i) = \prod_{i=1}^{n} \mathbb{E}\left[f_i(X_i)\right] ,$$

ce qui montre le résultat.
Montrons la réciproque. Soient $x_i \in E_i$, $i \in [\![1, n]\!]$. On applique (24) avec les fonctions indicatrices $f_i = \mathbf{1}_{\{x_i\}}$:

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) = \mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}\left[f_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right) ,$$

ce qui donne le résultat. ∎

4.4 Indépendance et covariance

S'il est facile de montrer que deux v.a. indépendantes ont une covariance nulle, il faut cependant résister à la tentation et ne pas croire en la réciproque : deux v.a. ayant une covariance nulle ne sont pas forcément indépendantes !

Proposition 18. *1.* Si deux v.a. réelles définies sur le même espace de probabilité fini sont indépendantes, alors leur covariance est nulle.
2. La réciproque est fausse en général.
3. Deux v.a. de Bernoulli définies sur le même espace de probabilité fini sont indépendantes si, et seulement si, leur covariance est nulle.

Démonstration. *1.* Soient X et Y deux v.a. indépendantes. Alors :

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0\,.$$

2. Il existe des v.a. X et Y qui ne sont pas indépendantes, et dont la covariance est nulle. En effet, considérons par exemple l'espace $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ avec $\Omega = \{-1,0,1\}$ et $\mathbb{P}(\{-1\}) = 1/4$, $\mathbb{P}(\{0\}) = 1/2$ et $\mathbb{P}(\{1\}) = 1/4$. La v.a. $X(\omega) = \omega$ est une v.a. réelle à valeurs dans $\{-1,0,1\}$ telle que :

$$\mathbb{P}(X=-1) = \frac{1}{4}\,, \quad \mathbb{P}(X=0) = \frac{1}{2}\,, \quad \mathbb{P}(X=1) = \frac{1}{4}\,\cdot$$

Introduisons la v.a réelle $Y = |X|$, qui est à valeurs dans $\{0,1\}$. On trouve successivement :

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{4}\times(-1) + \frac{1}{2}\times 0 + \frac{1}{4}\times 1 = 0\,,\\
\mathbb{E}[Y] &= \frac{1}{4}\times 1 + \frac{1}{2}\times 0 + \frac{1}{4}\times 1 = \frac{1}{2}\,,\\
\mathbb{E}[XY] &= \frac{1}{4}\times(-1)\times 1 + \frac{1}{2}\times 0\times 0 + \frac{1}{4}\times 1\times 1 = 0\,,
\end{aligned}$$

et donc $\mathrm{Cov}(X,Y) = 0$. Et pourtant X et Y ne sont pas indépendantes :

$$\mathbb{P}(X=1, Y=0) = 0 \neq \frac{1}{8} = \frac{1}{4}\times\frac{1}{2} = \mathbb{P}(X=1)\mathbb{P}(Y=0)\,.$$

3. Le sens direct est une conséquence du premier point. Montrons la réciproque. Soit un espace de probabilité fini $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, sur lequel sont définies deux v.a. de Bernoulli X et Y telle que $\mathrm{Cov}(X,Y) = 0$. Notons A l'événement $\{X=1\}$ et B l'événement $\{Y=1\}$. On a $X = \mathbf{1}_A$ et $Y = \mathbf{1}_B$, donc :

$$0 = \mathrm{Cov}(X,Y) = \mathbb{E}[\mathbf{1}_A\mathbf{1}_B] - \mathbb{E}[\mathbf{1}_A]\mathbb{E}[\mathbf{1}_B] = \mathbb{E}[\mathbf{1}_{A\cap B}] - \mathbb{E}[\mathbf{1}_A]\mathbb{E}[\mathbf{1}_B]\,,$$

la dernière égalité venant du fait que $\mathbf{1}_A(\omega)\mathbf{1}_B(\omega) = 1$ si, et seulement si, $\omega \in A$ et $\omega \in B$, soit $\omega \in A\cap B$. Comme $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$, on obtient donc que :

$$0 = \mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\,,$$

ce qui montre que les événements A et B sont indépendants, et on peut donc écrire :

$$\mathbb{P}(\{X=1\}\cap\{Y=1\}) = \mathbb{P}(X=1)\mathbb{P}(Y=1)\,.$$

Par la proposition 16 de la page 963, les couples d'événements (A, B^c), (A^c, B) et (A^c, B^c) sont indépendants, donc, en tenant compte du fait que $A^c = \{X=0\}$

et $B^c = \{Y = 0\}$, on obtient que pour tous $i, j \in \{0, 1\}$:

$$\mathbb{P}(\{X = i\} \cap \{Y = j\}) = \mathbb{P}(X = i)\mathbb{P}(Y = j)\,,$$

ce qui démontre l'indépendance de X et Y. ■

Le troisième point de la proposition montre qu'une covariance nulle pour deux v.a. de Bernoulli assure l'indépendance, mais une telle affirmation est spécifique aux v.a. de Bernoulli (ou aux variables aléatoires qui ne prennent chacune que deux valeurs) et ne peut pas être généralisée, comme le dit le deuxième point de la proposition et le contre-exemple associé (dans la démonstration).

Corollaire 19. Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite de v.a. réelles définies sur le même espace de probabilité fini. Si les X_i sont indépendantes deux à deux (c'est-à-dire que X_i et X_j sont indépendantes pour tous $i \neq j$), alors on a $\mathrm{Cov}(X_i, X_j) = 0$ pour tous $i \neq j$ et :

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)\,.$$

Attention : la réciproque est fausse !

Démonstration. Il suffit d'appliquer la formule (22) en tenant compte du *1.* de la proposition 18 pour obtenir le résultat. Pour se convaincre que la réciproque est fausse, on peut reprendre l'exemple fourni dans la preuve du *2.* de la proposition 18 : sur cet exemple, on a $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, et pourtant X et Y ne sont pas indépendantes. ■

Corollaire 20. Soit $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ un espace de probabilité fini sur lequel est définie une suite de n v.a. de Bernoulli indépendantes de même paramètre $p \in [0, 1]$, notées X_i, pour $i \in [\![1, n]\!]$. On note S la somme des X_i :

$$S = \sum_{i=1}^{n} X_i\,.$$

Alors S suit la loi binomiale de paramètres (n, p).

Démonstration. En introduisant les événements indépendants $A_i := \{X_i = 1\}$, on se retrouve exactement dans la situation de la proposition 3 de la page 971, ce qui donne immédiatement le résultat. ■

Application : on peut retrouver facilement les expressions de l'espérance et de la variance d'une v.a. binomiale $\mathcal{B}(n, p)$ données dans la proposition 10 de la page 981. En effet, en reprenant les notations du corollaire précédent et en utilisant la linéarité de l'espérance, on obtient :

$$\mathbb{E}[S] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\mathbb{E}[X_1] = np\,.$$

De plus, comme les v.a. X_i sont indépendantes, on a aussi :

$$\mathrm{Var}(S) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\mathrm{Var}(X_1) = np(1 - p)\,.$$

4.5 Loi faible des grands nombres

Cette section a pour but la démonstration d'un résultat intuitif qui stipule que, lorsqu'on répète un grand nombre de fois une expérience aléatoire dans des conditions identiques et indépendantes, les quantités obtenues par moyennisation sur les résultats de toutes les expériences se stabilisent autour d'une valeur moyenne déterministe. Ce résultat a d'ailleurs été utilisé bien avant sa démonstration complète, et c'est pour cela qu'on l'a appelé « loi », à l'instar d'autres lois de la physique.

Définition 9. Soient $n \in \mathbb{N}^*$ et $(X_i)_{i\in[\![1,n]\!]}$ une suite de v.a. réelles définies sur le même espace de probabilité fini, indépendantes et de même loi. La moyenne empirique de $(X_i)_{i\in[\![1,n]\!]}$ est la v.a. réelle :

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i\,. \tag{25}$$

La moyenne empirique $\bar{X}_n$ a pour espérance :

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu\,, \tag{26}$$

avec $\mu = \mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n]$. Ici, on a simplement utilisé la linéarité de l'espérance. La variance de la moyenne empirique est :

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\,(X_i) = \frac{\sigma^2}{n}\,, \tag{27}$$

avec $\sigma^2 = \mathrm{Var}(X_1) = \cdots = \mathrm{Var}(X_n)$. Ici, on a utilisé le fait que la variance d'une somme de v.a. indépendantes est la somme des variances. La quantité $\mathrm{Var}(\bar{X}_n)$ tendant vers 0 lorsque $n \to +\infty$, la variable $\bar{X}_n$ se concentre de plus en plus autour de la constante μ. Voilà le théorème limite le plus simple des probabilités. Le hasard se manifeste par des fluctuations. Mais si l'on répète l'expérience, de manière identique et indépendante, alors la moyenne empirique fluctue de moins en moins et converge vers l'espérance mathématique. L'inégalité de Bienaymé-Tchebychef (page 980) montre que, pour tout $\varepsilon > 0$:

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}\,, \tag{28}$$

la probabilité de déviation tend vers 0 lorsque $n \to +\infty$. Nous avons montré la loi faible des grands nombres.

Exemple. Considérons une suite $(X_i)_{i\in[\![1,n]\!]}$ de v.a. indépendantes et identiquement distribuées selon la loi de Bernoulli de paramètre p. L'événement $X_i = 1$ représente un « succès » lors de la $i^{ème}$ expérience. La probabilité de succès est p. La somme $\sum_{j=1}^{n} X_j$ suit une loi binomiale de paramètres (n, p) et représente le nombre de succès lors des n premières expériences. La moyenne empirique $\bar{X}_n$ représente la proportion de succès, elle a pour moyenne p et variance $p(1-p)/n$. La loi (faible) des grands nombres affirme que $\bar{X}_n$ est proche de p lorsque n est grand.

Exercice type corrigé

V.3.0 Soit X une variable aléatoire ne pouvant prendre que les valeurs 3, 4, 5 et 6.

1. Déterminer la loi de X sachant que :
$$\mathbb{P}(X < 5) = \frac{1}{6}, \quad \mathbb{P}(X > 5) = \frac{1}{2}, \quad \mathbb{P}(X \leqslant 3) = \mathbb{P}(X = 4).$$
2. Calculer l'espérance $\mathbb{E}[X]$.

Solution.

1. Il faut déterminer $\mathbb{P}(X = j)$ pour $j = 3, 4, 5, 6$.
Comme X ne prend pas de valeurs au dessus de 6, on peut dire que $\{X = 6\} = \{X > 5\}$, et donc :
$$\mathbb{P}(X = 6) = \frac{1}{2}\cdot$$
Comme $\{X < 5\}$ et $\{X \geqslant 5\}$ sont complémentaires on a :
$$\mathbb{P}(X \geqslant 5) = 1 - \mathbb{P}(X < 5) = 1 - \frac{1}{6} = \frac{5}{6}\cdot$$
Or $\{X \geqslant 5\} = \{X = 5\} \cup \{X = 6\}$ avec union disjointe, donc
$$\mathbb{P}(X = 5) = \mathbb{P}(X \geqslant 5) - \mathbb{P}(X = 6) = \frac{5}{6} - \frac{1}{2} = \frac{1}{3}\cdot$$
Comme X ne prend pas de valeur en dessous de 3, on peut dire que $\{X \leqslant 3\} = \{X = 3\}$, donc $\mathbb{P}(X = 3) = \mathbb{P}(X = 4)$. Or $\{X < 5\} = \{X = 3\} \cup \{X = 4\}$ avec union disjointe, donc :
$$\frac{1}{6} = \mathbb{P}(X < 5) = \mathbb{P}(X = 3) + \mathbb{P}(X = 4) = 2\mathbb{P}(X = 3),$$
ce qui donne :
$$\mathbb{P}(X = 3) = \mathbb{P}(X = 4) = \frac{1}{12}\cdot$$
2. On applique la définition de l'espérance :
$$\mathbb{E}[X] = \sum_{j=3}^{6} j\mathbb{P}(X = j) = 3 \times \frac{1}{12} + 4 \times \frac{1}{12} + 5 \times \frac{1}{3} + 6 \times \frac{1}{2} = \frac{21}{4}\cdot$$

Pour aller plus loin

Les corrigés sont disponibles en accès libre à partir des pages d'accueil du livre sur le site www.dunod.com

V.3.1 On lance deux dés équilibrés. On note X le plus grand des numéros obtenus. Déterminer la loi de la variable aléatoire X.

V.3.2 Un concierge a n clefs dont une seule ouvre sa porte, mais il ne sait plus laquelle. Il les essaie l'une après l'autre en éliminant après chaque essai la clef qui n'a pas convenu. Trouver le nombre moyen d'essais nécessaires pour trouver la bonne clef.

V.3.3 (Problème des rencontres) On considère une urne contenant n boules numérotées de 1 à n. On les extrait successivement sans remise et l'on observe le numéro de la boule tirée. On dit qu'il

y a rencontre au $k^{ème}$ tirage si la boule de numéro k est tirée au $k^{ème}$ tirage. Calculer la probabilité p_n pour qu'il y ait au moins une rencontre. Que devient p_n lorsque $n \to +\infty$?

V.3.4 Le problème des boîtes d'allumettes de Banach

Un fumeur a dans chacune de ses poches droite et gauche une boîte d'allumettes (de contenance N chacune). Lorsqu'il désire une allumette il choisit au hasard une de ses poches (chacune avec probabilité $1/2$). On considère le moment où, pour la première fois, le fumeur s'aperçoit que l'une des boîtes d'allumettes est vide. À ce moment, l'autre boîte peut contenir un nombre quelconque d'allumettes. On désigne par u_r la probabilité qu'elle en contienne r.

1. Calculer u_r, $0 \leqslant r \leqslant N$. Que vaut $\sum_{r=0}^{N} u_r$?

2. Calculer la probabilité u que les deux boîtes soient vides lorsque le fumeur s'aperçoit que l'une des boîtes d'allumettes est vide. Déterminer une expression approchée de u lorsque N est grand en utilisant la formule de Stirling $n! \sim (n/e)^n \sqrt{2\pi n}$.

V.3.5 Un joueur tire sur une cible de 10 cm de rayon, constituée de couronnes concentriques, délimitées par des cercles de rayons $1, 2, \ldots, 10$ cm, et numérotées respectivement de 10 à 1. La probabilité d'atteindre la couronne k est proportionnelle à l'aire de cette couronne, et on suppose que le joueur atteint sa cible à chaque lancer. Soit X la variable aléatoire qui à chaque lancer associe le numéro de la cible.

1. Quelle est la loi de probabilité de X ?

2. Le joueur gagne k euros s'il atteint la couronne numérotée k pour k compris entre 6 et 10, tandis qu'il perd 2 euros s'il atteint l'une des couronnes périphériques numérotées de 1 à 5. Le jeu est-il favorable au joueur ? (i.e., calculer l'espérance du gain et déterminer si cette espérance est positive ou négative).

V.3.6 On considère un dé truqué dont les faces sont numérotées de 1 à 6 et on note X la variable aléatoire donnée par le numéro de la face du dessus. On suppose que le dé est truqué de sorte que la probabilité d'obtenir une face est proportionnelle au numéro inscrit sur cette face.

1. Déterminer la loi de X, calculer son espérance.

2. On pose $Y = 1/X$. Déterminer la loi de Y et son espérance.

V.3.7 Soit X une v.a. réelle. Montrer que :

$$\mathrm{Var}(X) = \inf_{c \in \mathbb{R}} \mathbb{E}\left[(X-c)^2\right] ,$$

et que cette borne inférieure est atteinte précisément pour $c := \mathbb{E}[X]$.

V.3.8 Soit X une v.a. à valeurs dans un sous-ensemble de $\mathbb{N}$. Montrer que :

$$\mathbb{E}[X] = \sum_{n \geqslant 0} \mathbb{P}(X > n)\,. \tag{29}$$

V.3.9 Un devin moyen

On considère l'expérience de divination suivante. On dispose d'un jeu de 52 cartes distinctes, d'un manipulateur et d'un devin. Le devin ne peut à aucun moment voir le jeu ou le manipulateur, et doit deviner quelle est la carte se trouvant sur le dessus du paquet. Il annonce donc une carte au hasard, et le manipulateur retourne sans les montrer les cartes les unes après les autres jusqu'à tomber sur la carte annoncée par le devin. Après quoi ce dernier doit deviner la carte qui suit. Il annonce une carte au hasard parmi les 51 restantes, et le manipulateur continue de retourner les cartes à partir de l'endroit où il s'était arrêté. Ainsi de suite jusqu'à ce que tout le paquet soit retourné.

1. Donner un espace de probabilité correspondant à cette expérience.
2. Combien de cartes en moyenne le devin trouvera-t-il ?

V.3.10 On s'intéresse à une maladie génétique. Elle est portée par un gène particulier qui existe sous deux formes : l'allèle A (sain) et l'allèle B (malade). Il existe donc pour chaque individu trois génotypes possibles : 1 (AA), 2 (AB) et 3 (BB). Un individu est malade lorsqu'il porte le génotype (BB). Le but de l'exercice est de démontrer que la proportion de malades est constante au cours du temps. Pour cela, on s'intéresse à une population dont la proportion du génotype i, à la génération n, est notée $u_i(n)$, pour $i \in \{1, 2, 3\}$. On rappelle que chaque enfant reçoit un des deux allèles de chacun de ses parents (et ce de façon complètement aléatoire). On suppose aussi que les procréations dans la population se font complètement aléatoirement. On fixe $n \geqslant 0$ et on note E le génotype d'un enfant de la $n+1$-ième génération, P et M les génotypes respectifs du père et de la mère.

1. Calculer les probabilités conditionnelles $\mathbb{P}(E = 1 | (P, M) = (i, j))$ pour tous $i, j \in [\![1, 3]\!]$.
2. En déduire la loi de E en fonction de $u_i(n)$, $i \in [\![1, 3]\!]$.
3. On pose $\theta(n) = u_1(n) + \frac{1}{2}u_2(n)$.
   Exprimer $u_i(n+1)$ en fonction de $\theta(n)$ pour tout $i \in [\![1, 3]\!]$.
4. Démontrer que la proportion de malades ne varie plus à partir de la génération 2.

V.3.11 Les vaches laitières sont atteintes par une maladie M avec la probabilité $p = 0{,}15$. Pour dépister la maladie M dans une étable de n vaches, on fait procéder à une analyse de lait. Deux méthodes sont possibles :
Première méthode : On fait une analyse sur un échantillon de lait de chaque vache.
Deuxième méthode : On effectue d'abord une analyse sur un échantillon de lait provenant du mélange des n vaches. Si le résultat est positif, on effectue une nouvelle analyse, cette fois pour chaque vache.
On voudrait connaître la méthode la plus économique (=celle qui nécessite en moyenne le moins d'analyses). Pour cela, on note X_n la variable aléatoire du nombre d'analyses réalisées dans la deuxième méthode. On pose $Y_n = X_n/n$. Comme n est le nombre d'analyses que nécessite la première méthode, on conclut que la seconde méthode est préférable si, et seulement si, $\mathbb{E}[Y_n] < 1$.

1. Déterminer la loi de Y_n, et montrer que son espérance vaut : $1 + \frac{1}{n} - 0{,}85^n$.
2. Etudier la fonction $f(x) = ax + \ln x$, pour $a = \ln(0{,}85)$. Donner la liste des entiers n tels que $f(n) > 0$.
3. Montrer que $f(n) > 0$ équivaut à $\mathbb{E}[Y_n] < 1$. En déduire la réponse (en fonction de n) à la question posée.

Indications

IV.1 : Nombres réels suites numériques

IV.1.5 Les réels $\cos 2\pi/7$, $\cos 4\pi/7$ et $\cos 6\pi/7$ sont les solutions de :

$$8X^3 + 4X^2 - 4X - 1 = 0.$$

IV.1.7 On montrera que f est croissant.

IV.1.9 2. $E(a) = 99$.

IV.1.18 1. Utiliser :

$$\sin a - \sin b = 2 \sin \frac{a-b}{2} \cos \frac{a+b}{2}.$$

2. Noter $x_k := \mathrm{e}^{\frac{\pi}{2}+2k\pi}$ et utiliser $n_k := E(x_k) + 1$.

IV.1.25 1. On a $\operatorname{Ker} \Delta = \mathbb{Q}$ et $\Delta(\mathbb{Q}_r[X]) \subset \mathbb{Q}_{r-1}[X])$. En appliquant le théorème du rang, on montre $\Delta(\mathbb{Q}_r[X]) = \mathbb{Q}_{r-1}[X]$. 2. On utilise une somme télescopique et l'on obtient $S_k(n) = P_k(n)$. On cherche P_2 sous la forme $P_2 = ax^3 + bx^2 + cx$ et l'on trouve $a = \frac{1}{3}$, $b = \frac{1}{2}$, $c = \frac{1}{6}$. 3. On calcule P_k en résolvant un système linéaire. On trouve $P_3 = \frac{1}{4}x^4 + \frac{1}{2}x^3 + \frac{1}{4}x^2$.

IV.1.28 1. Le nombre de 4 cycles est 3, celui de 5-cycles est 6, celui de 6-cycles est 9. 2. $\sum_{\tau'|\tau} \omega(\tau') = 2^\tau$. 3. $\omega(17) = 7710$.

IV.1.29 1. Dans les 4 cas où $u_0 \neq 0$, ce sont des cycles. 3. f n'admet pas de 5-cycle positif.

IV.1.35 1. Partager $J := [0, 1[$ en Q intervalles semi-ouverts disjoints de même longueur et considérer $\{\langle m\alpha \rangle\}_{m \in [\![0,Q]\!]} \subset J$.

IV.1.38 La suite est une suite géométrique de raison b/a. Elle converge si, et seulement si, $b \leqslant a$.

IV.1.39 2. $\lim v = 1/2$.

IV.1.40 La suite u converge si, et seulement si, la suite $(v_n^{2^{-n}})$ est majorée.

IV.1.41 2. Si $\lim = \ell$, alors $\lim u = \ell_1$.

IV.1.42 On a $\ell = r$ si $\theta = 0$ et $\ell = r\frac{\sin\theta}{\theta}$ si $\theta \neq 0$.

IV.1.43 1. a) Les limites sont 5 ou 6. 1. b) On calcule $\frac{u_{n+1}-5}{u_{n+1}-6}$. 2. $\rho = 5$ ou $\rho = 6$.
3. On utilise $v_n = w_{n+1}/w_n$.
4. On raisonne par récurrence en utilisant 2. b).

IV.1.45 1.a) $v_n = \prod_{k=1}^{n} \cos \frac{\alpha}{2^k}$ et $u_n = v_n \cos \frac{\alpha}{2^n}$. 1.b) $\ell = \frac{\sin\alpha}{\alpha}$.
2. Si $a < b$: $\ell = \frac{\sqrt{b^2-a^2}}{\arccos(a/b)}$. Si $a > b$: $\ell = \frac{\sqrt{a^2-b^2}}{\operatorname{argcosh}(a/b)} = \frac{\sqrt{a^2-b^2}}{\ln\left(a/b+\sqrt{(a/b)^2-1}\right)}$.

IV.1.47 La période est 5.

IV.1.51 Écrire x sous la forme $x = \sin^2(\pi a/2)$.

IV.1.56 2. On reprend la méthode de la solution de l'exercice 39 de la page 589.

IV.1.57 Comparer α_n et $\alpha_{n+10\times k!}$.

IV.1.59 4. Si $P = a_d x^d + a_{d-1}x^{d-1} + \cdots + a_0$, alors $Q = a_0 S_\delta + a_1 S_{\delta-1} + \cdots + a_{\delta-1} S_1 + a_\delta$. 7. $P(Y) := 16Y^4 - 8Y^3 - 16Y^2 + 8Y + 1$.

IV.1.60 2. $a_{n,k} := (-1)^{k-n}\binom{n}{k-n} \in \mathbb{N}$.
3. Si $\pi = p/q \in \mathbb{Q}$, on en déduit $I_n > 1/q^n$, puis $n! \leqslant q^n \pi^{2n+1}$.

IV.1.61 1. d) Utiliser $S_k(1) = 1$.
2. $f_n = (-1)^n b_n$.
4. b) On utilise la formule du binôme, la question 3, un argument d'interversion de sommation et la formule de récurrence définissant les $\{f_n\}$. C'est la partie difficile de l'exercice.

IV.1.62 1. a) Les limites possibles de la suite sont 5, 6 ou 10. 1. e) La suite (u_n) tend vers 6. 2.c)
$v_n := \frac{\lambda_1 5^{n+1} + \lambda_2 6^{n+1} + \lambda_3 100^{n+1}}{\lambda_1 5^n + \lambda_2 6^n + \lambda_3 100^n}$.
3. Utiliser $\sigma_1 := p_1 + p_2 + p_3$, $\sigma_2 := p_1p_2 + p_2p_3 + p_3p_1$ et $\sigma_3 := p_1p_2p_3$.

IV.1.64 Soit (a, b, c) un 3-cycle de f, avec $a < b < c$. On note $I_0 := [a, b]$ et $I_1 := [b, c]$.
1. Utiliser $I_1 \subset f(I_1)$.
2. Montrer qu'il existe un segment $J \subset I_0$ tel que $f(J) = I_1$, puis utiliser $f \circ f$.
3. Construire une suite d'intervalles emboîtés :
$A_{\tau-2} \subset A_{\tau-3} \subset \cdots \subset A_1 \subset A_0 = I_1$, telle que, pour tout $i \in [\![0, \tau-3]\!]$, l'on ait $f(A_{i+1}) = A_i$.
Puis montrer l'existence d'un segment $A_{\tau-1} \subset A_{\tau-2}$ tel que $f^{\circ\tau-1}(A_{\tau-1}) = I_0$.

IV.2 : Fonctions réelles

IV.2.4 Appliquer le théorème des valeurs intermédiaires à la fonction $x \mapsto g(x) = f(x) - x$.

IV.2.6 1. Soient $a \in \mathbb{R}$, $\varepsilon > 0$ et τ une période de f. Alors l'intervalle $]a - \varepsilon, a + \varepsilon[$ contient un multiple entier (relatif) de τ. 2. Remarquer que, si $p \in \mathbb{N}^*$, $(\sqrt{2} - 1)^p$ est une période.

IV.2.9 On construit des exemples tels que $f(0) = 0$ en utilisant la définition d'une fonction par son graphe et la question 1.

IV.2.10 2. Les transformations affines de $\mathbb{R}$ forment un sous-groupe infini de G. 3. Soit $H \subset G$ un sous-groupe fini de G. On montre que si $f \in H$ est strictement croissante, elle est égale à l'identité. On en déduit que tout élement de H est d'ordre 2, puis que H a un ou deux éléments.

IV.2.11 La fonction f est dérivable en tout point de $\mathbb{R}$ exceptés -1, 0 et 1.

IV.2.14 Utiliser $f(x) = -1 + 2/(1 + x)$ et $f(x) = (1 - \cos 2x)/2$.

IV.2.16 2. Pour tout $\varepsilon > 0$, on définit une fonction $g_\varepsilon : [a, b] \to \mathbb{R}$ par $x \mapsto g_\varepsilon(x) := f(x) + \varepsilon x$ et l'on utilise 1.

IV.2.17 On montre que $f'(I)$ est convexe en utilisant les prolongements par continuité des fonctions taux d'accroissement φ_a et φ_b en a et b.

IV.2.22 On applique le théorème de Rolle à une fonction convenable de la forme $x \mapsto \varphi(x) := f(a) - 2f(a+x) + f(a+2x) - Ax^2$.

IV.2.23 Utiliser la formule de Taylor-Lagrange à l'ordre deux pour exprimer $f(x+y)-f(x)$ et $f(x-y)-f(x)$.

IV.2.24 On montre qu'il existe un réel a tel que, pour tout $n \in \mathbb{N}$, $f^{(n)}(a)=0$ et l'on applique la formule de Taylor-Lagrange au point a.

IV.2.28 Utiliser la proposition 59 pour la fonction $x \mapsto x^p$.

IV.2.29 2. Utiliser la proposition 59 avec $x_i := \frac{a_i^p}{(a_i+b_i)^p}$ et $\lambda_i := \frac{(a_i+b_i)^p}{\sum\limits_{j=1}^{n}(a_j+b_j)^p}$..

IV.2.30 On utilise l'inégalité de Young avec, pour tout $i \in [\![1,n]\!]$, $x_i := a_i/\|a\|_p$ et $y_i := b_i/\|b\|_p$. On somme ensuite sur i.

IV.3 : Fonctions transcendantes

IV.3.12 1. Utiliser $0 \leqslant \arctan x + \arctan y < \pi$.
2. $\arctan \frac{1}{\sqrt{3}+2} = \frac{\pi}{12}$.

IV.3.13 3. La suite $u_n(x)$ converge vers $\arctan x$ lorsque $n \to +\infty$.

IV.3.14 1. $\tan(4\arctan 1/5) = 120/119$.
2. $\tan\left(4\arctan\frac{1}{5} - \arctan\frac{1}{239}\right) = 1$.
5. $\ell = \pi$.

IV.3.21 3. $\left(\frac{u+i}{u-i}\right)^m \left(\frac{v+i}{v-i}\right)^n = i$.

IV.3.25 On a $f^2 - x^2 = 0$ et $g^2 - x^2 + 2x - 1 = 0$ et $h^4 - (4x^2-4x+2)h^2 + 4x^2 - 4x + 1 = 0$.

IV.3.27 On raisonne par l'absurde en supposant que $a_n(\ln)^n + a_{n-1}(\ln)^{n-1} + \cdots + a_0 = 0$, avec $a_0, \ldots, a_n \in \mathbb{R}[X]$, $a_n \neq 0$ et $n \in \mathbb{N}^*$ *minimal*.

IV.8 : Approximation

IV.8.1 $I = \frac{\ln 2}{2}$.

IV.8.4 2. On pourra utiliser les exercices 2 et 3 de la page 753 sur la première formule de la moyenne.

IV.8.8 $x - \dfrac{x^3}{6} + \dfrac{3x^5}{40} + \mathrm{o}(x^5)$.

IV.8.17 Utiliser la question 1 de l'exercice IV.8.0 de la page 903.
On obtient $\sinh \circ \sin - \sin \circ \sinh \sim \frac{1}{45}t^7$ et $\tanh \circ \tan - \tan \circ \tanh \sim -\frac{2}{45}t^7$.

IV.8.22 1. Utiliser $f(x) := x - a$.
2. et 3. Utiliser $f(x) := (x-a)^2$.
4. Utiliser la remarque de la page 896.
5. On peut choisir $M_4 = 0$.
6. On peut choisir $f : x^4/4!$ et $M_4 := 1$.

V.1 : Statistique descriptive

V.1.4 Pour la seconde question, on pourra remarquer que L est affine par morceaux.

V.1.5 On pourra partir de la définition de l'écart-type pour aborder la première question.

V.1.7 On pourra faire une récurrence pour montrer la question 3, et on peut obtenir ce résultat sans utiliser la question 2.

V.2 : Probabilités finies

V.2.2 On pourra essayer de construire une bijection de $\mathcal{C}(E,F)$, l'ensemble des applications strictement croissantes de E dans F, dans $\mathcal{P}_p(F)$, l'ensemble des parties de F à p éléments.

V.2.4 On pourra utiliser le développement limité de la fonction $\ln$ pour la seconde question.

V.2.9 On pourra introduire l'événement A : « le joueur est Flamand » et l'événement B : « le joueur est un amateur de frites hexagonales », et utiliser la formule de Bayes.

V.2.10 On pourra introduire l'événement $A =$ « le moteur des essuie-glaces est en panne », l'événement $B =$ « la commande manuelle des essuie-glaces est en panne » et l'événement $C =$ « il pleut ».

V.2.11 On pourra faire intervenir les événements $T_i =$ « le parapluie est dans le i-ème tiroir », pour $i \in [\![1,4]\!]$.

V.3 : Variables aléatoires

V.3.2 On peut noter V_k l'événement « la porte n'est toujours pas ouverte après k essais ». En remarquant qu'il s'agit d'une suite décroissante d'événements, dans le sens où $V_k \subset V_{k-1}$, exprimer la probabilité p_k qu'il faille exactement k essais pour ouvrir la porte en fonction de la probabilité de V_k. Puis calculer la probabilité de V_k. Appliquer ensuite la définition de l'espérance d'une v.a. réelle.

V.3.3 En notant A_k l'événement « il y a rencontre au $k^{ème}$ tirage », utiliser la formule d'inclusion-exclusion.
Pour le comportement de la probabilité p_n lorsque $n \to +\infty$, il faut utiliser la définition de l'exponentielle complexe.

V.3.4 On peut chercher le nombre de tirages qui ont été nécessaires pour qu'il reste exactement r allumettes dans l'autre boîte lorsque le fumeur s'aperçoit qu'une boîte est vide. Puis on peut examiner comment ont pu se répartir ces tirages (entre poche gauche et poche droite) pour que cet événement se produise.
La valeur de la somme $\sum_{r=0}^{N} u_r = 1$ peut s'obtenir sans calcul.

V.3.5 Pour le 2., on peut introduire Y la variable aléatoire correspondant au gain du joueur et on peut déterminer sa loi.

V.3.7 Développer l'expression de $\mathbb{E}[(X-c)^2]$, qui, en tant que fonction de c, est un polynôme du second degré avec un coefficient dominant strictement positif. En déduire l'existence, l'unicité, la position et la valeur du minimum de cette fonction.

V.3.8 La somme dans (29) est finie, car X est à valeurs dans un sous-espace fini de $\mathbb{N}$ donc $\mathbb{P}(X > n) = 0$ à partir d'un certain rang.

V.3.9 Pour la question 2, on pourra utiliser le fait que, si X est une v.a. entière positive, alors :

$$\mathbb{E}[X] = \sum_{k \geqslant 0} \mathbb{P}(X > k) = \sum_{k \geqslant 1} \mathbb{P}(X \geqslant k).$$

V.3.11 On peut commencer par identifier les valeurs possibles que peut prendre Y_n.

Index

A

abélien (groupe) 118
abscisse curviligne 488, 489
absolument convergente (série) 730
absorbant (élément —) 27
absurde (démonstration par l'—) 54
accroissements finis
    égalité des — 648
    inégalité des — 650, 794, 840
    théorème des — 794, 839
adaptée
    base — à une décomposition en somme directe) 197
    subdivision — à une fonction en escalier 740
addition
    des entiers naturels 66
    des entiers relatifs 93
    des nombres réels 594
    des nombres rationnels 102
    vectorielle 163
adhérence (valeur d' —) 563
adjacentes (suites —) 560
affectation 394, 401
    simultanée 394, 404
affine
    application — 432
    droite — 376, 427
    forme — 434
    géométrie — 425
    hyperplan — 376, 437
    plan — 376, 427
    projection — 435
    repère — 439
    sous-espace — 375, 429
    symétrie — 435
affinité 436
    orthogonale 455
affixe
    d'un point du plan 247
    d'un vecteur 247
Aleph zéro 43
algèbre sur un corps commutatif 217
algébrique (fonction —) 677
algébriquement clos (corps —) 299
algorithme(s)
    affectation 394
    complexité d'un — 407
    correction d'un — 399
    coût d'un — 399
    d'addition en base b 392
    d'Euclide 408
        pour les polynômes 294
    d'Euclide étendu 98
        pour les polynômes 296
    de Gauß 225
    de la division euclidienne des polynômes 289
    de tri 412
    en informatique 392
    sur les polynômes 417
    terminaison d'un — 398, 405
alterné(e)
    groupe — 130
    série — 732
amplitude d'une classe 919
analyse d'un algorithme 405
angle
    de droites 268
    de vecteurs 268
    doublement de l'— 576

anneau 132
commutatif 132
de Boole 137
intègre 137
ordonné 37
principal 144
produit 134
antécédent 13
antisymétrique
matrice — 216
relation — 29
appartenance 4
application 11
affine 432
caractéristique 24
constante 12
contractante 563
coordonnée 12
croissante 35
décroissante 35
identité 12
induite 14
itérée 525
linéaire 172
nulle 173
triviale 173
logistique 570
monotone 35
strictement (dé)croissante 35
strictement monotone 35
tente 529
vide 12
approximation décimale 560, 729
arc-cosinus 697
fonction — 696
arc-sinus 696
fonction — 696
arc-tangente 692
fonction — 692
Archimède
propriété d'— 69, 101, 518
archimédien 518
corps — 518
Argand-Cauchy (plan d'—) 247
argument
d'un nombre complexe 262
de la cotangente hyperbolique 703
de la diagonale 592
de la tangente hyperbolique 701
détermination principale de l'— 263
du cosinus hyperbolique 701
du sinus hyperbolique 701
principal 263, 483
arithmético-géométrique (suite —) 535
arithmétique (suite —) 531
arrangement(s) 76
nombre d'— 76
artinien (ensemble —) 37
associés (éléments —) 143
associative (loi —) 25
associativité
de l'addition des entiers 67
de la multiplication 68
du barycentre 429
asymptote 478, 485, 668
d'une hyperbole 462
asymptotique
caractère — de la notion de limite d'une suite 543
comportement — d'une suite 539
direction — 478, 485
attractif (point —) 564
automorphisme 27
d'anneau 140
d'espace vectoriel 175
de groupe 124
axe d'une symétrie 201
axiome(s)
d'extensionalité 4
pour les applications 12
de la paire 5
de séparation 7
des nombres réels 514
du choix 23
du singleton 5

B

babylonienne (tablette —) 558
barycentre 428
  associativité du — 429
base
  adaptée à une décomposition en somme directe 197
  canonique
    de $M_{m,n}(K)$ 209
    de $K^{(I)}$ 193
    de K^n 193
    de $K[X]$ 192, 283
    de $K_n[X]$ 287
  d'un espace vectoriel 192
  de Hamel 206
  duale 194
  orthonormée 453, 462
basique
  partie — d'un espace vectoriel 192
Bayes (formule de —) 958
Bernoulli
  loi de — 969
  nombres de — 532, 612, 884, 886
Bézout
  coefficients de — 98, 295
  théorème de — 97
    pour les polynômes 292
bicontinue (fonction) 637
bien ordonné (ensemble —) 38
Bienaymé-Tchebychef
  inégalité de — 980
bijection 15
bijective (application) 15
binaire
  algorithme — pour le PGCD 108
  relation — 28
binôme (formule du) 134
binomiale (loi —) 970
binomiaux (coefficients) 78
birégulier (point —) 477, 493, 499
boîte à moustaches 927
Bolzano-Weierstraß
  théorème de — 562, 784, 822
bon ordre 38
booléen 400
bordante (matrice) 362
borne
  inférieure 34, 510
  supérieure 34, 510
borné(e)
  partie — 34, 822
  suite — 784
boucle
  conditionnelle 402
  inconditionnelle 403
branche
  infinie 478, 485, 668
  parabolique 669
but d'une application 12

C

calcul formel (computer algebra) 391
canonique
  base — de $M_{m,n}(K)$ 209
  base — de $K[X]$ 192, 283
  base — de K^n 193
  base — de $K^{(I)}$ 193
  morphisme — 131
  structure affine — 427
Cantor
  procédé diagonal de — 24
  théorème de — 14
  théorème de —, Schröder, Bernstein 40
caractère 915
caractérisation séquentielle
  des bornes 544
caractéristique
  d'un anneau 142
  déterminant — d'un système linéaire 381
  fonction — 24, 76, 159
  polynôme — 371, 373
cardinal
  d'un ensemble 39
  d'un ensemble fini 70

cartésien(ne)
équation —
d'un plan 445
d'une droite 442, 449
produit — 11
d'une famille d'ensembles 22
repère — 439
Catalan (nombres de —) 26, 58
Cauchy
formule de la moyenne de — 651
produit de — 271, 731
théorème de — 156
Cauchy-Riemann (conditions de —) 848
Cauchy-Schwarz
inégalité de — 251, 453, 751, 781, 986
central(e)
endomorphisme — 200
matrice — 217
centre
d'un groupe 121
d'une homothétie 434
de $M_n(K)$ 218
de courbure 493, 494, 500
cercle 457
dans $\mathbb{C}$ 252
trigonométrique de $\mathbb{C}$ 257
unité 257
Cesaro (moyenne de —) 542
chaîne 39
Champernowne (nombre de —) 616
changement de variable 569, 764
Chasles
relation de — 269, 426, 748, 757, 796
chinois (théorème) 149
choix (axiome du —) 23
classe 918
amplitude d'une — 919
d'équivalence 32
fréquence d'une — 919
clos (corps algébriquement —) 299
clôture
réflexive transitive d'une relation 30
transitive d'une relation 30
codage 392, 396
coefficient(s)
au dessous de la diagonale 216
au dessus de la diagonale 216
constant d'un polynôme 284
d'indices (i, j) d'une matrice 207
d'un développement limité 871
d'une combinaison linéaire 165
de Bézout 295
de corrélation 932
de degré k d'un polynôme 284
diagonaux d'une matrice 216
dominant d'un polynôme 285
relations entre — et racines d'un polynôme 299
strictement au dessous de la diagonale 216
strictement au dessus de la diagonale 216
cofacteurs (d'une matrice) 361
collectivisante (propriété) 7
comatrice 361
combinaison linéaire 165
coefficients d'une — 165
combinaisons 78
avec répétitions 80
commutatif(ve)
anneau — 132
corps — 151
groupe — 118
ordonné 36
loi — 26, 113
commutativité
de l'addition des entiers 67
de la multiplication 68
compact 822
comparables (éléments —) 33
compatible(s)
relation d'ordre et loi de composition interne — 36
système — 379
complémentaire d'un sous-ensemble 9
complexe (nombre —) 244, 245

complexité
classe de — 407
d'un algorithme 407
en espace 407
en temps 407
composée de deux applications 16
compréhension (définition en —) 7
compteur 406
concave (fonction —) 658
conditions de Cauchy-Riemann 848
congruence (modulo un entier) 130
conique 458
à centre 464
conjonction 47
conjugaison 124, 569
dans $\mathbb{C}$ 248
formule de — 127
conjugué d'un nombre complexe 248
connecteur logique 47
constante d'Euler 329
construction à la règle et au compas 270
continu (puissance du —) 44
continue
fonction — 631, 846
fonction — par morceaux 743, 756
continuité 789, 821
continûment
derivable dérivable 645
différentiable 838
contractante
application — 563
fonction — 632
contradiction 49
convergente
série — 714
suite — 271, 542
convexe
fonction — 658
partie — de $\mathbb{R}$ 521
coordonnées
d'un point dans un repère 439
d'un vecteur dans une base 192
polaires 483
corestriction d'une application 12
corps 151
algébriquement clos 299
archimédien 518
commutatif 151
des fractions 153
des fractions rationnelles 315
finis 153, 300
ordonné 37
corrélation 932, 986
correction d'un algorithme 399, 406
correspondance 28
cosinus 256, 273
fonction — 688
fonction — hyperbolique 698
cotangente 691
fonction — 691
fonction — hyperbolique 698
couple 10
coupure 581
courbe 469
de classe $\mathcal{C}^k$ ou $\mathcal{C}^\infty$ 471
de l'espace 497
de niveau 820
définie par une équation 820
géométrique 488, 499
orientée 488, 499
paramétrée 470, 497
plane 469
polaire 483
rectifiable 486
régulière 472
support d'une — 470
courbure 491, 500
algébrique 492, 494
centre de — 493, 494, 500
rayon de — 493, 494, 500
coût
bilinéaire 421, 422
d'un algorithme 399
de l'exponentiation 410, 422
en espace 400
linéaire 399, 407

logarithmique 400, 407
maximum, ou le pire, d'un algorithme 407
moyen d'un algorithme 407, 409
polynomial 407
quadratique 419
covariance 932, 986
Cramer
formules de — 380
système de — 379
crible
d'Ératosthène 84
formule du — 136
croissante
application — 35
fonction — 628
suite — 35, 556
curviligne (abscisse —) 488, 489
cycle 126, 527, 557
cyclique (groupe) 122
cycloïde 503
cyclotomie 307

D

d'Alembert (règle de —) 723
d'Alembert-Gauß (théorème de —) 275
Darboux (théorème de —) 642, 673
décimale 582
décomposition en éléments simples d'une fraction rationnelle 320
décroissante
application — 35
fonction — 628
suite — 35, 556
définition
en compréhension 7
en extension 6
degré
d'un polynôme 285
d'une fraction rationnelle 316
demi-plan de Poincaré 279
démonstration
par contraposée 55
par disjonction de cas 55
par l'absurde 54
méthode de l'hypothèse auxiliaire 54
par récurrence à deux pas 55
par récurrence forte 56
par récurrence simple 55
dénombrable (ensemble —) 43, 74
dense (partie —) 519, 547
dérivabilité 791
dérivable
à droite 641
à gauche 641
continûment — 645
fonction — 640, 846
dérivation(s)
numérique 908, 910
supérieures d'un polynôme 305
sur $\mathbb{C}$ 846
dérivé(e) 640
à droite 641
à gauche 641
d'un polynôme 301
d'une fraction rationnelle 322
logarithmique 315
$n^{ème}$ 644
partielle 826
suivant un vecteur 828
déterminant 442
caractéristique 381
d'un endomorphisme 364
d'une matrice carrée 352
de Vandermonde 360
de n vecteurs 365
détermination principale
de l'argument 263
du logarithme 704
développement limité 871
coefficients du — 871
d'une fraction rationnelle 323
partie régulière du — 871

diagonale
argument de la — 592
d'une matrice 216
matrice — 216
diagonalisable 462
diagonaux
coefficients — d'une matrice 216
diagramme 18
à secteurs 921
en bâtons 919
dichotomie 634
différence
de deux ensembles 9
de deux entiers naturels 67
symétrique de deux ensembles 28, 137
différentiabilité 826
différentiable 829
continûment — 838
différentielle 829
dilatation
matrice de — 229
sur les colonnes 227
sur les lignes 226
dimension 195
d'un espace affine 427
d'un espace vectoriel 333
d'un sous-espace affine 375
espace vectoriel de — finie 332
finie 440
directe
base — 364
somme — de deux sous-espaces vectoriels 196
direction
asymptotique 478, 485
d'un espace affine 426
d'un sous-espace affine 375
d'une symétrie 201
directrice 458
disjoints (ensembles —) 8
disjonction 47
disque
fermé dans $\mathbb{C}$ 252
ouvert dans $\mathbb{C}$ 252
distance 451, 452, 780
distributivité 68
divergente
série — 714
suite — 542
diviser pour régner (divide and conquer) 410
divisible (groupe —) 203
division
euclidienne 81, 94
des polynômes 288
par puissances croissantes 321
dominant (terme, coefficient — d'un polynôme) 285
domination 869
donnée 916
doublement de l'angle 576
droite
affine 427
vectorielle 171, 334
dual(e)
base — 194
espace — 184
dyadique (nombre) 110

E

écart-type 927, 979
échange
des colonnes 227
des lignes 226
matrice d'— 228
échantillon 916
taille d'un — 916
échelonnée (matrice —)
en colonnes 210
en lignes 210
écriture
d'un élément dans une base 192
décimale 583
décimale impropre 583
trigonométrique d'un nombre complexe 262
effectif d'une classe 918, 919

effective (résolution) 391
élément
idempotent 137
inversible 116
d'un anneau 137
involutif 116
irréductible 144
neutre 26, 115
nilpotent 137
régulier 116
simple 770
symétrisable 116
élémentaire(s)
matrice — 209
opérations — 226
ellipse 458, 459
emboîtés
propriété des segments — 516, 599
endomorphisme 27
d'un espace vectoriel 175
engendré
idéal — par une partie 140
sous-anneau — par une partie 140
sous-espace vectoriel —
par une famille 170
par une partie 170
sous-groupe — par une partie 122
ensemble
d'arrivée d'une application 12
de définition 619
d'une fraction rationnelle 322
de départ d'une application 12
de représentants 32
d'une classe d'équivalence 32
dénombrable 74
des parties d'un ensemble 8
fini 42, 70
fondamental 941
image d'une application 13
infini 42, 70
ordonné 33
vide 5
entière (partie) 103
entiers
de Gauß 278
de Von Neumann 43
naturels 63
relatifs 91
énumération d'un ensemble 43
équation
affine 438
cartésienne
d'un plan 445
d'une droite 442, 449
différentielle 797
du premier ordre 798, 802
du second ordre 811
homogène 799, 802
logistique 608
solution maximale d'une — 799, 802
du 2ème degré dans $\mathbb{C}$ 255
du 3ème degré dans $\mathbb{C}$ 265
du 4ème degré dans $\mathbb{C}$ 277
paramétrique
d'un plan 447
d'une droite 443
équilatère (hyperbole —) 462, 698
équipotents (ensembles —) 39
équiprobabilité 946
équivalence
classe d'— 32
logique 47
relation d'— 29
équivalentes
fonctions — 861
matrices — 220
Ératosthène (crible d'—) 84
escalier (fonction en —) 740
espérance 976
espace
affine 426
d'arrivée d'une application linéaire 172
de départ d'une application linéaire 172
de probabilité fini 944
dual 184

espace(s) vectoriel(s) 162
isomorphes 174
nul 164
trivial 164
étoile
Chérifienne 261
de David 261
Euclide
algorithme d'— 89, 294, 408
étendu 98, 296, 409
lemme d'— 85, 97, 145
pour les polynômes 293
euclidienne (division —) 81, 94
pour les polynômes 288
Euler
constante d'— 329, 727
fonction indicatrice d'— 89, 146
formules d'— 265
théorème d'— 147
évaluation
d'un polynôme (coût de l'—) 284
paresseuse 421
événement(s) 942
disjoints 943
élémentaire 943
incompatibles 943
indépendants 959, 962
excentricité 458
exponentiation dichotomique (ou babylonienne, chinoise, indienne) 410
exponentielle 680
complexe 256, 272
d'une matrice 803
de base a 683
fonction — 680
expression polynomiale 281
extension (définition en —) 6
extensionalité (axiome d'—) 4
pour les applications 12
externe (multiplication — dans un espace vectoriel) 162
extremum 647

F

Faa di Bruno (formule de —) 646
factorielles 76
factorisation (d'un entier) 86
famille 20
à support fini 163
canoniquement associée à un ensemble 21
d'ensembles 21
de polynômes à degrés étagés 287
de polynômes à valuations étagées 290
extraite 21
génératrice d'un espace vectoriel 185
liée d'un espace vectoriel 190
libre d'un espace vectoriel 187
vide 21
Faulhaber (formule de —) 613
Fermat
dernier théorème de — 138
descente infinie de — 66
nombre de — 85, 146
petit théorème de — 147
fermé 821, 822
Fibonacci (suite de —) 108, 550
théorème de Lamé 409
fibre (d'une application) 72
fini(e)
dimension — 195
ensemble — 42, 70
groupe — 118
fixe (point —) 526
fonction(s)
algébrique 677
arc-cosinus 696
arc-sinus 696
arc-tangente 692
argument de la tangente hyperbolique 701
argument du cosinus hyperbolique 701
argument du sinus hyperbolique 701
bicontinue 637
caractéristique 24, 76, 159

concave 658
continue 631, 846
contractante 632
convexe 658
cosinus 688
cosinus hyperbolique 698
cotangente 691
croissante 628
d'erreur 901
de Dirichlet 619
de Möbius 158
de plusieurs variables 819, 821
décroissante 628
dérivable 640, 846
dérivée 642
exponentielle 654, 680
exponentielle complexe 256, 272
graphe d'une — 819
holomorphe 846
impaire 620
indicatrice d'Euler 89, 146
lipschitzienne 632
logarithme népérien 678
monotone 628
paire 620
périodique 620
polynomiale 296
puissance(s) 684
réelle 619
racine 684
rationnelle 322
sinus 688
sinus hyperbolique 698
strictement monotone 628
symétriques élémentaires 299
tangente 691
tangente hyperbolique 698
transcendante 677
vectorielle 779, 789

forme
affine 434
indéterminée 867, 891
infixe d'une loi de composition 24
linéaire 181
linéaire coordonnée relative à une base 194

formule(s)
d'Euler 265
d'inversion de Möbius 158
de Bayes 958
de Cardan, de Scipion del Ferro 265
de Faa di Bruno 646
de Faulhaber 613
de Grassmann 339
de la moyenne de Cauchy 651
de Leibniz 645
pour les polynômes 302
de Mac-Laurin 652
de Machin 709
de Madhava-Gregory 709
de Moivre 267, 689
de Newton 134
de Serret-Frenet 493, 501
de Taylor 652, 856
pour les polynômes 304
de Taylor-Lagrange 652
de Taylor-Young 655, 859
des marginales 984
des probabilités composées 956
des probabilités totales 957
du binôme 134
du crible 136
du multinôme 158

foyer 458

fractile 927

fraction rationnelle 315
décomposition en éléments simples d'une — 320
degré, degré total d'une — 316
dérivée d'une — 322
développement limité d'une — 323
ensemble de définition d'une — 322
forme irréductible d'une — 316
ordre d'un pôle d'une — 318
ordre d'un zéro d'une — 318
ordre d'une — 318
partie entière d'une — 319
pôle d'une — 318
résidu en un pôle d'une — 320
valeur d'une — 317
zéro d'une — 318

fractions (corps des —) 153
fréquence d'une classe 918, 919
Fubini (formule de —) 72

G

Gauß
  algorithme de — 225
  entiers de — 278
  théorème de — 88, 97
    pour les polynômes 293
gendarmes (théorème des —) 555, 628
génératrice
  famille — d'un espace vectoriel 185
  partie — d'un espace vectoriel 185
  partie — d'un groupe 122
géométrie
  affine 425
  euclidienne 451
géométrique
  courbe — 488, 499
  série — 716
  suite — 532
grands nombres (loi des —) 991
graphe 29, 841
  d'une application 13
  d'une fonction 668, 819
Grassmann (formule de —) 339
groupe 118
  abélien 118
    divisible 203
    sans torsion 203
  alterné 130
  commutatif 118
    ordonné 36
  cyclique 122
  des éléments inversibles 137
  fini 118
  linéaire $GL_n(K)$ 219
  linéaire $\mathcal{GL}_K(V)$ 184
  monogène 122
  produit 118
  symétrique 125
    d'un ensemble 118
  trivial 118

H

Hamel (base de —) 206
harmonique (série) 719
hauteur d'une matrice colonne 210
hélice 498
hérédité 38, 55
Héron (suite de —) 558
histogramme 920
Hölder (inégalité de —) 675
holomorphe (fonction —) 846
homéomorphisme 470, 637
homogène
  équation linéaire — 375
  système linéaire — 377
homographie 329, 535
homographique
  suite — 535
homothétie 173, 434
Horner (schéma de —) 285
hyperbole 458, 460
  équilatère 462, 698
hyperbolique
  (co)sinus — 698
  (co)tangente — 698
  rotation — 708
hypergéométrique (loi —) 975
hyperplan
  affine 437
  vectoriel 181, 334
hypothèse du continu 44

I

idéal 139
  engendré par une partie 140
  principal 144
idempotent (élément) 27, 137
identité
  application — 12
  du parallélogramme 426
Im *voir* image

image
d'un élément 11
d'un morphisme 123
d'un sous-ensemble par une application 13
d'une application 13
d'une application linéaire 178
réciproque 17
imaginaire
partie — d'un nombre complexe 245
pur 245
impair(e)
fonction — 620
polynôme — 285
implication
logique 47
matérielle 48
inégalité
de Bienaymé-Tchebychef 980
de Cauchy-Schwarz 986
incertitude 894
inclusif (ou — des logiciens) 48
inclusion 7
indépendance linéaire 187
indéterminée 281
indexé(e)
famille d'ensembles — 21
famille — 20
terme — 20
indice 20
inductif (ensemble —) 39
induction
noetherienne 37
principe d'— 37
induite (loi —) 25, 117
inégalité
de Cauchy-Schwarz 251, 453, 781
de Hölder 675
de Minkowski 675
de Taylor-Lagrange 654, 859
de Young 667
des accroissements finis 650, 794, 840
pour les polynômes 306
du triangle dans $\mathbb{C}$ 251
triangulaire 452, 487, 517, 781
infimum 635
infini(e)
branche — 478, 485
ensemble — 42, 70
infixe
forme — d'une loi de composition 24
notation — d'une relation binaire 29
inflexion (point d'—) 475
initialisation (d'une variable) 404
injection 15
canonique 12
injective (application) 15
instruction 401
conditionnelle 402
d'un algorithme 395
intégrale
d'une fonction continue par morceaux 747
d'une fonction en escalier 741
d'une fonction vectorielle 795
de Wallis 773
intégration par parties 761
intègre (anneau —) 137
intérieur 628
d'un intervalle 520
point — 520
interpolation
de Lagrange 340
linéaire 905
polynômes d'— 298
intersection
d'un plan et d'une droite 450
d'une famille d'ensembles 22
de deux droites 444
de deux ensembles 8
de deux plans 448
intervalle 63, 520
invariant de boucle 406
inverse 26
d'un élément 116
inversible (élément —) 26, 116

inversion
d'une matrice par opérations élémentaires 235
d'une permutation 130
théorème d'— locale 837
involutif
élément — 116
endomorphisme — 201
irréductible
élément — 144
forme — d'un rationnel 104
forme — d'une fraction rationnelle 316
polynôme — 290
isomorphes (espaces vectoriels —) 174
isomorphisme 27
d'anneaux 140
d'espaces vectoriels 174
de groupes 124, 680, 681
itérées (d'une application) 65, 525

J

jacobienne (matrice —) 830, 831

K

Ker *voir* noyau
K-espace vectoriel 162
Knuth (Donald E.) 409
Koenig (théorème de —) 926
Kronecker (notation de —) 56, 166, 193, 209

L

Lagrange
polynôme d'interpolation de — 298, 340
théorème de — 121, 125
Lamé (théorème de —) 409
Landau (notations de —) 867, 869
largeur d'une matrice ligne 210
Leibniz
formule de — 645
pour les polynômes 302
règle de — pour les polynômes 302
lemme
d'Euclide 97
pour les polynômes 293
de Gauß 88
pour les polynômes 293
de Zorn 39
lexicographique (ordre —) 34, 106
libre
famille — d'un espace vectoriel 187
partie — d'un espace vectoriel 187
variable — 53
lié(e)(s)
complexes positivement — 249
complexes réellement — 249
famille — d'un espace vectoriel 190
partie — d'un espace vectoriel 190
variable — 53
lieu géométrique 248
limite
à droite 624
à gauche 624
d'une fonction 622, 789
d'une suite 541, 782
linéaire
application — 172
combinaison — 165
forme — 181
groupe — $GL_n(K)$ 219
groupe — $\mathcal{GL}_K(V)$ 184
partie — d'une application affine 432
système — 440
linéairement (in)dépendants (éléments — d'un espace vectoriel) 187, 190
linéarisation d'une expression trigonométrique 265
Liouville (nombre de —) 314, 590, 610
lipschitzienne (fonction —) 632
liste 401
logarithme
binaire 682
de base a 682
décimal 682
détermination principale du — 704
népérien 678

logique
connecteur 47
équivalence 47
implication 47
propositionnelle 46
logiquement équivalentes
propositions — 48
logistique
application — 570
équation différentielle — 608
suite — 526
loi(s)
binomiale 970
de Bernoulli 969
de composition interne
notation additive 26
notation infixe 24
notation multiplicative 26
de groupe 118
de Morgan 10
des grands nombres 991
hypergéométrique 975
induite 25, 117
multinomiale 985
produit 25, 117
longueur d'une courbe 486, 499
Lorentz (transformation de —) 708

M

Mac-Laurin (formule de —) 652
Machin (formule de —) 709
Madhava-Gregory (formule de —) 709
majorant 34
majore 33
strictement 33
majorée (partie —) 34
marginales (formule des —) 984
matérielle (implication —) 48
matrice(s) 207
antisymétrique 216
bordante 362
carrée d'ordre n 216
colonne 207
d'échange 228
d'un endomorphisme 347
d'un système linéaire 378
d'une application linéaire 343
d'une famille de vecteurs 342
de dilatation 229
de passage 349
de Sylvester 326
de transvection 228
de type (m, n) 207
de Vandermonde 239
diagonale 216
échelonnée (en lignes, en colonnes) 210
élémentaire 209
équivalentes 220
identité d'ordre n 214
inversible 219
jacobienne 830, 831
ligne 207
méthode de multiplication de Strassen 422
nilpotente 218
régulière 219
scalaire 214
symétrique 216
transposée 215
triangulaire inférieure 216
triangulaire inférieure stricte 216
triangulaire supérieure 216
triangulaire supérieure stricte 216
unipotente 218
maximal (élément —) 34
maximum 34, 825, 844
global 647
local 647
strict 647
médiane 922
Mersenne (nombres de —) 84

méthode
de Ferrari 277
de Newton-Cotes 894
de quadrature 893
de Simpson 899
de variation de la constante 800, 805, 814
des rectangles 894
des rectangles médians 895
des trapèzes 897
mineur (d'une matrice) 358
minimal (élément —) 34
minimum 34, 825, 844
global 647
local 647
strict 647
Minkowski (inégalité de —) 675
minorant 34
minore 33
strictement 33
minorée (partie —) 34
mise à jour (d'une variable) 404
mixte (produit —) 367, 446
Möbius
fonction de — 158
formule d'inversion de — 158
modalité 916
mode 922
modèle des coûts bilinéaires 421, 422
module
d'un nombre complexe 250
d'une subdivision 740
Moivre (formule de —) 267, 689
monogène (groupe) 122
monôme 284
monotone
application — 35
fonction — 628
suite — 35, 556
morphisme 27
canonique 131
d'anneaux 140
de groupes 123
de signature 129
trivial 123
moyenne 923
arithmético-géométrique 607
de Cesaro 542
inégalité de la — 750
première formule de la — 753
valeur — 750
muette (lettre ou variable —) 51, 53
multinomiale (loi —) 985
multiple
d'un élément 113
point — 482
multiplication
des entiers naturels 68
des entiers relatifs 94
des nombres réels 597
des nombres rationnels 100
dichotomique des polynômes 419
externe (dans un espace vectoriel) 162

N

naturels (entiers —) 63
négatif 513
négligeable 867
neutre (élément —) 26, 115
Newton
formule de — 134
polynômes de — 287
sommes de — 300
Newton-Cotes (méthode de —) 894
nilpotent(e)
élément — 137
matrice — 218
niveau (courbe de —) 820
noetherien (ensemble —) 37
nombre(s)
algébriques 312
complexe 244, 245
d'inversions d'une permutation 416
d'or 550, 581
décimal 582
de Bernoulli 532, 612, 884, 886
de Catalan 26
de Fermat 85, 146

de Liouville 314, 590, 610
de Mersenne 84
dyadique 110
entier naturel 63
entier relatif 91
parfait 107
premier 83
réel 588
rationnel 99
transcendants 312
univers 616
non contradiction (principe de —) 49
normal (paramétrage —) 488, 489, 500
norme
algébrique d'un nombre complexe 249
d'un vecteur 452, 780
d'une application linéaire 825
notation(s)
de Kronecker 56, 166, 193, 209
de Landau 867, 869
infixe d'une relation binaire 29
noyau
d'un morphisme 123
d'une application linéaire 177
numération en base b 81
n-uple 11
n-uplet 11

O

opérations élémentaires sur les lignes et sur les colonnes d'une matrice 226
opposé (d'un élément) 26, 116
orbite 525
périodique 528
ordinaire (point —) 475
ordonné (ensemble totalement —) 63
ordre 33, 894
d'un élément 125
d'un groupe 118
d'un polynôme 290
de grandeur 392
lexicographique 34
produit 34
strict 33
total 33
orientation d'un espace vectoriel réel 364
origine (d'un repère) 439
orthogonalité 451
orthonormée (base —) 453
osculateur (plan —) 498
ou inclusif des logiciens 48
ouvert 821

P

pair(e) 5
fonction — 620
polynôme — 285
parabole 458
parallèle 430
paramétrage
d'une courbe plane 469
du cercle 695
normal 488, 489, 500
paramétrée (courbe —) 470, 497
paramétrique(s) (équation(s) —)
d'un plan 447
d'une droite 443
paresseuse (évaluation) 421
parfait (nombre) 107
partie 7
basique d'un espace vectoriel 192
ensemble des —s 8
entière 582, 624
d'une fraction rationnelle 319
génératrice d'un espace vectoriel 185
imaginaire d'un nombre complexe 245
liée d'un espace vectoriel 190
libre d'un espace vectoriel 187
linéaire 432
négative d'une fonction 620
positive d'une fonction 620
principale 582
réelle d'un nombre complexe 245
régulière
d'un développement limité 871
partielle (dérivée —) 826
partition d'un ensemble 23

Pascal
relation de — 79
triangle de — 79
période 620
périodique (fonction —) 620
permutation
d'un ensemble 77
des colonnes 227
des lignes 226
PGCD 88, 97, 144
pivot
d'une matrice colonne ou d'une matrice ligne 210
dans l'algorithme de Gauß 230
plan
affine 427
d'Argand-Cauchy 247
osculateur 498
tangent 841, 842
vectoriel 171, 334
plus grand élément 34
plus petit élément 34
point
attractif 564
birégulier 477, 493, 499
d'arrêt 478
d'inflexion 475
de rebroussement
de première espèce 475
de seconde espèce 475
fixe 14, 125, 526
image d'un complexe 247
intérieur 520
multiple 482
ordinaire 475
régulier 472, 499
répulsif 564
stationnaire 472
super attractif 564
polaires (coordonnées —) 483
pôle d'une fraction rationnelle 318, 771
polygone régulier 267
polynôme(s)
à coefficients dans un anneau 282
algorithmes sur les — 417
application, fonction associée 296
caractéristique 371, 373
constant 282
coût de l'évaluation d'un — 284
cyclotomiques 309
d'approximation 653, 872
d'interpolation de Lagrange 298
de Newton 287
de Tchebychef 267, 295, 310
degré d'un — 285
dérivé d'un — 301
dissocié 298
division par puissances croissantes 321
étrangers 292
impair 285
multiplication
dichotomique des — 419
nul 282
pair 285
premiers entre eux 292
réductible, irréductible 290
racines d'un — 297
radical d'un — 305
réciproques 611
scindé 298
unitaire 285
polynomial(e)
expression — 281
fonction —, application — 296
population 915
positif 513
PPCM 90, 99
prédécesseur 63
prédicat 50
premier(s)
nombres — entre eux 88
nombre — 83
primitive
d'une fonction 678, 758
racine — de l'unité 260

principal(e)
anneau — 144
argument — 263
détermination — de l'argument 263
équation — 381
idéal — 144
inconnue — 381
principe
d'induction 37
de non contradiction 49
de récurrence 43, 55
du tiers exclu 49
probabilité 944
conditionnelle 955
uniforme 946
problèmes du millénaire (sept —) 50
procédé diagonal de Cantor 24
produit
anneau — 134
cartésien 11
d'une famille d'ensembles 22
d'espaces vectoriels 182
de Cauchy 271
de deux matrices 211
groupe — 118
loi — 117
mixte 367, 446
scalaire 451, 780
vectoriel 367, 446
projecteur 180
projection 199
affine 435
$i^{ème}$ 12
orthogonale 455
prolongement par continuité 631
proportionnels
éléments — d'un espace vectoriel 168
proposition (en logique) 46
propre
sous-espace — 369
valeur — 369, 370
vecteur — 369, 370
puissance
d'un élément 113
d'un entier 69
du continu 44
fonction — 684
Pythagore (théorème de —) 453

Q

quadrature (méthode de —) 893
quadruplet 11
quantificateur 51
existentiel 51
universel 51
quaternions 277, 327
quintuplet 11
quotient 100
dans la division des polynômes 288
euclidien 81, 94

R

racine(s)
carrée d'un nombre complexe 253
d'un polynôme 296, 297
de l'unité 258
fonction(s) — 684
fonction — $n^{ème}$ 684
multiple 297
multiplicité d'une — 297
$n^{èmes}$ d'un nombre complexe 264
ordre d'une — 297
primitive de l'unité 260
relations entre coefficients et — d'un polynôme 299
simple 297
radical d'un polynôme 305
raison
d'une série géométrique 716
d'une suite arithmétique 531
d'une suite géométrique 532
rang
d'un système linéaire 378
d'une application linéaire 338
d'une famille de vecteurs 341
d'une matrice 224

rapport d'une homothétie 173, 434
rationnel (nombre —) 99
rayon de courbure 493, 494, 500
réalisation 941
rebroussement (point de —) 475
réciproque(s)
d'une application 15
image — 17
polynômes — 611
recouvrement d'un ensemble 23
rectangles
méthode des — 894
méthode des — médians 895
rectifiable (courbe —) 486
récurrence
à deux pas 55
construction par — 64
finie 64
forte 56, 64
principe de — 43, 55
simple 55
suite définie par — 785
théorème de — 64
réductible (polynôme —) 290
réelle (partie — d'un nombre complexe) 245
réflexive (relation —) 29
règle
de l'Hôpital 651
de Leibniz 645
pour les polynômes 302
régression linéaire 930
régularité
dans un groupe 119
de l'addition 100
de l'addition des entiers 67
de la multiplication 68, 100
régulier(ère)
courbe — 472
élément — 26, 116
matrice — 219
point — 472, 499
polygone — 267
relatifs (entiers) 91
relation(s)
antisymétrique 29
binaire 28
d'équivalence 29
engendrée par une relation 30
d'ordre 29
total 33
de Chasles 269, 426, 796
entre coefficients et racines d'un polynôme 299
linéaire 167
triviale 167
plus faible 30
plus forte 30
réflexive 29
symétrique 29
transitive 29
repère
affine 439
cartésien 439
de Serret-Frenet 491, 494, 501
mobile 483
origine d'un — 439
répétition
définie 403
indéfinie 402
représentants (ensemble de —) 32
représentation
algébrique d'un nombre complexe 245
cartésienne d'un nombre complexe 245
décimale d'un nombre réel 585
trigonométrique d'un nombre complexe 256, 262
répulsif (point —) 564
résidu d'une fraction rationnelle en un pôle 320
résolution effective 391
reste
d'une série 717
dans la division euclidienne 81, 94
des polynômes 288
de Taylor 653

restriction d'une application 12
résultant 326
rétraction (d'une application injective) 40
réunion 8
d'une famille d'ensembles 21
Riemann
série de — 724
sommes de — 754, 795
Rolle (théorème de —) 648
rotation hyperbolique 708
RSA (méthode —) 148

S

SAT (problème de la satisfiabilité) 50
satisfiable (formule logique —) 49
scalaire 162
matrice — 214
produit — 451
schéma de Horner 285
scindé (polynôme —) 298
section (d'une application surjective) 40
segment(s) 248, 516, 520, 638, 658, 839
emboîtés 516
semi-convergente (série —) 730
séparable (polynôme —) 298
séparation (axiome de —) 7
séquencement 402
série 714
absolument convergente 730
alternée 732
convergente 714
de Riemann 724
divergente 714
géométrique 716
harmonique 719
restes d'une — convergente 717
semi-convergente 730
sommes partielles d'une — 714
Serret-Frenet
formules de — 493, 501
repère de — 491, 494, 501
signature
d'une permutation 129
morphisme de — 129
simplifiable (élément —) 26
Simpson (méthode de —) 899
singleton 5
sinus 256, 273
fonction — 688
fonction — hyperbolique 698
somme(s)
d'idéaux 140
d'une série 714
de deux entiers 66
de Newton 300
de Riemann 754, 795
de sous-groupes 122
des puissances d'entiers
consécutifs 532
des sous-espaces vectoriels 171
directe de deux sous-espaces
vectoriels 196
partielles d'une série 714
source d'une application 12
sous-anneau 138
engendré par une partie 140
sous-corps 153
sous-ensemble 7
sous-espace affine 429
sous-espace vectoriel 168
engendré par une partie 170
engendré une famille 170
sous-groupe(s) 95, 120
engendré par une partie 122
finis de $\mathbb{C}^*$ 260
sous-suite 20, 523
stable
partie — 117
par une application 14
pour une loi de composition 25
sous-espace vectoriel — 368
stationnaire
point — 472
suite — 37, 523

strictement (dé)croissante
application — 35
fonction — 628
suite — 35
strictement monotone
application — 35
fonction — 628
suite — 35
structure
d'espace vectoriel induite 168
de contrôle 395, 402
boucle conditionnelle 395, 402
boucle inconditionnelle 395, 402
boucle pour 395, 402
branchement conditionnel 395, 402
itération 395, 402
séquencement 395, 402
structure affine canonique 427
subdivision 740
à pas constant 740
adaptée à une fonction en escalier 740
successeur 63
suite(s) 20, 64, 522
adjacentes 560
arithmético-géométrique 535
arithmétique 531
bornée 784
comportement asymptotique
d'une — 539
convergente 542
croissante 35, 556
de Fibonacci 108, 550
de Héron 558
décroissante 35, 556
divergente 542
extraite 20, 523
finie 20
géométrique 532
homographique 535
logistique 526
monotone 35, 556
récurrente 785
stationnaire 523
strictement (dé)croissante 35
strictement monotone 35
tronquée 522
vectorielle 779
super attractif (point —) 564
supplémentaires
sous-espaces vectoriels — 198, 455
support
d'une courbe 470, 497
d'une famille 116
fini 116
supremum 635
surface
définie par une équation 820
paramétrée 820
surjection 14
surjective (application) 14
symétrie 201
affine 435
orthogonale 455
symétrique 26
d'un élément 116
différence — 137
groupe — 118, 125
matrice — 216
relation — 29
symétrisable (élément) 116
système
binaire 82
complet d'événements 957
décimal 82
linéaire 377, 440
compatible 379
de Cramer 379
homogène 377

T

table de vérité 47
tableau 401
des inversions d'une permutation 422
taille
d'un échantillon 916
de l'écriture d'un entier en base b 392, 395, 396

tangent (plan —) 842
tangente 691, 841
à une courbe 471, 484, 498
fonction — 691
fonction — hyperbolique 698
tautologie 49
taux d'accroissement 640
Taylor (formule(s) de —) 652, 856
avec reste intégral 857, 858
pour les polynômes 304
Taylor-Lagrange
formule de — 652
inégalité de — 654, 859
théorème de — 652
Taylor-Young (formule de —) 655, 859
Tchebychef (polynômes de —) 267, 295, 310
tendre
vers l'infini 543, 578
vers zéro 540
tente (application —) 529
terme
constant d'un polynôme 284
de degré k d'un polynôme 284
dominant d'un polynôme 285
terminaison d'un algorithme 398, 405
Thalès (théorème de —) 173
théorème
d'inversion locale 837
de Bézout 97
pour les polynômes 292
de Bolzano-Weierstraß 784, 822
de Cantor 14
de Cantor, Schröder, Bernstein 40
de Cauchy 758
de d'Alembert-Gauß 274, 306
de Darboux 642, 673
de Gauß 88
pour les polynômes 293
de Koenig 926
de la base incomplète 440
de Pythagore 453
de Rolle 648
de Taylor-Lagrange 652
de Thalès 173
de Wilson 153, 327
de Zermelo 38
des accroissements finis 794, 839
des gendarmes 628
du rang 440
fondamental
de l'analyse 758
de l'arithmétique des polynômes 293
tiers exclu (principe du —) 49
tiroirs (principe des —) 71
topologie de $\mathbb{C}$ 271
torsion 203, 500, 501
total (ordre —) 33
transcendante (fonction —) 677
transformation (de Lorentz) 708
transitive (relation —) 29
translation 119
transposée d'une matrice 215
transposition 126
transvection
matrice de — 228
sur les colonnes 227
sur les lignes 226
trapèzes (méthode des —) 897
tri
à bulle 128, 414
algorithmes de — 412
coût moyen optimal 413
par sélection 414
triangulaire
inégalité — 517
inégalité — 452
matrice — 216
tribulle 414
trigonométrique
cercle — de $\mathbb{C}$ 257
représentation — d'un nombre complexe 256, 262
triplet 11
trivial(aux)
diviseurs — 144
groupe — 118

troncature 583
tronquée (suite —) 522
type 400

U

uniforme (probabilité —) 946
union de deux ensembles 8
unipotente (matrice —) 218
unitaire (polynôme —) 285
univers (nombre —) 616

V

valeur
  absolue 104, 517
  d'adhérence 563
  d'une fraction rationnelle 317
  de vérité 47
  intermédiaire 634
  propre 369, 370, 462, 807
valuation 87
  d'un polynôme 290
Vandermonde
  déterminant de — 360
  matrice de — 239
variable(s) 394
  affectation d'une — 394, 401
  aleatoire(s) 967
    indépendantes 987
  auxiliaire 394
  contenu, valeur d'une — 394
  dans un prédicat 50
  initialisation d'une — 404
  libre 53
  liée 53
  locale 394
  mise à jour d'une — 404
  muette 51, 53
  propositionnelle 47
variance 925, 979
variation de la constante
  méthode de — 800, 805, 814
variations d'une fonction 668
vecteur 162
  colonne 163
  directeur 443
  image d'un complexe 247
  ligne 163
  propre 369, 370, 462, 807
vectoriel(le)
  addition — 163
  fonction — 779, 789
  hyperplan — 181
  produit — 367, 446
  suite — 779
vide
  ensemble — 5
  famille — 21
vitesse 472
voisinage
  au — de 621, 625

W

Wallis (intégrale de —) 773
Weierstraß
  caractérisation de — 624, 625, 631
Wilson (théorème de —) 153

Y

Young (inégalité de —) 667

Z

Zermelo (théorème de —) 38
zéro d'une fraction rationnelle 318
Zorn (lemme de —) 39