Sous la direction de
Jean-Pierre RAMIS
André WARUSFEL
Xavier BUFF • Emmanuel HALBERSTADT
François MOULIN • Monique RAMIS
Jacques SAULOY

Mathématiques

Tout-en-un pour la Licence

2

2e 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, professeur émérite à l'Institut de Mathématiques de Toulouse (Université Paul Sabatier).

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, membre junior de l'Institut Universitaire de France, professeur à l'Institut de Mathématiques de Toulouse, directeur de l'Institut de Recherches sur l'Enseignement des Mathématiques de Toulouse.

Emmanuel Halberstadt, ancien maître de conférences à l'UPMC (Paris), 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, maître de conférences à l'Institut de Mathématiques de Toulouse.

Illustration de couverture : © Ian 2010 - Fotolia.com

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1er juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique s'est généralisée dans les établissements d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée. Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).

© Dunod, 2007, 2014

5 rue Laromiguière, 75005 Paris
www.dunod.com

ISBN 978-2-10-071392-9

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2° et 3° a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).
Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code 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 (700) 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 v

Avant-propos xv

I Algèbre

I.1 Compléments d'algèbre 3

1 Quotients . . . 3
1.1 Quotient d'un ensemble par une relation d'équivalence . . . 3
1.2 Passage au quotient d'une loi de composition interne . . . 6
1.3 Groupes quotients . . . 9
1.4 Anneaux quotients . . . 14
1.5 Espaces vectoriels quotients . . . 17
2 Anneaux commutatifs . . . 18
2.1 Idéaux . . . 18
2.2 Polynômes sur un anneau commutatif . . . 22
3 Algèbres sur un corps commutatif . . . 26
4 Séries formelles . . . 30
4.1 La K-algèbre des séries formelles . . . 31
4.2 Convergence dans $K[[X]]$ . . . 39
Exercices . . . 49

I.2 Actions de groupes 57

1 Généralités . . . 57
1.1 Définitions et exemples . . . 57
1.2 Espaces affines et actions de groupes . . . 62
2 Orbites, stabilisateurs . . . 65
2.1 Orbites . . . 65
2.2 Points fixes. Stabilisateur . . . 68
2.3 Classes modulo un sous-groupe . . . 71
3 Quelques applications des actions de groupes . . . 75
3.1 Problèmes de classification . . . 75
3.2 Groupes de symétries . . . 75
3.3 Applications au groupe symétrique . . . 78
3.4 Applications aux groupes finis . . . 83
3.5 Une application à la combinatoire . . . 86
Exercices . . . 88

I.3 Algèbre bilinéaire 95

1 Dualité . . . 95
1.1 Formes linéaires et hyperplans . . . 95
1.2 Orthogonalité . . . 97
1.3 Transposition . . . 99
1.4 Dualité en dimension finie . . . 100
2 Applications multilinéaires . . . 104
2.1 Définitions et exemples . . . 104
2.2 Retour sur le déterminant . . . 105
2.3 Formes bilinéaires . . . 108
2.4 Formes bilinéaires symétriques, antisymétriques . . . 111
3 Formes quadratiques . . . 116
3.1 Généralités sur les formes quadratiques . . . 116
3.2 Décomposition LU, décomposition de Gauß . . . 121
4 Formes quadratiques sur un espace vectoriel réel . . . 128
4.1 Formes positives, définies positives — inégalité de Cauchy-Schwarz . . . 129
4.2 Signature d'une forme quadratique réelle, théorème d'inertie de Sylvester . . . 131
4.3 Matrices symétriques réelles définies positives, décomposition de Cholesky . . . 133
Exercices . . . 136

I.4 Espaces préhilbertiens 141

1 Espaces vectoriels préhilbertiens réels, espaces euclidiens . . . 141
1.1 Produit scalaire, norme euclidienne . . . 142
1.2 Orthogonalité, projecteurs orthogonaux, symétries orthogonales . . . 144
1.3 Distance d'un point à un sous-espace de dimension finie, inégalité de Bessel . . . 152
1.4 Orthogonalisation de Gram-Schmidt . . . 155
1.5 Adjoint d'un endomorphisme d'un espace préhilbertien réel . . . 161
1.6 Groupe orthogonal . . . 165
1.7 Endomorphismes symétriques et applications . . . 169
2 Formes sesquilinéaires — Formes hermitiennes . . . 186
2.1 Formes sesquilinéaires . . . 187
2.2 Formes hermitiennes . . . 191
3 Espaces vectoriels préhilbertiens complexes, espaces hermitiens . . . 196
3.1 Produit scalaire, norme hermitienne . . . 196
3.2 Orthogonalité, distance d'un point à un sous-espace de dimension finie, inégalité de Bessel . . . 196
3.3 Adjonction, groupe unitaire, endomorphismes hermitiens . . . 199
3.4 Réduction d'un endomorphisme normal, applications . . . 203
Exercices . . . 206

I.5 Réduction des matrices 213

1 Rappels et compléments . . . 213
1.1 Sommes directes de sous-espaces vectoriels . . . 213
1.2 Calculs matriciels par blocs . . . 217
1.3 Polynômes d'endomorphismes . . . 220
1.4 Trace d'une matrice carrée, d'un endomorphisme . . . 222
2 Diagonalisabilité — Trigonalisabilité . . . 224
2.1 Sous-espaces propres, sous-espaces stables . . . 224

2.2 Théorème de Cayley-Hamilton . . . 230
2.3 Endomorphismes et matrices diagonalisables . . . 231
2.4 Endomorphismes et matrices trigonalisables . . . 238
3 Réduction : résultats généraux . . . 242
3.1 Sous-espaces caractéristiques . . . 242
3.2 Endomorphismes nilpotents, matrices nilpotentes . . . 246
3.3 Décomposition de Dunford . . . 249
3.4 Réduction de Jordan . . . 253
Exercices . . . 255

I.6 Groupes classiques 261

1 Le groupe linéaire $GL_n(K)$ . . . 261
1.1 Propriétés algébriques . . . 262
1.2 Propriétés géométriques . . . 265
1.3 Propriétés topologiques et différentielles . . . 272
2 Groupe orthogonal, groupe unitaire . . . 280
2.1 Propriétés algébriques . . . 280
2.2 Propriétés géométriques . . . 283
2.3 Propriétés topologiques et différentielles . . . 286
2.4 Les groupes $O_3(\mathbb{R})$ et $SO_3(\mathbb{R})$ . . . 289
3 Homographies . . . 296
3.1 Le groupe projectif et la droite projective . . . 296
3.2 Le groupe spécial projectif réel et le demi-plan de Poincaré . . . 301
Exercices . . . 305

I.7 Polynômes à plusieurs indéterminées 313

1 Calcul dans $K[X_1, \ldots, X_n]$ . . . 313
1.1 Construction de $K[X_1, \ldots, X_n]$ . . . 313
1.2 Règles de calcul . . . 318
1.3 Dérivations . . . 324
1.4 Polynômes symétriques . . . 329
2 Fonctions polynomiales . . . 335
2.1 Prolongement des identités algébriques . . . 335
2.2 Résultant et discriminant . . . 338
Exercices . . . 344

I.8 Structures discrètes et récursivité 353

1 Mots et langages . . . 353
1.1 L'algèbre des mots . . . 355
1.2 Langages . . . 357
2 Graphes . . . 362
2.1 Graphes non orientés . . . 363
2.2 Graphes orientés . . . 367
2.3 Arbres . . . 372
3 Récursion . . . 375
3.1 Exemples expliqués . . . 376
3.2 Analyse d'algorithmes récursifs . . . 380
3.3 Récursion et induction structurelle . . . 382
Exercices . . . 383

II Analyse

II.1 Espaces vectoriels normés 393

1 Espaces vectoriels normés, espaces métriques . . . 393
1.1 Normes . . . 394
1.2 Espaces métriques . . . 396
1.3 Limites de suites dans un espace métrique . . . 401
1.4 Parties ouvertes, parties fermées . . . 403
1.5 Applications continues . . . 408
1.6 Applications (multi)linéaires continues . . . 415
1.7 Normes équivalentes . . . 421
2 Espaces métriques complets . . . 422
2.1 Suites de Cauchy — Espaces complets . . . 422
2.2 Théorème du point fixe . . . 427
2.3 Séries dans un espace vectoriel normé . . . 429
3 Espaces métriques compacts . . . 432
3.1 Définition et premières propriétés . . . 432
3.2 Fonctions continues sur un compact . . . 435
4 Espaces vectoriels normés de dimension finie . . . 440
4.1 Théorèmes fondamentaux . . . 440
4.2 Normes matricielles . . . 443
5 Connexité. Convexité . . . 444
5.1 Parties convexes . . . 444
5.2 Espaces connexes . . . 448
5.3 Fonctions convexes . . . 452
5.4 Inégalités de convexité . . . 457
6 Espaces de Hilbert . . . 460
6.1 Théorème de projection . . . 460
6.2 Bases hilbertiennes . . . 463
Exercices . . . 467

II.2 Suites et séries de fonctions 475

1 Suites de fonctions . . . 475
1.1 Convergence simple, convergence uniforme . . . 475
1.2 Convergence uniforme et continuité . . . 479
1.3 Intégration et dérivation . . . 483
1.4 Théorème d'approximation de Weierstraß . . . 487
2 Séries de fonctions . . . 489
2.1 Passage des suites de fonctions aux séries de fonctions . . . 489
2.2 Convergence normale . . . 492
3 Séries entières . . . 500
3.1 Domaine de convergence . . . 501
3.2 Opérations algébriques sur les séries entières . . . 509
3.3 Dérivation terme à terme d'une série entière . . . 514
3.4 Fonctions développables en série entière . . . 521
3.5 Exponentielle d'une matrice carrée . . . 529
Exercices . . . 534

II.3 Intégration 543

1 Intégrale de Riemann d'une fonction réelle sur un segment 549
1.1 Subdivisions d'un segment 549
1.2 Applications en escalier à valeurs dans un espace vectoriel 550
1.3 Fonctions intégrables, définition de l'intégrale et propriétés 552
2 Sommes de Darboux et de Riemann 570
2.1 Sommes de Darboux 570
2.2 Sommes de Riemann d'une fonction 577
3 Intégrale de Riemann d'une application à valeurs dans un espace vectoriel de dimension finie 580
3.1 Définitions 580
3.2 Propriétés 581
3.3 Sommes de Riemann 582
4 Intégrale de Riemann d'une application à valeurs dans un espace de Banach 584
4.1 Applications en escalier 584
4.2 Applications intégrables au sens de Riemann 585
4.3 Intégrale d'une application intégrable 585
4.4 Sommes de Riemann d'une application à valeurs dans un espace de Banach . 590
5 Applications réglées 590
5.1 Définition et premières propriétés 590
5.2 Intégrabilité des fonctions réglées 594
5.3 Caractérisation des fonctions intégrables au sens de Riemann 596
6 Propriétés de l'intégrale fonction de sa borne supérieure 598
6.1 Compléments sur la dérivation 599
6.2 L'application intégrale 600
6.3 Le théorème des accroissements finis 602
6.4 Primitives et théorème fondamental 605
6.5 Changement de variables, intégration par parties 607
7 Intégration sur un intervalle quelconque 611
7.1 Intégrales impropres au sens de Riemann 611
7.2 Changement de variables, intégration par parties 622
7.3 Intégration des relations de comparaison 624
8 Théorèmes de convergence 626
8.1 Interversion des limites et des intégrales 626
8.2 Continuité et dérivabilité des fonctions définies par une intégrale 642
Exercices 649

II.4 Séries de Fourier 659

1 Coefficients de Fourier 660
1.1 La famille des exponentielles 661
1.2 La famille des sinus et cosinus 664
1.3 Propriétés des coefficients de Fourier 665
1.4 Taille des coefficients de Fourier 666
1.5 Problème inverse 668
2 Convergence de la série de Fourier d'une fonction continue 670
2.1 Produit de convolution 670
2.2 Formule de Parseval 672

2.3 Théorème de Dirichlet . . . 674
2.4 Théorème de Féjer . . . 676
2.5 Une fonction continue qui n'est pas la somme de sa série de Fourier . . . 679
3 Série de Fourier d'une fonction réglée . . . 680
3.1 Coefficients de Fourier, formule de Parseval . . . 681
3.2 Théorèmes de Dirichlet et de Féjer . . . 685
3.3 Phénomène de Gibbs . . . 688
4 Quelques applications . . . 690
4.1 Équation de la chaleur . . . 690
4.2 Théorème ergodique . . . 693
4.3 Inégalité isopérimétrique . . . 695
4.4 Séries lacunaires . . . 696
Exercices . . . 699

II.5 Fonctions de plusieurs variables 703

1 Théorème d'inversion locale – Théorème des fonctions implicites . . . 704
1.1 Rappels . . . 704
1.2 Énoncés des théorèmes . . . 705
1.3 Exemples . . . 709
1.4 Démonstrations des théorèmes . . . 714
2 Sous-variétés de $\mathbb{R}^n$ . . . 718
2.1 Définition et exemples . . . 718
2.2 Sous-variétés paramétrées . . . 721
2.3 Espaces tangents . . . 724
3 Développement de Taylor . . . 728
3.1 Fonctions de classe $\mathcal{C}^k$ . . . 730
3.2 Les formules de Taylor . . . 735
3.3 Classification des points critiques d'une fonction . . . 739
4 Calcul différentiel . . . 743
4.1 Intégrales curvilignes . . . 743
4.2 Dérivée d'une forme différentielle de degré 1 . . . 747
4.3 Formule de Green-Riemann . . . 750
4.4 Quelques explications . . . 754
Exercices . . . 756

II.6 Fonctions analytiques 761

1 Définition et premières propriétés . . . 761
1.1 Retour sur les séries entières . . . 762
1.2 Fonctions analytiques . . . 766
2 Principe du prolongement analytique $-$ Principe des zéros isolés . . . 768
2.1 Principe du prolongement analytique . . . 769
2.2 Principe des zéros isolés . . . 771
3 Logarithme complexe et racines . . . 774
3.1 Logarithme complexe . . . 774
3.2 Racines $k^{èmes}$ . . . 777
3.3 Théorème d'inversion locale . . . 779
3.4 Analyticité du logarithme complexe et des racines $k^{èmes}$ . . . 782
4 Séries entières et séries de Fourier . . . 783
4.1 Formule de Cauchy pour un cercle . . . 783

4.2 Analyticité des fonctions holomorphes . . . 785
4.3 Inégalités de Cauchy . . . 787
4.4 Principe du maximum . . . 789
4.5 Suites de fonctions analytiques . . . 791
Exercices . . . 793

II.7 Équations différentielles 799

1 Équations différentielles linéaires sur $\mathbb{R}$ . . . 800
1.1 Généralités . . . 804
1.2 Équations à coefficients constants . . . 811
1.3 Vectorialisation et existence des solutions globales . . . 819
2 Existence et comportement des solutions . . . 826
2.1 Le théorème de Cauchy-Lipschitz linéaire . . . 826
2.2 Aspects qualitatifs . . . 836
2.3 Équations à coefficients périodiques . . . 846
3 Équations différentielles non linéaires . . . 850
3.1 Généralités . . . 850
3.2 Existence et comportement des solutions . . . 858
3.3 Systèmes différentiels autonomes . . . 865
Exercices . . . 875

II.8 Méthodes numériques 881

1 Calcul numérique . . . 881
1.1 Rappels et compléments sur le calcul en base b . . . 882
1.2 Calcul en représentation flottante . . . 884
1.3 Performances des méthodes itératives . . . 889
1.4 La formule de Stirling . . . 890
2 Résolution approchée de l'équation $f(x) = 0$ . . . 892
2.1 Principes généraux . . . 892
2.2 Les principales méthodes . . . 898
3 Interpolation et approximation polynomiales . . . 905
3.1 Interpolation polynomiale . . . 905
3.2 Approximation polynomiale . . . 909
4 Calcul approché d'intégrales . . . 914
4.1 Première approche . . . 916
4.2 Méthodes de quadratures . . . 919
4.3 Méthode de Gauß . . . 929
5 Analyse matricielle . . . 931
5.1 Méthodes directes . . . 932
5.2 Méthodes itératives . . . 941
Exercices . . . 949

Bibliographie 955

Indications 956

Index 989

Avant-propos

Cet ouvrage est le deuxième d'une série de trois, conçue pour couvrir les programmes de mathématiques de la plupart des Licences scientifiques.
Dans cette nouvelle édition nous avons, pour certains sujets (en particulier l'intégration), choisi une pédagogie plus progressive. Nous avons ainsi tenu compte, d'une part, de la mise en place des nouveaux programmes de l'enseignement secondaire et des modifications corrélatives des enseignements universitaires et, d'autre part, des remarques de nos lecteurs et de nos collègues enseignants. Cette présentation étant plus détaillée, certains sujets habituellement enseignés au niveau de la deuxième année de licence sont maintenant traités dans le troisième volume de la série, qui couvre par ailleurs un « tronc commun » des programmes de mathématiques au niveau de la troisième année.

Ce cours est illustré d'exemples et applications, il propose de plus au fil du texte de nombreux exercices corrigés qui permettront à l'étudiant de s'entraîner au fur et à mesure de son apprentissage, des notices historiques et un index très complet. On trouvera aussi à la fin de chaque « module » des exercices supplémentaires[1] avec des indications de solutions. Une correction détaillée d'une grande partie de ces exercices est accessible sur le site de l'éditeur.

Nos livres sont conçus comme une aide à l'enseignement oral dispensé par nos collègues dans les cours et travaux dirigés. L'ordre de lecture n'est pas complètement 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 aussi 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 dépendant de la filière choisie, 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 exagérément sur les ordinateurs et les logiciels, à « se débrouiller » devant un problème abstrait ou issu des applications. Nous avons, à cette fin, mis en place une approche adaptée. Nous insistons aussi sur les exigences de rigueur (définitions précises, démonstrations rigoureuses), mais les choses sont mises

[1] Les plus difficiles sont marqués d'une ou deux étoiles.

en place de façon progressive et pragmatique, et nous proposons des exemples riches, dont 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, ni 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. De nos jours, leurs interactions se poursuivent vigoureusement (avec la physique, l'informatique, la mécanique, la chimie, la biologie, l'économie...). Nous souhaitons accompagner ce mouvement au niveau de l'enseignement des premières années d'université et aider à la mise en place, ici ou là, de filières scientifiques pluridisciplinaires contenant une composante mathématique pure ou appliquée. En particulier nous avons introduit de solides initiations aux probabilités et statistique ainsi qu'à l'algorithmique.

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 d'entre nous pour correction lors d'un nouveau tirage, de nous en excuser.

Jean-Pierre Ramis, André Warusfel

Vous pouvez accéder aux corrigés des exercices supplémentaires à partir de la page de présentation de l'ouvrage sur le site de l'éditeur www.dunod.com*. Les corrigés sont au format pdf et permettent une recherche classique par mots clef. Ils peuvent être lus, enregistrés ou imprimés en partie comme en totalité.*

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 XVII^{e} siècle (Fermat), au XVIII^{e} (Euler) et au XIX^{e} (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 XX^{e} 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 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 XVI^{e} 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 XIX^{e}, 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 XIX^{e} siècle, a développé des techniques « effectives » d'algèbre linéaire pour l'astronomie (trajectoire de Ceres en 1801). Au XX^{e} 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...).

Ce cours de seconde année approndit nettement deux aspects fondamentaux de l'algèbre, d'une part l'algèbre générale en systématisant les structures quotients (par exemple à propos des anneaux et des polynômes), et d'autre part en traitant l'algèbre linéaire et bilinéaire qui constitue le fonds classique d'une seconde année d'université. Il faut également signaler un important module consacré à la récursivité, complétant l'initiation à l'algorithmique du premier volume.

Compléments d'algèbre

I.1

Nous approfondissons ici, en vue d'applications dans ce volume, les notions d'algèbre de base de [L1] concernant les groupes, les anneaux, les espaces vectoriels et les polynômes.

1 Quotients

Cette section fait appel au vocabulaire de la section « Relations d'équivalence » du module « Fondements » de [L1], qu'il est donc nécessaire de bien maîtriser.

1.1 Quotient d'un ensemble par une relation d'équivalence

Soit E un ensemble muni d'une relation d'équivalence $\mathcal{R}$. On notera indifféremment $x\mathcal{R}y$ ou $x \equiv y \pmod{\mathcal{R}}$, ou encore, pour alléger l'écriture, $x \sim y$. Pour certains raisonnements ou calculs portant sur les éléments de E, il n'y a pas lieu de distinguer entre deux éléments équivalents, le résultat obtenu (ou les propriétés étudiées) ne dépendant pas du choix d'un élément particulier dans une classe d'équivalence donnée. Le but de cette section est de se donner des moyens (vocabulaire et méthodes) de raisonner et de calculer sur des objets considérés *à équivalence près*.

Exemple. Pour « vérifier » l'opération $54\,321 \times 6\,789 = 368\,785\,269$, on considère chaque entier naturel n comme « équivalent » à la somme $s(n)$ de ses chiffres. Ainsi, $54\,321 \sim 15 \sim 6$ et $6\,789 \sim 30 \sim 3$. On effectue le calcul sur les équivalents plus simples : $6 \times 3 = 18 \sim 9$. Par ailleurs, $368\,785\,269 \sim 54 \sim 9$. On estime alors que le résultat est plausible : il n'a pas été réfuté par le test dit de la « preuve par neuf ». L'idée sous-jacente est qu'il existe une relation d'équivalence $\sim$ telle que, d'une part, $\forall n \in \mathbb{N}\ ,\ n \sim s(n)$ et que, d'autre part, $a \sim a'$ et $b \sim b' \Rightarrow aa' \sim bb'$. Alors $54\,321 \sim 6$ et $6\,789 \sim 3 \Rightarrow 54\,321 \times 6\,789 \sim 6 \times 3 = 18 \sim 9$. C'est, bien entendu, la relation $\equiv_9$ de congruence modulo 9 qui convient.

Théorème et définition 1. Soit E un ensemble muni d'une relation d'équivalence $\mathcal{R}$, également notée $\sim$. Il existe alors un ensemble $\overline{E}$ et, pour chaque élément x de E, un élément $\overline{x}$ de $\overline{E}$ de telle sorte que l'application $\pi : x \mapsto \overline{x}$ soit surjective de E sur $\overline{E}$ et que l'on ait l'équivalence suivante : $\forall x, y \in E\ ,\ \pi(x) = \pi(y) \Leftrightarrow x \sim y$.
L'ensemble $\overline{E}$ est appelé *ensemble quotient de E par la relation d'équivalence $\mathcal{R}$*, et il est noté $E/\mathcal{R}$. L'application π est appelée *projection canonique* de E sur $\overline{E}$. L'élément $\overline{x}$ de $\overline{E}$ est appelé *classe de x* (dans $\overline{E}$) ; on dit que l'élément x est un *représentant* de $\overline{x}$ (dans E).

Autrement dit, on demande que soient vérifiés les axiomes suivants :
- Deux éléments quelconques de E ont la même classe si, et seulement s'ils sont équivalents : $\forall x, y \in E\ ,\ \overline{x} = \overline{y} \Longleftrightarrow x \sim y$.
- Tout élément de $\overline{E}$ est la classe d'un élément de E ; noter cependant que l'on ne requiert *pas* l'unicité du représentant $x \in E$ de $\overline{x} \in \overline{E}$.

Remarquons d'ailleurs qu'en toute rigueur, c'est l'ensemble $\overline{E}$ *muni de la projection canonique* π que l'on devrait appeler le quotient.

Démonstration. Il y a aux moins deux méthodes pour construire l'ensemble quotient $\overline{E} = E/\mathcal{R}$ et la projection canonique $\pi : E \to \overline{E}$. Dans la première méthode, on choisit, dans chaque classe d'équivalence, un élément arbitraire que l'on note $\overline{x}$. C'est possible grâce à l'axiome du choix (module « Fondements » de [L1]). L'ensemble $\overline{E} := \{\overline{x} \mid x \in E\}$ est donc, dans ce cas, un *ensemble de représentants*.
Dans la deuxième méthode, on pose $\overline{x} := \mathrm{cl}(x)$. L'objet $\overline{x}$ est donc *l'ensemble des éléments de E équivalents à x*, donc un sous-ensemble de E, et un représentant de $\overline{x}$ en est un élément. Dans ce cas : $\overline{E} = E/\mathcal{R} := \{\mathrm{cl}(x) \mid x \in E\} \subset \mathcal{P}(E)$. Pour chacune de ces deux constructions, le lecteur vérifiera sans peine que l'application $\pi : E \to \overline{E}$ est surjective et que $\pi(x) = \pi(y) \Leftrightarrow x \sim y$. ∎

L'avantage de la première construction est que les symboles $\overline{x}$ désignent des objets « concrets ». L'avantage de la deuxième construction est qu'elle ne fait pas jouer un rôle privilégié à un élément particulier de $\mathrm{cl}(x)$, ceux-ci sont vraiment équivalents ! L'inconvénient (surtout psychologique) est que le symbole $\overline{x}$ peut être parfois vu comme désignant un *élément* de $\overline{E}$, parfois comme désignant un *sous-ensemble* de E.

Corollaire 2. Soient $a \in \overline{E}$ et $x \in \pi^{-1}(a)$ un représentant de a dans E. Alors l'ensemble $\pi^{-1}(a)$ de tous les représentants de a dans E est la classe d'équivalence $\mathrm{cl}(x)$.

Démonstration. Notons que x existe bien puisque π est surjective. De plus, on a les équivalences logiques : $y \in \pi^{-1}(a) \Leftrightarrow \pi(y) = a = \pi(x) \Leftrightarrow y \sim x$, de sorte que, par définition de la classe d'équivalence de x, $\pi^{-1}(a) = \{y \in E \mid y \sim x\} = \mathrm{cl}(x)$. ∎

Exemple. Pour la relation $\equiv_9$, on a $\overline{E} = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}, \overline{5}, \overline{6}, \overline{7}, \overline{8}\}$. Selon la première construction, on prend, par exemple, $\overline{n} := n \mod 9$ (reste de la division euclidienne) ;

dans ce cas, $\overline{E} = [\![0, 8]\!]$ et $\overline{54321} = 6$. Selon la seconde construction, $\overline{n}$ est l'ensemble $\{n + 9k \mid k \in \mathbb{Z}\}$. Ainsi, $\overline{54321} = \overline{15} = \overline{6}$ est-il à la fois une sorte de nombre et une partie de $\mathbb{Z}$.

1.1.1 Passage au quotient d'une application

On reprend les notations précédentes E, $\mathcal{R}$, $\sim$ et $\pi : E \to \overline{E}$. Soit de plus $f : E \to F$ une application dont l'ensemble d'arrivée F est quelconque.

Théorème et définition 3 (Théorème de factorisation pour les applications).
Les assertions suivantes sont équivalentes :
(i) Deux éléments équivalents de E ont même image : $x \sim y \Rightarrow f(x) = f(y)$; de manière équivalente, l'application f est constante sur chaque classe $\mathrm{cl}(x)$.
(ii) Il existe $\overline{f} : \overline{E} \to F$ telle que $\overline{f} \circ \pi = f$, c'est-à-dire : $\forall x \in E\ ,\ \overline{f}(\overline{x}) = f(x)$.
L'application $\overline{f}$ est alors unique. Nous dirons que la relation $\mathcal{R}$ est *compatible* avec l'application f, et que l'application $\overline{f}$ a été obtenue à partir de l'application f par *passage au quotient* par la relation $\mathcal{R}$.

Démonstration. S'il existe une application $\overline{f}$ satisfaisant (ii), on a les implications logiques : $x \sim y \Rightarrow \overline{x} = \overline{y} \Rightarrow f(x) = \overline{f}(\overline{x}) = \overline{f}(\overline{y}) = f(y)$.

Supposons réciproquement (i) vérifiée. On définit alors l'application $\overline{f}$ de la manière suivante. Pour tout $a \in \overline{E}$, on choisit un représentant arbitraire $x \in E$ de a et l'on est obligé (pour respecter la condition requise) de poser $\overline{f}(a) = f(x)$, ce qui entraîne d'ailleurs l'unicité de $\overline{f}$. Le point crucial est que *cette définition ne dépend pas du choix du représentant* : pour tout autre représentant y, on a $\overline{x} = a = \overline{y}$, donc $x \sim y$, donc $f(x) = f(y)$.

Par construction, on a alors bien $\forall x \in E\ ,\ \overline{f}(\overline{x}) = f(x)$, autrement dit, $\overline{f} \circ \pi = f$. ∎

Pratiquement, on procède comme suit : soit $a \in \overline{E}$. Pour définir $\overline{f}(a)$, on choisit tout d'abord un représentant arbitraire $x \in E$ de a et l'on pose $\overline{f}(a) := f(x) \in F$; dans un deuxième temps, *il faut vérifier que $f(x)$ ne dépend pas du représentant x choisi*.

Corollaire 4. On suppose E et F respectivement munis des relations d'équivalence $\mathcal{R}$ et $\mathcal{S}$, abrégées $\sim$ et $\equiv$. Soit $f : E \to F$ une application telle que deux éléments équivalents de E ont des images équivalentes : $x \sim y \Rightarrow f(x) \equiv f(y)$. Il existe alors une unique application $\overline{f} : \overline{E} := E/\mathcal{R} \to \overline{F} := F/\mathcal{S}$ telle que :

$$\forall x \in E\ ,\ \overline{f}(\overline{x}) = \overline{f(x)}.$$

Démonstration. On applique le théorème à l'application $x \mapsto \overline{f(x)}$ de E dans $\overline{F}$. ∎

Exemple. Prenons $E = \mathbb{R}$ muni de la relation $\equiv_{2\pi}$ de congruence modulo 2π ; l'ensemble quotient est alors noté $\mathbb{R}/2\pi\mathbb{Z}$ (module sur les nombres complexes de [L1]). La

relation $\equiv_{2\pi}$ est compatible avec l'application $\theta \mapsto e^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$ de $\mathbb{R}$ dans $\mathbb{C}$, car les fonctions cosinus et sinus sont 2π-périodiques. On peut donc en déduire, par passage au quotient, une application $\overline{\theta} \mapsto e^{\mathrm{i}\theta}$ de $\mathbb{R}/2\pi\mathbb{Z}$ dans $\mathbb{C}$. En revanche, l'application $\theta \mapsto e^{\mathrm{i}\theta/2}$ ne passe pas au quotient et l'application $\overline{\theta} \mapsto e^{\mathrm{i}\theta/2}$ de $\mathbb{R}/2\pi\mathbb{Z}$ dans $\mathbb{C}$ *n'est pas définie* : on ne saurait en effet décider si $\overline{0} = \overline{2\pi}$ a pour image $e^{\mathrm{i}0/2} = 1$ ou $e^{\mathrm{i}2\pi/2} = -1$ (voir également l'exercice I.1.1 de la page 50).

Exercice 1.
Soit $f : E \to F$ une application. On définit sur E la relation $x \sim y \Leftrightarrow f(x) = f(y)$. Démontrer que c'est une relation d'équivalence compatible avec l'application f et que l'application $\overline{f} : \overline{E} \to F$ est injective. Elle induit donc une bijection de $\overline{E}$ sur $\operatorname{Im} f$.

Solution. On a vu dans [L1] que $\sim$ est une relation d'équivalence dont les classes d'équivalence sont exactement les ensembles de la forme $f^{-1}(y)$, où $y \in F$. Selon la deuxième construction de $\overline{E}$, on peut identifier l'élément $\overline{x}$ de $\overline{E}$ à $f^{-1}(y)$, où $y = f(x)$; on a alors $\overline{f}(\overline{x}) = y$. L'unique antécédent de y par $\overline{f}$ est donc la classe $\overline{x}$ de n'importe quel antécédent x de y par f.

1.2 Passage au quotient d'une loi de composition interne

On reprend les notations précédentes E, $\mathcal{R}$, $\sim$ et $\pi : E \to \overline{E}$. Soit de plus $\star$ une loi de composition interne sur E. On souhaite munir $\overline{E}$ d'une loi de composition interne $\diamond$ telle que π soit un morphisme, autrement dit : $\forall x, y \in E\ ,\ \overline{x} \diamond \overline{y} = \overline{x \star y}$. Pour que cela soit possible, il est évidemment *nécessaire* que $\overline{x \star y}$ ne change pas si l'on remplace x, y par x', y' tels que $\overline{x} = \overline{x'}$ et $\overline{y} = \overline{y'}$.

Définition 1. On dit que la relation $\sim$ est *compatible avec la loi* $\star$, ou encore que c'est une *congruence* si : $\forall x, y, x', y' \in E\ ,\ x \sim x'$ et $y \sim y' \Longrightarrow x \star y \sim x' \star y'$.

Il suffit en fait de vérifier séparément deux propriétés un peu plus faibles, la *compatibilité à gauche* : $y \sim y' \Rightarrow x \star y \sim x \star y'$, et la *compatibilité à droite* : $x \sim x' \Rightarrow x \star y \sim x' \star y$. En effet, des hypothèses $x \sim x'$ et $y \sim y'$, on peut alors déduire respectivement $x \star y \sim x' \star y$ et $x' \star y \sim x' \star y'$, d'où, par transitivité, $x \star y \sim x' \star y'$. Si la loi $\star$ est commutative, ces deux compatibilités partielles sont bien sûr équivalentes.

Proposition et définition 5. Si la relation $\sim$ est compatible avec la loi $\star$, l'ensemble quotient $\overline{E}$ peut être muni d'une unique loi de composition interne $\diamond$ telle que la projection canonique π est un morphisme de $(E, \star)$ dans $(\overline{E}, \diamond)$. La loi $\diamond$ est appelée *loi quotient* (de la loi $\star$ par la relation d'équivalence $\sim$).

Notons que la conclusion traduit l'équation $\overline{x} \diamond \overline{y} = \overline{x \star y}$.

Démonstration. Soient $a, b \in \overline{E}$ et soient $x, y \in E$ des représentants respectifs de a, b. Pour que π soit un morphisme, $\diamond$ doit être définie par la formule $a \diamond b = \overline{x \star y}$, d'où

l'unicité de la loi $\diamond$. Il reste à voir que cette définition en est bien une, c'est-à-dire qu'elle ne dépend pas du choix des représentants. Si $a = \overline{x} = \overline{x'}$ et $b = \overline{y} = \overline{y'}$, on a $x \sim x'$ et $y \sim y'$, d'où, par hypothèse, $x \star y \sim x' \star y'$ et $\overline{x \star y} = \overline{x' \star y'}$. ∎

Exemple. La relation $\equiv_9$ est compatible avec l'addition et la multiplication dans $\mathbb{Z}$; on peut donc additionner et multiplier des classes. Par exemple, comme $\overline{4} = \overline{13}$ et $\overline{17} = \overline{-1}$,on peut indifféremment calculer $\overline{4}+\overline{17} = \overline{21}$ ou $\overline{13}+\overline{-1} = \overline{12}$ et obtenir le même résultat $\overline{21} = \overline{12}$.

Exercice 2.

Montrer que la relation $\equiv_{2\pi}$ est compatible avec l'addition dans $\mathbb{R}$. Donner un exemple d'addition de classes. Cette relation est-elle compatible avec la multiplication dans $\mathbb{R}$?

Solution. Soient x, x', y, y' des réels tels que $x' \equiv_{2\pi} x$ et $y' \equiv_{2\pi} y$. Il existe donc des entiers relatifs a, b tels que $x' - x = a \times 2\pi$ et $y' - y = b \times 2\pi$, d'où $(x' + y') - (x + y) = (a + b) \times 2\pi$, d'où $x' + y' \equiv_{2\pi} x + y$: la relation $\equiv_{2\pi}$ est bien compatible avec l'addition. On reconnaît dans $(\mathbb{R}/\equiv_{2\pi}, +)$ le groupe $\mathbb{R}/2\pi\mathbb{Z}$ rencontré dans l'étude des nombres complexes. Par exemple, comme $\overline{\pi} = \overline{-\pi}$ et $\overline{3\pi/2} = \overline{-\pi/2}$, on peut indifféremment calculer $\overline{\pi} + \overline{3\pi/2} = \overline{5\pi/2}$ ou $\overline{-\pi} + \overline{-\pi/2} = \overline{-3\pi/2}$ et obtenir le même résultat : $\overline{5\pi/2} = \overline{-3\pi/2}$.

Avec les notations ci-dessus, aucun calcul ne permet de conclure que $x'y' - xy$ est un multiple entier de 2π. De fait, si $x = y = 0$, $x' = a \times 2\pi$ et $y' = b \times 2\pi$, on a $x'y' - xy = (2\pi ab) \times 2\pi$ et $2\pi ab$ n'est en général pas un entier. La relation $\equiv_{2\pi}$ n'est donc pas compatible avec la multiplication. Ainsi, on ne saurait dire ce que vaut $\overline{\pi} \times \overline{3\pi/2}$ ou $\overline{-\pi} \times \overline{-\pi/2}$ car $\overline{3\pi^2/2}$ et $\overline{\pi^2/4}$ ne sont pas égaux.

PROPRIÉTÉS HÉRITÉES PAR LA LOI QUOTIENT. *Avec les mêmes notations, les propriétés qui suivent se déduisent de la surjectivité du morphisme $\pi : (E, \star) \to (\overline{E}, \diamond)$:*

1. *Si $\star$ est commutative (resp. associative), $\diamond$ est commutative (resp. associative).*
2. *Si e est un élément neutre (resp. neutre à gauche, resp. neutre à droite, resp. absorbant) pour $\star$, alors $\overline{e}$ est un élément neutre (resp. neutre à gauche, resp. neutre à droite, resp. absorbant) pour $\diamond$.*
3. *Si y est inverse de x (resp. inverse à droite, resp. inverse à gauche) pour $\star$, alors $\overline{y}$ est inverse de $\overline{x}$ (resp. inverse à droite, resp. inverse à gauche) pour $\diamond$.*
4. *En particulier, si $\star$ est une loi de groupe, la loi quotient $\diamond$ est une loi de groupe.*
5. *Dans le cas d'une relation d'équivalence compatible avec deux lois $\star$ et $\top$, telle que $\top$ est distributive (resp. à gauche, resp. à droite) par rapport à $\star$, la loi quotient $\overline{\top}$ est distributive (resp. à gauche, resp. à droite) par rapport à $\overline{\star}$.*
6. *La régularité ne s'hérite pas : exercice I.1.3 de la page 50.*

Théorème 6 (Théorème de factorisation pour les morphismes).
Soit $\sim$ une relation d'équivalence sur E compatible avec la loi $\star$. Notons $\overline{E}$ et $\overline{\star}$ le quotient et la loi quotient et $\pi : E \to \overline{E}$ la projection canonique.
Soit par ailleurs $f : (E, \star) \to (F, \diamond)$ un morphisme. Pour que f soit compatible avec $\sim$, il faut, et il suffit, qu'il existe un morphisme $\overline{f} : (\overline{E}, \overline{\star}) \to (F, \diamond)$ tel que $f = \overline{f} \circ \pi$.
Le morphisme $\overline{f}$ est alors unique.

Démonstration. On applique la proposition 3 de la page 5, ce qui fournit l'application $\overline{f}$ (qui est unique). Il reste à voir que f est un morphisme si, et seulement si, $\overline{f}$ l'est. Dans le sens direct, c'est évident (composé de deux morphismes). Réciproquement, si f est un morphisme, on a, avec des notations évidentes :

$$\overline{f}(\overline{x}\,\overline{y}) = \overline{f}(\overline{xy}) = f(xy) = f(x)f(y) = \overline{f}(\overline{x})\overline{f}(\overline{y}).$$

Comme π est surjective, $\overline{x}$ et $\overline{y}$ sont arbitraires dans $\overline{E}$ et $\overline{f}$ est un morphisme. ∎

Théorème 7 (Théorème d'isomorphisme). (i) Soit $f : (E, \star) \to (F, \diamond)$ un morphisme. La relation $\sim$ sur E définie par $x \sim y \Leftrightarrow f(x) = f(y)$ est une relation d'équivalence compatible avec la loi $\star$. On notera $\overline{E}$ l'ensemble quotient, $\overline{\star}$ la loi induite et $\overline{f} : \overline{E} \to F$ l'application obtenue par passage au quotient.
(ii) Le sous-ensemble $F' = \operatorname{Im} f$ de F est stable pour la loi $\diamond$.
(iii) L'application $\overline{f}$ induit un isomorphisme de $(\overline{E}, \overline{\star})$ sur $(F', \diamond)$.

Démonstration. (i) D'après l'exercice 1 de la page 6, $\sim$ est une relation d'équivalence. Si $x \sim y$ et $x' \sim y'$, alors $f(x) = f(y)$ et $f(x') = f(y')$, donc

$$f(x \star x') = f(x) \diamond f(x') = f(y) \diamond f(y') = f(y \star y'),$$

d'où $x \star x' \sim y \star y'$: on a bien la compatibilité.
(ii) Deux éléments quelconques de F' sont de la forme $f(x), f(y)$ et leur composé est $f(x) \diamond f(y) = f(x \star y) \in F'$: cet ensemble est bien un sous-ensemble stable.
(iii) Le morphisme $\overline{f} : \overline{E} \to F$ induit par corestriction un morphisme $\overline{f} : \overline{E} \to F'$, qui est bijectif d'après l'exercice 1 de la page 6. ∎

Exemple. L'application $t \mapsto e^{it}$ est un morphisme de $(\mathbb{R}, +)$ dans $(\mathbb{C}^*, \times)$ et la relation $\sim$ associée est $\equiv_{2\pi}$. L'image est le cercle unité $\mathbb{U}$ et l'on obtient par passage au quotient un isomorphisme de $(\mathbb{R}/\equiv_{2\pi}, +) = \mathbb{R}/2\pi\mathbb{Z}$ sur $(\mathbb{U}, \times)$. D'après les propriétés d'héritage de la page précédente, il s'agit d'un isomorphisme de groupes.

1.2.1 Retour sur la construction de $\mathbb{Z}$

Soit M un ensemble non vide muni d'une loi de composition interne associative et commutative notée additivement, et disposant d'un élément neutre 0. On dit alors que $(M, +)$ est un *monoïde*

commutatif. Notons additivement la loi produit sur $M \times M$: $(a_1, b_1)+(a_2, b_2) = (a_1+a_2, b_1+b_2)$. Ainsi $(M \times M, +)$ est un monoïde commutatif d'élément neutre $(0, 0)$. Définissons sur $M \times M$ une relation $\sim$ par : $(a, b) \sim (a', b') \Leftrightarrow \exists c \in M \ : \ a + b' + c = a' + b + c$. Par exemple, $(a, a) \sim (0, 0)$ pour tout a. On démontre alors que $\sim$ est une relation d'équivalence compatible avec l'addition de $M \times M$. Notons G l'ensemble quotient et $+$ la loi quotient ; alors $(G, +)$ est un groupe commutatif, dans lequel l'élément neutre est $\overline{(0, 0)}$ et l'opposé de la classe $\overline{(a, b)}$ est $\overline{(b, a)}$. L'application $i : a \mapsto \overline{(a, 0)}$ est un morphisme de $(M, +)$ dans $(G, +)$. Deux éléments $a, b \in M$ ont même image si, et seulement si, $\exists c \in M \ : \ a + c = b + c$. En particulier, si tous les éléments de M sont simplifiables, le morphisme i est injectif. Le groupe G est appelé *groupe des fractions du monoïde* M. Dans le cas du monoïde $(\mathbb{N}, +)$, on obtient ainsi le groupe $G = \mathbb{Z}$ et le morphisme injectif $\mathbb{N} \to \mathbb{Z}$ ([L1], module « Arithmétique »).

1.2.2 Retour sur la construction du corps des fractions

Soient A un anneau commutatif et S une partie *multiplicative* de A, autrement dit : S est stable pour la multiplication et $1 \in S$: ainsi, $(S, \times)$ est (en un sens évident) un sous-monoïde du monoïde commutatif $(A, \times)$. On suppose de plus que $0 \notin S$ (il en résulte d'ailleurs que S ne contient aucun élément nilpotent). On définit sur $A \times S$ deux lois de composition interne, notées additivement et multiplicativement : $(a_1, s_1)+(a_2, s_2) = (a_1 s_2+a_2 s_1, s_1 s_2)$ et $(a_1, s_1) \times (a_2, s_2) = (a_1 a_2, s_1 s_2)$. La deuxième loi est simplement la loi produit définie à partir des monoïdes commutatifs $(A, \times)$ et $(S, \times)$. La première loi, plus compliquée, est modelée sur le calcul des fractions. On définit également sur $A \times S$ une relation $\sim$ comme suit : $(a, s) \sim (a', s') \iff \exists t \in S \ : \ t(as'-a's) = 0$. On démontre alors que $\sim$ est une relation d'équivalence compatible avec les lois $+$ et $\times$ sur $A \times S$. Notons B l'ensemble quotient et $+$ et $\times$ les lois quotients. Alors $(B, +, \times)$ est un anneau commutatif dont les éléments neutres sont $\overline{(0, 1)}$ (pour l'addition) et $\overline{(1, 1)}$ (pour la multiplication). L'opposé de $\overline{(a, s)}$ est $\overline{(-a, s)}$. L'application $i : a \mapsto \overline{(a, 1)}$ est un morphisme d'anneaux de A dans B, de noyau : $\operatorname{Ker} i = \{a \in A \mid \exists s \in S \ : \ sa = 0\}$. Si S ne contient aucun diviseur de zéro dans A, en particulier si A est intègre, le morphisme i est injectif. L'anneau B est noté $S^{-1}A$ et l'on dit que c'est un *anneau de fractions de* A. L'image de $(a, s) \in A \times S$ dans $S^{-1}A$ est notée $s^{-1}a$ ou $\dfrac{a}{s}$.
Lorsque A est intègre, en prenant $S = A \setminus \{0\}$, on reconnaît dans $S^{-1}A$ le corps des fractions de A ([L1], module « Groupes, anneaux, corps »).

1.3 Groupes quotients

1.3.1 Relations compatibles et sous-groupes distingués

Lemme 8. (i) Soient G un groupe et H un sous-groupe de G. Alors la relation $\sim$ définie par $a \sim b \Leftrightarrow a^{-1}b \in H$ (resp. $ab^{-1} \in H$) est une relation d'équivalence sur G compatible à gauche (resp. à droite) avec la loi interne ; la classe d'équivalence de $a \in G$ est sa *classe à gauche* aH (resp. sa *classe à droite* Ha).
(ii) Réciproquement, toutes les relations d'équivalence sur G compatibles à gauche (resp. à droite) avec la loi interne sont obtenues de cette manière.

Démonstration. (i) Nous ne traiterons que le cas des classes à gauche, l'autre côté étant entièrement similaire. De l'équivalence logique $a^{-1}b \in H \Leftrightarrow aH = bH$ (dont la preuve facile est laissée au lecteur), on déduit que $\sim$ est une relation d'équivalence et que la classe

d'équivalence de $a \in G$ est aH. De l'implication évidente $aH = bH \Rightarrow caH = cbH$, on déduit que $\sim$ est compatible à gauche avec la loi interne.
(ii) Soit $\sim$ une relation d'équivalence sur G compatible à gauche avec la loi interne. Soit H la classe de l'élément neutre e. Le sous-ensemble H contient e et l'on a les implications : $a, b \in H \Longrightarrow a \sim e, b \sim e \Longrightarrow a^{-1}b \sim a^{-1}e \sim a^{-1}a \Longrightarrow a^{-1}b \in H$, ce qui entraîne que H est un sous-groupe de G. On a maintenant les équivalences logiques : $a \sim b \Longleftrightarrow a^{-1}a \sim a^{-1}b \Longleftrightarrow a^{-1}b \in H$, ce qui achève la démonstration. ∎

Théorème et définition 9. Les propriétés suivantes sont équivalentes :
(i) Les relations d'équivalence sur G respectivement définies par $a^{-1}b \in H$ et par $ab^{-1} \in H$ coïncident.
(ii) Pour tout $a \in G$, la classe à gauche et la classe à droite coïncident : $aH = Ha$.
(iii) Le sous-groupe H est invariant par les *automorphismes intérieurs* $x \mapsto axa^{-1}$ ($a \in G$) : quel que soit $a \in G$, on a $aHa^{-1} = H$.
Le sous-groupe H de G est alors dit *distingué* dans G. On parle également de sous-groupe *invariant* ou de sous-groupe *normal*. On note alors : $H \lhd G$.

Démonstration. C'est immédiat en vertu du lemme 8 de la page précédente. ∎
Précisons le sens de la troisième propriété. Pour tout $a \in G$ fixé, l'application $x \mapsto axa^{-1}$ est un automorphisme de G ([L1], module « Groupes, anneaux, corps ») ; c'est ce que l'on appelle un automorphisme intérieur. Pratiquement, il suffit de vérifier que, pour tout $a \in G$, on a l'inclusion $aHa^{-1} \subset H$. En effet, l'inclusion correspondante pour a^{-1} s'écrit $a^{-1}Ha \subset H$, qui entraîne $H \subset aHa^{-1}$ et l'on obtient finalement l'égalité. Les sous-groupes G et $\{e\}$ de G sont évidemment distingués. Si G est commutatif, tout automorphisme intérieur est égal à l'identité et tout sous-groupe est distingué.

Exercice 3.
Montrer que le centre Z du groupe G en est un sous-groupe distingué.
Solution. Rappelons ([L1]) que $Z = \{a \in G \mid \forall b \in G\ ,\ ab = ba\}$. Tout élément de Z est fixé par tout automorphisme intérieur, *a fortiori*, le sous-groupe Z est globalement invariant par tout automorphisme intérieur.

Proposition 10. Le noyau $\operatorname{Ker} f$ d'un morphisme de groupes $f : G \to G'$ est distingué. La relation associée est définie par $a \sim b \Leftrightarrow f(a) = f(b)$.

Démonstration.
Si $x \in \operatorname{Ker} f$, alors $f(x) = e'$ (neutre de G'), donc $f(axa^{-1}) = f(a)e'f(a)^{-1} = e'$, donc $axa^{-1} \in \operatorname{Ker} f$: c'est bien un sous-groupe distingué. De plus, on a les équivalences :
$$a^{-1}b \in \operatorname{Ker} f \Longleftrightarrow f(a^{-1}b) = e' \Longleftrightarrow f(a) = f(b).$$ ∎

Corollaire 11. Soit $n \in \mathbb{N}^*$. Le groupe alterné $\mathfrak{A}_n$ est distingué dans le groupe symétrique $\mathfrak{S}_n$. Soit E un ensemble fini non vide. Le groupe alterné $\mathfrak{A}(E)$ est distingué dans le groupe symétrique $\mathfrak{S}(E)$.

Démonstration. C'est le noyau du morphisme signature. ∎

Définition 2. Soit $n \in \mathbb{N}^*$. On appelle *groupe spécial linéaire d'ordre* n, et l'on note $SL_n(K)$ le sous-groupe du groupe linéaire $GL_n(K)$ formé des matrices de déterminant 1. De même, si E est un K-espace vectoriel de dimension finie, on appelle *groupe spécial linéaire de* E, et l'on note $\mathcal{SL}(E)$ le sous-groupe du groupe linéaire $\mathcal{GL}(E)$ formé des automorphismes de déterminant 1.

Corollaire 12. Le groupe spécial linéaire $SL_n(K)$ est un sous-groupe distingué du groupe linéaire $GL_n(K)$. Le groupe spécial linéaire $\mathcal{SL}(E)$ est un sous-groupe distingué du groupe linéaire $\mathcal{GL}(E)$.

Démonstration. C'est, dans chaque cas, le noyau du morphisme déterminant. ∎

1.3.2 Quotient par un sous-groupe distingué

Théorème 13. (i) Soit $H \lhd G$ un sous-groupe distingué. Les égalités $a^{-1}b \in H$ et $ab^{-1} \in H$ définissent une même relation, notée $a \equiv b \pmod{H}$, qui est une relation d'équivalence sur G compatible (à gauche et à droite) avec la loi interne ; la classe d'équivalence de $a \in G$ est $aH = Ha$. Réciproquement, toutes les relations d'équivalence sur G compatibles (à gauche et à droite) avec la loi interne sont obtenues de cette manière à partir d'un sous-groupe distingué H.
(ii) La loi quotient fait de l'ensemble quotient $\overline{G}$ un groupe.
La projection canonique $\pi : G \to \overline{G}$ est un morphisme surjectif de noyau H.

Démonstration.
(i) C'est immédiat à partir du lemme 8 de la page 9 et du théorème 9 de la page ci-contre.
(ii) On sait déjà que la projection canonique est un morphisme surjectif, et, d'après les propriétés d'héritage de la de la page 7, $\overline{G}$ est un groupe. Le noyau de π est la classe de l'élément neutre, c'est-à-dire H. ∎

Définition 3. La relation $a \equiv b \pmod{H}$ est appelée *congruence modulo* H.
Le groupe $\overline{G}$ est appelé *groupe quotient de* G *par* H et noté G/H, ou, parfois $\dfrac{G}{H}$.

Théorème 14 (Théorème de factorisation pour les groupes).
Soient $H \lhd G$ un sous-groupe distingué et $\pi : G \to G/H$ la projection canonique.
Soit $f : G \to G'$ un morphisme de groupes.
Alors, pour que H soit inclus dans $\operatorname{Ker} f$, il faut, et il suffit, qu'il existe un morphisme de groupes : $\overline{f} : G/H \to G'$ tel que $f = \overline{f} \circ \pi$.

Démonstration.
L'inclusion $H \subset \operatorname{Ker} f$ équivaut à l'implication $a^{-1}b \in H \Rightarrow f(a) = f(b)$, c'est-à-dire à la compatibilité de la congruence modulo H avec f : on peut donc appliquer le théorème 6 de la page 8. ∎

Théorème 15 (Premier théorème d'isomorphisme pour les groupes). Soit $f : G \to G'$ un morphisme de groupes. Soit $\pi : G \to G/\operatorname{Ker} f$ la projection canonique. Il existe alors un unique morphisme de groupes $\overline{f} : G/\operatorname{Ker} f \to G'$ tel que $f = \overline{f} \circ \pi$. Ce morphisme est injectif et induit par corestriction un isomorphisme $\overline{f} : G/\operatorname{Ker} f \to \operatorname{Im} f$.

Démonstration. On peut, au choix, appliquer le théorème 7 de la page 8 au cas des groupes ou bien appliquer le théorème 14 de la page précédente au cas de $H = \operatorname{Ker} f$. ∎

Proposition 16. Soient $H \lhd G$ et $\pi : G \to G/H$ la projection canonique. On définit une bijection entre les sous-groupes de G contenant H et les sous-groupes de G/H de la manière suivante : à tout sous-groupe G' de G contenant H, on associe son image $\pi(G') \subset G/H$; à tout sous-groupe K de G/H, on associe son image réciproque $G' = \pi^{-1}(K) \subset G$. Avec ces notations, on a alors : $\overline{G'} := \pi(G') = G'/H$.

Démonstration. L'inclusion $G' \subset \pi^{-1}\big(\pi(G')\big)$ est vraie pour toute partie G' de G. Pour un sous-groupe G' contenant H, on montre l'inclusion réciproque : si $x \in \pi^{-1}\big(\pi(G')\big)$, alors $\pi(x) \in \pi(G')$, donc $\pi(x) = \pi(g)$ avec $g \in G'$; alors $x^{-1}g \in \operatorname{Ker}\pi = H \subset G'$, d'où $x^{-1}g \in G'$, d'où $x \in G'$ (car $g \in G'$). On a donc $\pi^{-1}\big(\pi(G')\big) \subset G'$, d'où $G' = \pi^{-1}\big(\pi(G')\big)$. Puisque π est surjective, l'égalité $F = \pi\big(\pi^{-1}(F)\big)$ est vraie pour toute partie F de G/H. Pour tout sous-groupe G' de G contenant H, $F := \pi(G')$ est un sous-groupe de G/H. Pour tout sous-groupe F de G/H, $G' := \pi^{-1}(F)$ est un sous-groupe de G contenant $\pi^{-1}(e) = H$.
Ainsi, les applications $F \mapsto \pi^{-1}(F)$ et $G' \mapsto \pi(G')$ sont des bijections réciproques l'une de l'autre entre les sous-groupes de G/H et les sous-groupes de G contenant H. La restriction-corestriction de π à $G' \to \overline{G'}$ est surjective de noyau H, donc (théorème 15) induit un isomorphisme $G'/H \simeq \overline{G'}$ qui permet l'identification. ∎

Théorème 17 (Deuxième théorème d'isomorphisme pour les groupes). La correspondance de la proposition 16 induit une bijection entre les sous-groupes distingués de G contenant H et les sous-groupes distingués de G/H. Si $G' \lhd G$, on a de plus un isomorphisme : $\dfrac{G/H}{G'/H} \simeq \dfrac{G}{G'}$.

Démonstration. Si G' est distingué, l'égalité $aG'a^{-1} = G'$ donne, par application du morphisme π, l'égalité $bFb^{-1} = F$, où $F := \pi(G')$ et où $b := \overline{a}$ est un élément quel-

conque de G/H (puisque π est surjective) ; F est donc distingué. Si $F = G'/H$ est distingué dans $\overline{G} = G/H$, le morphisme composé $G \to \overline{G} = G/H \to \overline{G}/F$ est surjectif de noyau G', qui est donc distingué ; on applique alors le premier théorème d'isomorphisme. ∎

Exercice 4.
Quels sont les sous-groupes de $GL_n(K)$ (resp. de $\mathfrak{S}_n$) contenant $SL_n(K)$ (resp. $\mathfrak{A}_n$) ?
Solution. Les premiers sont en bijection avec les sous-groupes de $GL_n(K)/SL_n(K)$. Or, $SL_n(K)$ est le noyau du morphisme surjectif $\det : GL_n(K) \to K^*$. Pour tout sous-groupe F de K^*, on définit donc un sous-groupe $G' := \det^{-1}(F)$, et tous les sous-groupes de $GL_n(K)$ contenant $SL_n(K)$ s'obtiennent de cette façon. Comme K^* est commutatif, ils sont distingués. Un argument analogue s'applique à la seconde question, avec la signature ε à la place du déterminant. Comme $\operatorname{Im}\varepsilon = \{+1, -1\}$ n'admet comme sous-groupes que $\{+1\}$ et lui-même, les seuls sous-groupes de $\mathfrak{S}_n$ contenant $\mathfrak{A}_n$ sont $\mathfrak{S}_n$ et $\mathfrak{A}_n$.

1.3.3 Le cas d'un groupe commutatif

Dans le cas d'un groupe commutatif G, tout sous-groupe H est distingué et l'on peut former le quotient G/H, qui est commutatif d'après les propriétés d'héritage de la de la page 7. Pratiquement tous les quotients que nous rencontrerons cette année sont de ce type : cela concerne en particulier les anneaux et les espaces vectoriels (sections suivantes). Décrivons rapidement le mode d'emploi d'un tel quotient ; on suppose que les groupes sont notés additivement. Lorsque l'on veut additionner deux éléments quelconques a et b de G/H, on en choisit des *représentants*, c'est à dire des éléments x et y dans G tels que $a = \overline{x}$ et $b = \overline{y}$. On a alors $a + b = \overline{x + y}$. Cette somme ne dépend pas du choix des représentants x, y. Nous avons déjà rencontré deux exemples de tels calculs : le calcul des congruences modulo n (il correspond au cas du sous-groupe $n\mathbb{Z}$ de $\mathbb{Z}$; le théorème 18 l'explicite) ; et le calcul des congruences modulo 2π, pour les arguments de nombres complexes (il correspond au cas du sous-groupe $2\pi\mathbb{Z}$ de $\mathbb{R}$, le groupe quotient étant $\mathbb{R}/2\pi\mathbb{Z}$).

Théorème 18. (i) Soit $n \in \mathbb{N}^*$. Les sous-groupes de $\mathbb{Z}/n\mathbb{Z}$ sont exactement les groupes $d\mathbb{Z}/n\mathbb{Z}$, où $d \in \mathbb{N}^*$ est un diviseur de n. Le quotient de $\mathbb{Z}/n\mathbb{Z}$ par $d\mathbb{Z}/n\mathbb{Z}$ est isomorphe à $\mathbb{Z}/d\mathbb{Z}$. Le nombre de ces sous groupes est donc le nombre de diviseurs de n dans $\mathbb{N}^*$.
(ii) Soit G un groupe cyclique d'ordre n. Pour tout diviseur d de n, G admet un unique sous-groupe H d'ordre d, celui-ci est cyclique ainsi que le quotient G/H. Tous les sous-groupes de G s'obtiennent ainsi. Dans le cas où G est le groupe μ_n des racines $n^{èmes}$ de l'unité dans $\mathbb{C}$, le sous-groupe d'ordre d est μ_d.

Démonstration. (i) Puisque tous les sous-groupes de $\mathbb{Z}$ sont de la forme $m\mathbb{Z}$ avec $m \in \mathbb{N}$, les sous-groupes de $\mathbb{Z}/n\mathbb{Z}$ sont les $m\mathbb{Z}/n\mathbb{Z}$, où m est un diviseur de n (proposition 16 de la page 12). Selon le théorème 17 de la page précédente, le quotient de $\mathbb{Z}/n\mathbb{Z}$ par $m\mathbb{Z}/n\mathbb{Z}$ s'identifie à $\mathbb{Z}/m\mathbb{Z}$. L'assertion (ii) est alors immédiate. ∎

1.4 Anneaux quotients

Nous ne considérerons que des anneaux commutatifs.

Théorème 19. (i) Soit I un idéal de l'anneau commutatif A. La congruence modulo le sous-groupe I est compatible avec la multiplication de A. Soit réciproquement $\sim$ une relation d'équivalence sur A compatible avec l'addition et la multiplication. Alors la classe de 0 est un idéal I et $\sim$ est la congruence modulo I.
(ii) La loi quotient de la multiplication de I fait alors du groupe quotient $(A/I, +)$ un anneau commutatif. La projection canonique $\pi : A \to A/I$ est un morphisme surjectif d'anneaux, de noyau I.

On dit que l'image $\overline{a} := \pi(a) \in A/I$ s'obtient par *réduction modulo* I de $a \in A$.

Démonstration. Presque tout cet énoncé fait partie du théorème 13 de la page 11.
(i) Notons $\sim$ la congruence modulo l'idéal I. Alors, si $a \sim b$ et si c est quelconque, $a-b \in I \Rightarrow c(a-b) \in I$ (car I est un idéal), donc $ca \sim cb$, ce qui suffit par commutativité. Si $\sim$ est une relation d'équivalence sur A compatible avec l'addition et la multiplication, on sait déjà que la classe I de 0 est un sous-groupe et que $\sim$ est la congruence modulo I. Reste à voir que c'est un idéal. Si $a \in I$ et $c \in A$, alors $a \sim 0 \Rightarrow ca \sim c0 \Rightarrow ca \in I$.
(ii) Les propriétés d'héritage de la de la page 7 entraînent que $(A/I, +, \times)$ est un anneau commutatif, et il est immédiat que π est un morphisme surjectif d'anneaux, de noyau I. ∎

Exemples. Si l'idéal est trivial : $I = 0$ (notation abusive abrégée pour $I := \{0\}$), alors $A/I = A$. Si $I = A$, le quotient est l'anneau trivial 0 (notation abusive abrégée pour $\{0\}$). Si $A = \mathbb{Z}$ et $I = n\mathbb{Z}$ ($n \in \mathbb{N}^*$), le quotient est l'anneau $\mathbb{Z}/n\mathbb{Z}$.

Exercice 5.
Décrire tous les anneaux quotients de l'anneau $\mathbb{Z}$ et ceux de l'anneau $K[X]$.
Solution. Les idéaux de $\mathbb{Z}$ sont les $n\mathbb{Z}$, où $n \in \mathbb{N}$. Les anneaux quotients de $\mathbb{Z}$ sont donc les anneaux de classes de congruence $\mathbb{Z}/n\mathbb{Z}$ (donc $\mathbb{Z}$ si $n = 0$).
De même, les idéaux de $K[X]$ sont tous principaux et de la forme $<P>$, où P est soit 0 soit un polynôme unitaire. Si $P = 0$, l'idéal est nul et le quotient est $K[X]$. Si $P = 1$, l'idéal est $K[X]$ et le quotient est trivial. Supposons que P est de degré n. De la division euclidienne, on déduit que tout polynôme de $K[X]$ est congru modulo P à un unique polynôme de $K_{n-1}[X]$. Comme on le verra à la section 1.5, de la décomposition $K[X] = <P> \oplus K_{n-1}[X]$ ([L1], module sur les polynômes), on peut déduire l'isomorphisme des *espaces vectoriels*, donc des *groupes* $K[X]/<P>$ et $K_{n-1}[X]$ (proposition 28 de la page 18). Cependant, la multiplication dans $K_{n-1}[X]$ doit alors s'effectuer modulo P : pour multiplier A et B selon la loi quotient, on calcule AB puis on prend le reste $AB \mod P$. Ce type de calcul est illustré dans l'exercice 6 de la page suivante.

Théorème 20 (Théorème de factorisation pour les anneaux). Soient I un idéal de A et $\pi : A \to A/I$ la projection canonique. Soit $f : A \to A'$ un morphisme d'anneaux. Alors, pour que I soit inclus dans $\operatorname{Ker} f$, il faut, et il suffit, qu'il existe un morphisme d'anneaux : $\overline{f} : A/I \to A'$ tel que $f = \overline{f} \circ \pi$.

Démonstration. On invoque le théorème 14 de la page 11 ; il ne reste qu'à vérifier que $\overline{f}$ est bien un morphisme pour la multiplication, ce qui est facile. ∎

Exemple. L'application $P \mapsto P(\mathrm{i})$ de $\mathbb{R}[X]$ dans $\mathbb{C}$ est un morphisme surjectif d'anneaux. Son noyau contient l'idéal $<X^2+1>$. D'après le théorème, on a donc par passage au quotient un morphisme surjectif d'anneaux $\mathbb{R}[X]/<X^2+1> \to \mathbb{C}$.

Corollaire 21. Soit $f : A \to B$ un morphisme d'anneaux et soit J un idéal de B. Alors $f^{-1}(J)$ est un idéal de A et l'on a un morphisme injectif : $A/f^{-1}(J) \to B/J$. En particulier, si $A \subset B$ est un sous-anneau et J un idéal de B, alors $A \cap J$ est un idéal de A et l'on a un morphisme injectif : $A/(A \cap J) \to B/J$.

Démonstration. Le noyau du morphisme composé $A \to B \to B/J$ est l'idéal $f^{-1}(J)$. Si $A \subset B$ et si f est l'inclusion de A dans B, $f^{-1}(J) = A \cap J$. ∎

Théorème 22 (Premier théorème d'isomorphisme pour les anneaux). Soient $f : A \to A'$ un morphisme d'anneaux et $\pi : A \to A/\operatorname{Ker} f$ la projection canonique. Il existe alors un unique morphisme d'anneaux $\overline{f} : A/\operatorname{Ker} f \to A'$ tel que $f = \overline{f} \circ \pi$. Ce morphisme est injectif et induit un isomorphisme de $A/\operatorname{Ker} f$ sur $\operatorname{Im} f$.

Démonstration. On invoque le théorème 15 de la page 12 ; il ne reste qu'à vérifier que $\overline{f}$ est bien un morphisme pour la multiplication, ce qui est facile. ∎

Exercice 6.

Appliquer ce qui précède au morphisme $P \mapsto P(\mathrm{i})$ de $\mathbb{R}[X]$ dans $\mathbb{C}$ et à l'image de $\mathbb{Q}[X]$.

Solution. De la division euclidienne $P = (X^2+1)Q + (a+bX)$, on déduit les équivalences $P(\mathrm{i}) = 0 \Leftrightarrow a + b\mathrm{i} = 0 \Leftrightarrow a = b = 0$. Le noyau est donc l'idéal $<X^2+1>$, d'où un isomorphisme $\mathbb{R}[X]/<X^2+1> \simeq \mathbb{C}$. On peut donc décrire $\mathbb{C}$ comme le groupe $\mathbb{R}_1[X] = \mathbb{R} + \mathbb{R}X$, muni de la multiplication pour laquelle le produit de $a+bX$ et de $c+dX$ est le reste de la division de $(a+bX)(c+dX)$ modulo X^2+1. On vérifie sans peine que ce reste vaut $(ac-bd)+(ad+bc)X$. C'est donc la classe de X modulo X^2+1 qui joue le rôle de i. L'image de $\mathbb{Q}[X]$ est évidemment $\mathbb{Q}[\mathrm{i}]$.

En prenant, dans le corollaire 21, $A := \mathbb{Q}[X]$, $B := \mathbb{R}[X]$ et $J := (X^2+1)\mathbb{R}[X]$, on trouve $A \cap J = (X^2+1)\mathbb{Q}[X]$ (pourquoi ?) et un morphisme injectif

$$\mathbb{Q}[X]/<X^2+1> \to \mathbb{R}[X]/<X^2+1>,$$

qui s'identifie à l'inclusion $\mathbb{Q}[\mathrm{i}] \subset \mathbb{C}$.

Théorème 23 (Deuxième théorème d'isomorphisme pour les anneaux). Soient A un anneau, I un idéal de A et $\pi : A \to A/I$ la projection canonique. La correspondance du théorème 17 de la page 12 établit une bijection entre les idéaux de A contenant I et les idéaux de A/I. Si $J \subset A$ est un idéal contenant I et si $\pi(J) = J/I$ est l'idéal de A/I qui lui correspond, on a un isomorphisme d'anneaux : $\dfrac{A/I}{J/I} \simeq \dfrac{A}{J}$.

Démonstration. Il suffit de vérifier que les constructions correspondantes pour les sous-groupes distingués sont bien compatibles avec la multiplication, ce qui est facile. ∎

Exercice 7.

Décrire les idéaux et les anneaux quotients de l'anneau $\mathbb{Z}/n\mathbb{Z}$ (où $n \in \mathbb{N}$). On détaillera le cas particulier $n := 6$. Décrire de même les idéaux et les anneaux quotients de l'anneau $A := K[X]/<P>$ (où $P \in K[X]$ est un polynôme unitaire de degré $n \in \mathbb{N}^*$). On détaillera le cas $n := 2$ (en admettant que la caractéristique de K n'est pas 2).

Solution. Les idéaux de $\mathbb{Z}/n\mathbb{Z}$ sont les $m\mathbb{Z}/n\mathbb{Z}$, où m divise n, et les quotients correspondants sont les anneaux $\mathbb{Z}/m\mathbb{Z}$. Par exemple, les idéaux de $\mathbb{Z}/6\mathbb{Z}$ sont tous ses sous-groupes : $1\mathbb{Z}/6\mathbb{Z} = \mathbb{Z}/6\mathbb{Z}$, $2\mathbb{Z}/6\mathbb{Z}$, $3\mathbb{Z}/6\mathbb{Z}$ et $6\mathbb{Z}/6\mathbb{Z} = \{0\}$; et les anneaux quotients correspondants sont les anneaux $\mathbb{Z}/1\mathbb{Z} = \{0\}$, $\mathbb{Z}/2\mathbb{Z}$, $\mathbb{Z}/3\mathbb{Z}$ et $\mathbb{Z}/6\mathbb{Z}$.

Les idéaux de A sont les $<Q>/<P>$, où Q est un polynôme unitaire qui divise P, et les quotients correspondants sont les $K[X]/<Q>$. Si $P = X^2 + pX + q$, les diviseurs $Q = 1$ et $Q = P$ donnent respectivement les idéaux $K[X]/<P> = A$ et $<P>/<P> = 0$, et les quotients correspondants sont 0 et A. Pour d'éventuels autres idéaux et quotients, une discussion est nécessaire :

1. Si $\Delta := p^2 - 4q$ n'est pas un carré dans K, le polynôme P est irréductible et A n'admet pas d'autre idéal que 0 et lui-même. Nous verrons plus loin qu'en fait, A est un corps (exemple de la page 20).
2. Si $\Delta = 0$, P admet pour seul diviseur unitaire non trivial $X + p/2$, et A admet un seul idéal autre que 0 et lui-même.
3. Si Δ est un carré non nul, $P = (X-a)(X-b)$ avec $a \neq b$ et l'anneau A admet deux idéaux et deux quotients non triviaux. De plus, comme aucun des deux facteurs de P ne divise l'autre, ces deux idéaux ne sont pas contenus l'un dans l'autre. Nous verrons plus loin que A est le produit de deux corps (exercice I.1.25 de la page 53).

1.5 Espaces vectoriels quotients

Théorème et définition 24. Soit W un sous–espace vectoriel d'un K-espace vectoriel V. Le groupe V/W admet une unique structure d'espace vectoriel telle que l'application $\pi : V \to V/W$ soit linéaire. On définit ainsi *l'espace vectoriel quotient* V/W et la *projection canonique* π.

Démonstration. Puisque l'on veut avoir $\forall x \in V\ ,\ \pi(\lambda x) = \lambda\pi(x)$, on doit définir la loi externe sur V/W en posant $\lambda a := \overline{\lambda x}$, où $x \in V$ est un représentant de a. D'après les implications : $x \sim x' \Rightarrow x - x' \in W \Rightarrow \lambda x - \lambda x' \in W \Rightarrow \lambda x \sim \lambda x'$, la relation $\sim$ sur V est *compatible avec la loi externe* et cette définition ne dépend pas du choix du représentant x de a. On vérifie sans peine que l'on obtient bien un espace vectoriel. Les règles de calcul y sont donc les suivantes : $\overline{x} + \overline{y} = \overline{x+y}$ et $\lambda\overline{x} = \overline{\lambda x}$. ∎

Comme dans le cas des anneaux, les principaux théorèmes concernant les quotients de groupes s'étendent facilement au cas des espaces vectoriels. Il s'agit simplement de vérifier chaque fois des compatibilités avec la loi externe, comme nous venons de le faire. Nous ne détaillerons donc pas la démonstration des théorèmes qui suivent, laissant ce soin au lecteur.

Théorème 25 (Théorème de factorisation pour les espaces vectoriels). Soient $W \subset V$ un sous–espace vectoriel et $\pi : V \to V/W$ la projection canonique. Soit $f : V \to V'$ une application linéaire. Alors, pour que W soit inclus dans $\operatorname{Ker} f$, il faut, et il suffit, qu'il existe une application linéaire : $\overline{f} : V/W \to V'$ telle que $f = \overline{f} \circ \pi$.

Ce théorème admet une utile généralisation que nous détaillerons (une fois n'est pas coutume). Soit $f : V \to V'$ une application linéaire et soient $W \subset V$ et $W' \subset V'$ des sous-espaces vectoriels tels que $f(W) \subset W'$. Si $x, x' \in V$ sont deux représentants de $a \in V/W$, on vérifie aisément que $f(x) - f(x') \in W'$. En posant $\overline{f}(\overline{x}) := \overline{f(x)}$, on définit donc sans ambiguïté une application linéaire $\overline{f} : V/W \to V'/W'$ (*application induite par passage au quotient*). En particulier, *si f est un endomorphisme de V et W un sous-espace stable par f, on obtient ainsi un endomorphisme $\overline{f}$ de V/W.*

Théorème 26 (Premier théorème d'isomorphisme pour les espaces vectoriels). Soit $f : V \to V'$ une application linéaire. Le morphisme de groupes $\overline{f} : V/\operatorname{Ker} f \to V'$ (théorème 15 de la page 12) est linéaire. En particulier, les espaces vectoriels $V/\operatorname{Ker} f$ et $\operatorname{Im} f$ sont isomorphes.

Théorème 27 (Deuxième théorème d'isomorphisme pour les espaces vectoriels). Soient $G \subset F \subset E$ des sous-espaces vectoriels. L'isomorphisme de groupes $\dfrac{E/G}{F/G} \simeq \dfrac{E}{F}$ (théorème 17 de la page 12) est linéaire, donc un isomorphisme d'espaces vectoriels.

1.5.1 Quotients et supplémentaires

Proposition 28. Soit W' un supplémentaire de W dans V. Alors la restriction à W' de l'application $\pi : V \to V/W$ est un isomorphisme.

Démonstration. Le noyau de $\pi_{|W'}$ est $W' \cap \operatorname{Ker}\pi = W' \cap W = \{0\}$. Son image est le sous-espace $\operatorname{Im}\pi_{|W'} = \pi(W') = \pi(W) + \pi(W') = \pi(W + W') = \pi(V) = V/W$. ∎

On obtient ainsi une nouvelle preuve du fait que tous les supplémentaires de W sont isomorphes entre eux ([L1]). On peut d'ailleurs vérifier que l'isomorphisme entre deux supplémentaires obtenu par l'une ou l'autre méthode est le même.

Définition 4. On appelle *codimension* du sous–espace vectoriel W de V, et l'on note $\operatorname{codim}_V W$ la dimension (si elle existe) du quotient V/W. On dit alors que W est *de codimension finie dans* V.

La codimension de W est relative à l'espace ambiant V, et il est absolument nécessaire d'indiquer celui-ci dans la notation. Bien sûr, la codimension dépend également du corps de base, mais la notation ci-dessus ne le mentionne pas : il faudra donc si nécessaire indiquer soigneusement le corps de référence pour éviter les ambiguïtés.

De nombreuses propriétés des sous-espaces vectoriels d'un espace vectoriel de dimension finie s'étendent à des sous-espaces vectoriels de codimension finie d'un espace vectoriel quelconque, et cela aura son utilité en analyse. En voici un échantillon. D'après ce qui précède, la codimension de W est égale à la dimension de l'un quelconque de ses supplémentaires ; par exemple, les sous-espaces vectoriels de codimension 1 sont les hyperplans. Soient W_1 et W_2 deux sous-espaces vectoriels de codimension finie de V. Alors $W_1 \cap W_2$ est de codimension finie et l'on a l'inégalité : $\operatorname{codim}_V(W_1 \cap W_2) \leqslant \operatorname{codim}_V W_1 + \operatorname{codim}_V W_2$ (notant π_1, π_2 les projections canoniques de V sur $V/W_1, V/W_2$, le noyau de l'application linéaire $(\pi_1, \pi_2) : V \to V/W_1 \times V/W_2$ est en effet $W_1 \cap W_2$). L'intersection de k hyperplans est donc un sous–espace vectoriel de codimension inférieure ou égale à k. Sa codimension est exactement k si, et seulement si, des formes linéaires définissant ces hyperplans sont linéairement indépendantes (exercice I.1.20 de la page 52).

Proposition 29. Soit $f : V \to V'$ une application linéaire.
(i) Le noyau $\operatorname{Ker} f$ est de codimension finie si, et seulement si, $\operatorname{Im} f$ est de dimension finie et, dans ce cas : $\operatorname{codim}_V(\operatorname{Ker} f) = \dim(\operatorname{Im} f)$.
(ii) Si V' est de dimension finie, $\operatorname{Ker} f$ est de codimension finie et $\operatorname{codim}_V(\operatorname{Ker} f) \leqslant \dim V'$, avec égalité si, et seulement si, f est surjective.

Démonstration. Cela découle du théorème 26 de la page précédente. ∎

2 Anneaux commutatifs

2.1 Idéaux

Nous rappelons ou introduisons ici quelques constructions utiles. Le lecteur pourra s'exercer à leur application à l'aide de l'exercice I.1.22 de la page 52. Rappelons que l'on note $\mathbb{F}_p$ le

corps fini $\mathbb{Z}/p\mathbb{Z}$ ($p \in \mathbb{N}$ premier). Soit A un anneau commutatif.

1. L'idéal engendré par une partie $X \subset A$ est, par définition, le plus petit idéal de A contenant X, c'est-à-dire l'intersection des idéaux de A contenant X. Il est égal à l'ensemble $\{\sum a_k x_k \mid \forall k\ ,\ a_k \in A, x_k \in X\}$ des combinaison linéaires d'éléments de X à coefficients dans A. Par exemple, l'idéal engendré par $\bigcup I_k$ est $\sum I_k$.
2. Soient I et J deux idéaux de A. Alors $I \cap J$ est un idéal de A, ainsi que $I + J := \{i + j \mid i \in I, j \in J\}$. Plus généralement, si (I_k) est une famille d'idéaux, l'intersection $\bigcap I_k$ et la somme $\sum I_k$ sont des idéaux ; cette dernière est l'ensemble des sommes $\sum i_k$ de familles (i_k) à support fini et telles que $\forall k\ ,\ i_k \in I_k$. On note IJ l'idéal engendré par $\{ij \mid i \in I, j \in J\}$ (*produit* des idéaux I et J). On a : $IJ = \{\sum i_k j_k \mid \forall k\ ,\ i_k \in I, j_k \in J\}$. On vérifie facilement que $IJ \subset I \cap J$. La multiplication des idéaux est commutative, associative et admet pour élément neutre A ; elle est distributive par rapport à l'addition. Les puissances d'un idéal forment une suite décroissante : $I^0 = A \supset I^1 = I \supset \cdots \supset I^n \supset \cdots$
3. Si A est un sous-anneau de B et I un idéal de A, l'idéal de B engendré par I est $IB = \{\sum i_k b_k \mid \forall k\ ,\ i_k \in I, b_k \in B\}$.
4. L'image $\overline{J}$ de l'idéal J dans l'anneau quotient $\overline{A} := A/I$ est un idéal de $\overline{A}$. Le morphisme composé $A \to \overline{A} \to \overline{A}/\overline{J}$ est surjectif, de noyau $I + J$, d'où un isomorphisme : $\overline{A}/\overline{J} \simeq A/(I + J)$. Retenons le principe : pour quotienter l'anneau A par $I + J$, on peut quotienter successivement par I puis par (l'image de) J.

Exercice 8.
Quels sont les idéaux d'un corps commutatif K ? Réciproque ?
Solution. Les seuls idéaux d'un corps K sont 0 et K. En effet, soit I un idéal non nul, et soit $x \in I \setminus \{0\}$. Alors $1 = x^{-1}x \in I$; or, le seul idéal d'un anneau qui contienne 1 est l'anneau tout entier. Supposons réciproquement que l'anneau K n'a pour idéaux que 0 et lui-même. Soit $x \in K \setminus \{0\}$. L'idéal principal Kx est non nul, donc égal à K, donc il contient 1, donc x est inversible : l'anneau K est un corps.

2.1.1 Idéaux maximaux et idéaux premiers

Un idéal *propre* de l'anneau commutatif A est un idéal I de A distinct de A. Il revient au même de dire que I ne contient pas 1, ou qu'il ne contient aucun élément inversible. Les idéaux propres I sont ceux tels que l'anneau quotient A/I est non trivial.

Théorème et définition 30. On dit qu'un idéal I de A est *maximal* s'il possède les propriétés équivalentes suivantes :
(i) L'idéal I est maximal pour l'inclusion parmi les idéaux propres de A.
(ii) L'idéal I est propre et, pour tout $x \in A \setminus I$, l'idéal $I + Ax$ est égal à A.
(iii) L'anneau quotient A/I est un corps.

Démonstration. Si I est maximal et si $x \in A \setminus I$, l'idéal $I + Ax$ contient strictement I, donc il est égal à A (par maximalité). Supposons que, pour tout $x \in A \setminus I$, l'idéal $I + Ax$

est égal à A. Soit $\overline{x}$ un élément non nul de l'anneau quotient A/I. Alors $x \in A \setminus I$, et, par hypothèse, il existe $a \in A$ et $i \in I$ tels que $i + ax = 1$, ce qui entraîne que $\overline{ax} = 1$ dans A/I. Ainsi, tout élément non nul de A/I est inversible et cet anneau est un corps. Supposons enfin que A/I est un corps. Ses seuls idéaux sont donc 0 et lui-même. D'après le théorème 23 de la page 16, il y a une bijection entre les idéaux de A contenant I et les idéaux de A/I : le seul idéal propre de A contenant I est donc I lui-même. ∎

Corollaire 31. L'idéal 0 est maximal si, et seulement si, A est un corps. Les idéaux maximaux de A/I sont les $\mathfrak{M}/I$, où $\mathfrak{M}$ est un idéal maximal de A contenant I.

Exemples.

1. Les idéaux de $\mathbb{Z}$ sont les $n\mathbb{Z}$ ($n \in \mathbb{N}$) et $n\mathbb{Z} \subset m\mathbb{Z} \Leftrightarrow m|n$. Les idéaux maximaux de $\mathbb{Z}$ sont donc les $p\mathbb{Z}$, p premier. Les anneaux quotients correspondants sont les corps finis $\mathbb{F}_p$. Soit $m \in \mathbb{N} \setminus \{0, 1\}$. Les idéaux maximaux de $\mathbb{Z}/m\mathbb{Z}$ sont les $p\mathbb{Z}/m\mathbb{Z}$, p premier divisant m. Les quotients correspondants sont les $\mathbb{F}_p$.
2. Les idéaux de $K[X]$ sont de la forme $<P>$ (P unitaire) et $<P> \subset <Q> \Leftrightarrow Q|P$. Les idéaux maximaux de $K[X]$ sont donc les $<P>$, P irréductible. L'anneau quotient $K[X]/<P>$ est alors une extension de K de degré $n := \deg P$, c'est-à-dire un corps L tel que $K \subset L$ et $\dim_K L = n$. Soit Q un polynôme unitaire. Les idéaux maximaux de $K[X]/<Q>$ sont les $<P>/<Q>$, P irréductible divisant Q. Les quotients correspondants sont les $K[X]/<P>$.
3. Soit $A = \mathcal{F}(E, K)$ l'anneau des applications d'un ensemble E non vide dans le corps K. Pour tout $x \in E$, l'application $f \mapsto f(x)$ est un morphisme surjectif d'anneaux de A sur K. Son noyau est l'idéal $\mathfrak{M}_x$ des applications nulles en x. D'après le premier théorème d'isomorphisme, on a : $A/\mathfrak{M}_x \simeq K$, et $\mathfrak{M}_x$ est maximal.

Théorème 32 (Théorème de Krull). Soit A un anneau et soit I un idéal propre de A. Alors I est contenu dans un idéal maximal.

Démonstration. Considérons l'ensemble E dont les éléments sont les idéaux propres de A, ordonné par inclusion. Cet ensemble est inductif ([L1], module « Fondements »). En effet, si (I_k) est une famille totalement ordonnée d'idéaux propres, il est clair que $\bigcup I_k$ est un idéal ; de plus, aucun des I_k ne contient 1, donc $\bigcup I_k$ non plus. D'après le lemme de Zorn, tout élément de E est donc majoré par un élément maximal. ∎

Théorème et définition 33. On appelle *idéal premier* de A un idéal propre I possédant les propriétés équivalentes suivantes :
(i) On a l'implication : $\forall x, y \in A\ ,\ xy \in I \Rightarrow x \in I \quad \text{ou} \quad y \in I$.
(ii) L'anneau quotient A/I est intègre.

Démonstration. Notons $\overline{x}$ la classe de $x \in A$ dans $\overline{A} := A/I$, alors $\overline{x} = 0 \Leftrightarrow x \in I$; et tous les éléments de $\overline{A}$ sont de la forme $\overline{x}$, $x \in A$. L'implication de la définition se réécrit donc : $\forall \overline{x}, \overline{y} \in \overline{A}\ ,\ \overline{x}\,\overline{y} = 0 \Rightarrow \overline{x} = 0 \quad \text{ou} \quad \overline{y} = 0$, ce qui caractérise l'intégrité de $\overline{A}$. ∎

Exemples. Les idéaux premiers de $\mathbb{Z}$ sont 0 et les $p\mathbb{Z}$, p premier. Les idéaux premiers de $K[X]$ sont 0 et les $<P>$, P irréductible.

Corollaire 34. Tout idéal maximal est premier. Tout anneau non trivial admet donc des idéaux premiers. L'idéal 0 de A est premier si, et seulement si, A est intègre. Les idéaux premiers de A/I sont les idéaux $\mathcal{P}/I$, où $\mathcal{P}$ est un idéal premier de A.

Exercice 9.

Montrer que les éléments nilpotents de A forment un idéal et que celui-ci est contenu dans l'intersection des idéaux premiers de A.

Solution. Si x est nilpotent, il est évident que ax l'est quel que soit a (car, l'anneau étant commutatif, on a $(ax)^n = a^n x^n$). D'autre part, si $x^n = y^m = 0$, on déduit de la formule du binôme de Newton que $(x+y)^{m+n-1} = 0$ (dans chacun des termes $x^i y^j$ avec $i+j = m+n-1$, on a $i \geqslant n$ ou $j \geqslant m$). On a donc bien un idéal (appelé *radical* ou *nilradical* de A). Si $x^m = 0$ et si $\mathcal{P}$ est un idéal premier, alors $\overline{x}$ est nilpotent dans l'anneau intègre $A/\mathcal{P}$, donc nul, et $x \in \mathcal{P}$. (Voir également l'exercice I.1.31 de la page 54.)

Proposition 35. Soient $f : A \to B$ un morphisme d'anneaux et $\mathcal{Q} \subset B$ un idéal premier. Alors $f^{-1}(\mathcal{Q})$ est un idéal premier de A. Si par exemple $A \subset B$, pour tout idéal premier $\mathcal{Q}$ de B, l'idéal $\mathcal{P} := \mathcal{Q} \cap A$ est premier.

Démonstration. L'idéal $\mathcal{P} := f^{-1}(\mathcal{Q})$ est le noyau du morphisme composé $A \to B \to B/\mathcal{Q}$, d'où une injection de $A/\mathcal{P}$ dans l'anneau intègre $B/\mathcal{Q}$ ∎

La proposition ne se transpose pas aux idéaux maximaux : 0 est un idéal maximal de $\mathbb{Q}$, mais son intersection avec $\mathbb{Z}$ n'est pas un idéal maximal de $\mathbb{Z}$.

2.1.2 Idéaux étrangers, théorème chinois

Définition 5. On dit que les idéaux I et J sont *étrangers* si $I + J = A$, autrement dit, s'il existe $i \in I$ et $j \in J$ tels que $i + j = 1$. On dit aussi que I *est étranger à* J.

Exemples.

1. Tout idéal est étranger à A. Le seul idéal étranger à lui-même est A. Un idéal maximal est étranger à tout idéal qu'il ne contient pas ; deux idéaux maximaux distincts sont étrangers.
2. Dans l'anneau $\mathbb{Z}$, les idéaux $m\mathbb{Z}$ et $n\mathbb{Z}$ sont étrangers si, et seulement si, m et n sont premiers entre eux (théorème de Bézout dans $\mathbb{Z}$).
3. Dans l'anneau $K[X]$, les idéaux $<P>$ et $<Q>$ sont étrangers si, et seulement si, P et Q sont premiers entre eux (théorème de Bézout dans $K[X]$).

Proposition 36. Si I est étranger à J et à K, alors I est étranger à JK. En particulier, si I et J sont étrangers, I^m et J^n sont étrangers pour $m, n \in \mathbb{N}$.

Démonstration. On a $i_1 + j = 1$ et $i_2 + k = 1$, avec $i_1, i_2 \in I$, $j \in J$ et $k \in K$.
Alors $i + jk = 1$, où $i := i_1 i_2 + i_1 k + i_2 j \in I$ et $jk \in JK$.
La deuxième assertion s'ensuit par récurrence sur m et n. ∎

Théorème 37 (Théorème chinois).
Soient I et J deux idéaux étrangers de l'anneau A.
Le morphisme d'anneaux $x \mapsto (x \mod I, x \mod J)$ de A dans $(A/I) \times (A/J)$ est surjectif de noyau $I \cap J$. On obtient par passage au quotient un isomorphisme : $A/(I \cap J) \to (A/I) \times (A/J)$. On a de plus l'égalité $I \cap J = IJ$.

Démonstration. Soient $i \in I$ et $j \in J$ tels que $i + j = 1$.
Tout élément de $(A/I) \times (A/J)$ est de la forme $(a \mod I, b \mod J)$, où $a, b \in A$. L'élément $x = aj + bi$ est tel que $x - a = (b - a)i \in I$ et $x - b = (a - b)j \in J$, d'où $(x \mod I, x \mod J) = (a \mod I, b \mod J)$, ce qui prouve la surjectivité. L'élément x est dans le noyau si, et seulement si, $(x \mod I, x \mod J) = (0, 0)$, c'est-à-dire si, et seulement si, $x \in I$ et $x \in J$. Le noyau est donc $I \cap J$. On applique alors le premier théorème d'isomorphisme. L'inclusion $IJ \subset I \cap J$ est vraie pour des idéaux quelconques. Soit $x \in I \cap J$; alors $xi \in IJ$ et $xj \in IJ$, d'où $x = xi + xj \in IJ$. ∎

Corollaire 38. Soient $I_1, \ldots, I_k$ des idéaux étrangers deux à deux.
On a alors un isomorphisme : $A/(I_1 \cap \cdots \cap I_k) \to A/I_1 \times \cdots \times A/I_k$ et une égalité : $I_1 \cap \cdots \cap I_k = I_1 \cdots I_k$.

Démonstration. Pour $k = 2$, c'est le théorème ; on passe de $k - 1$ à k en remarquant que I_k est étranger à $I_1 \cap \cdots \cap I_{k-1} = I_1 \cdots I_{k-1}$ (proposition 36). ∎

Réciproquement, si le morphisme $A \to (A/I) \times (A/J)$ est surjectif, soit i un antécédent de $(0, 1)$: alors $i \in I$ et $i - 1 \in J$, d'où $1 \in I + J$, et I et J sont étrangers.

2.2 Polynômes sur un anneau commutatif

On fixe un anneau commutatif A. Soit B une *extension* de A, c'est-à-dire un anneau commutatif dont A est un sous-anneau. Pour tout élément x de B, le *sous-anneau de B engendré par A et x* (ou encore *engendré sur A par x*) est, par définition, le plus petit sous-anneau de B contenant A et x, ou encore l'intersection des sous-anneaux de B contenant A et x. On le note parfois $A[x]$, mais on prendra garde à l'ambiguïté de cette notation (qui sous-entend l'extension B). L'anneau $A[x]$ contient évidemment toutes les expressions de la forme $\sum_{i \geqslant 0} a_i x^i$, où la suite $(a_i)_{i \in \mathbb{N}}$ de $A^{\mathbb{N}}$ est à support fini (*expressions polynomiales en x à coefficients dans A*). Par ailleurs, l'ensemble de ces expressions forme clairement un sous-anneau de B contenant A et x, et l'on a donc : $A[x] = \left\{ \sum_{i \geqslant 0} a_i x^i \;\middle|\; (a_i)_{i \in \mathbb{N}} \in A^{(\mathbb{N})} \right\}$.

Exemple. Dans l'extension $\mathbb{C}$ de $\mathbb{Z}$, le sous-anneau engendré par $\mathbb{Z}$ et i est l'anneau $\mathbb{Z}[\mathrm{i}]$ des entiers de Gauß. Naturellement, un élément de $\mathbb{Z}[\mathrm{i}]$ admet de nombreuses écritures comme expression polynomiale en i à coefficients dans $\mathbb{Z}$: par exemple, $0 = 1 + \mathrm{i}^2 = 1 - \mathrm{i}^4 = \cdots$

Théorème et définition 39. Il existe une extension B de A et un élément X de B tels que tout élément b de B admet une écriture *unique* : $b = \sum_{i\geqslant 0} a_i X^i$, où la suite $(a_i)_{i\in\mathbb{N}}$ de $A^{\mathbb{N}}$ est à support fini. L'anneau B est appelé *anneau des polynômes à coefficients dans A en l'indéterminée X* et on le note $A[X]$. On dit que l'anneau $A[X]$ est *librement engendré sur A par X*.

Démonstration. Nous plagions la construction de l'anneau des polynômes sur un corps. On pose $B = A^{(\mathbb{N})}$, que l'on munit de la loi de groupe produit, qui est commutative, et de la multiplication : $(a_i)_{i\in\mathbb{N}}(b_i)_{i\in\mathbb{N}} = (c_i)_{i\in\mathbb{N}}$, où : $\forall n \in \mathbb{N}$, $c_n := \sum_{i+j=n} a_i b_j$. On a bien un anneau, l'élément neutre de la multiplication étant la suite $(1, 0, \ldots)$. Si l'on pose $X := (0, 1, 0, \ldots)$, on vérifie que $(a_i)_{i\in\mathbb{N}} = \sum_{i\geqslant 0} a_i X^i$, et que c'est bien l'unique écriture de $(a_i)_{i\in\mathbb{N}}$ comme expression polynomiale en X à coefficients dans A. ∎

Lorsque A est un corps, on retrouve la définition de [L1]. On vérifie facilement les faits suivants : si A est un sous-anneau de B, $A[X]$ est un sous-anneau de $B[X]$; si $f : A \to B$ est un morphisme d'anneaux, l'application $\sum_{i\geqslant 0} a_i X^i \mapsto \sum_{i\geqslant 0} f(a_i) X^i$ est un morphisme d'anneaux de $A[X]$ dans $B[X]$.

2.2.1 Règles de calcul

Définition 6. Le *degré* d'un polynôme non nul $P := \sum_{i\geqslant 0} a_i X^i$ est l'entier

$$\deg P := \max\{i \in \mathbb{N} \mid a_i \neq 0\}.$$

Par convention, le degré du polynôme nul est $\deg 0 := -\infty$.

Le *coefficient dominant* et le *terme dominant* du polynôme $P := \sum_{i\geqslant 0} a_i X^i$ de degré $n \geqslant 0$ sont respectivement $\mathrm{cd}(P) := a_n$ et $\mathrm{td}(P) := a_n X^n$.

Pour tous $P, Q \in A[X]$, on a $\deg(P + Q) \leqslant \max(\deg P, \deg Q)$, avec égalité si, et seulement si, les termes dominants ne s'annulent pas, par exemple si $\deg P \neq \deg Q$. En revanche, l'égalité $\deg PQ = \deg P + \deg Q$ ne va pas de soi si A n'est pas intègre :

Exercice 10.

Étudier l'exemple des polynômes $P := \overline{2}X + \overline{1}$ et $Q := \overline{3}X + \overline{1}$ sur $\mathbb{Z}/6\mathbb{Z}$. Montrer que, dans l'anneau $\mathbb{Z}/4\mathbb{Z}[X]$, le polynôme non constant $\overline{1} + \overline{2}X$ est inversible.

Solution. On a $\overline{2} + \overline{3} = -\overline{1}$ et $\overline{2}.\overline{3} = \overline{0}$, d'où $PQ = -X + \overline{1}$, de degré $1 < 1 + 1$.

Dans $\mathbb{Z}/4\mathbb{Z}[X]$, le polynôme $\overline{1} + \overline{2}X$ est son propre inverse car :

$$(\overline{1} + \overline{2}X)^2 = \overline{1} + \overline{4}X + \overline{4}X^2 = \overline{1}.$$

Pour abréger les écritures, nous adopterons fréquemment la convention suivante : si P est un polynôme non nul, la notation $P = aX^m + \cdots$ signifie que $aX^m = \mathrm{td}(P)$; ainsi, $a \neq 0$ et les points $\cdots$ désignent « des termes de degré strictement inférieur ».

Lemme 40. Si $\mathrm{cd}(P)$ ou $\mathrm{cd}(Q)$ n'est pas diviseur de 0 dans A, alors on a les règles : $\deg(PQ) = \deg(P) + \deg(Q)$, $\mathrm{cd}(PQ) = \mathrm{cd}(P)\,\mathrm{cd}(Q)$ et $\mathrm{td}(PQ) = \mathrm{td}(P)\,\mathrm{td}(Q)$.

Démonstration. Écrivons $P = aX^m + \cdots$ et $Q = bX^n + \cdots$ (donc $a, b \neq 0$).
Alors $ab \neq 0$ (hypothèse sur a et b) et $PQ = abX^{m+n} + \cdots$ ∎

Proposition 41.
Supposons l'anneau A intègre ; alors $A[X]$ l'est également. On a les règles :
$\deg(PQ) = \deg(P) + \deg(Q), \mathrm{cd}(PQ) = \mathrm{cd}(P)\,\mathrm{cd}(Q)$ et $\mathrm{td}(PQ) = \mathrm{td}(P)\,\mathrm{td}(Q)$.
En particulier, si un polynôme P divise un polynôme R tel que $\deg R < \deg P$, alors $R = 0$. Les éléments inversibles de $A[X]$ sont les polynômes constants $a \in A^*$ (*i.e.* les éléments inversibles de A).

Démonstration. On peut invoquer, au choix, le lemme ou bien le fait que $A[X]$ est un sous-anneau de $K[X]$, où K est le corps des fractions de A. ∎

Proposition 42. Si L est le corps des fractions de l'anneau commutatif intègre A, alors $L(X)$ est le corps des fractions de $A[X]$.

Démonstration. Puisque $A[X]$ est un sous-anneau du corps $L(X)$, il suffit de voir que tout élément de $L(X)$ est quotient de deux éléments de $A[X]$, ce qui est facile et laissé au lecteur. ∎

Proposition 43.
(i) Soit P un polynôme de terme dominant aX^n ($a \in A \setminus \{0\}$, $n \in \mathbb{N}$). Pour tout polynôme $F \in A[X]$, il existe alors un entier k et des polynômes Q et R tels que :

$$a^k F = PQ + R \quad \text{et} \quad \deg R < \deg P.$$

(ii) Si l'on suppose de plus $a = \mathrm{cd}(P)$ inversible, il existe des polynômes uniques Q et R tels que $F = PQ + R$ et $\deg R < \deg P$.

Démonstration.
(i) Si $\deg F < \deg P$, on peut prendre $k := 0$, $Q := 0$ et $R := F$. Si $\deg F \geqslant \deg P$, on écrit $\mathrm{td}\, F = bX^p$ et l'on introduit le polynôme $F_1 := aF - bX^{p-n}P$, qui est de degré strictement inférieur à p ; on peut alors itérer le procédé. On voit même que k peut être choisi égal à $\deg F - \deg P + 1$.
(ii) L'existence découle de (i) et d'une division par a^k. Supposons $PQ + R = PQ_1 + R_1$, avec $\deg R, \deg R_1 < \deg P$. Alors le lemme 40, appliqué à l'égalité $P(Q - Q_1) = R_1 - R$ entraîne (par contraposée) $Q - Q_1 = R_1 - R = 0$. ∎

Exemple. Prenons $A := \mathbb{Z}$, $P := 2X + 1$ et $F := X^3 + X^2 + X + 1$. Alors :
$F_1 := 2F - X^2P = X^2 + 2X + 2$, $F_2 := 2F_1 - XP = 3X + 4$ et $F_3 := 2F_2 - 3P = 5$.
Finalement :

$$8F = 4X^2P + 4F_1 = 4X^2P + 2XP + 2F_2$$
$$= 4X^2P + 2XP + 3P + F_3 = (4X^2 + 2X + 3)P + 5.$$

2.2.2 Idéaux de $A[X]$

Commençons par une motivation pour l'étude des idéaux de $A[X]$.

Exemple. L'application $\sum_{n\geqslant 0} a_n X^n \mapsto \sum_{n\geqslant 0} a_n \mathrm{i}^n$ est un morphisme surjectif de l'anneau des polynômes $\mathbb{Z}[X]$ sur l'anneau des entiers de Gauß $\mathbb{Z}[\mathrm{i}]$. Déterminons son noyau. Pour tout polynôme $P \in \mathbb{Z}[X]$, on peut écrire $P = (X^2 + 1)Q + (aX + b)$ avec $Q \in \mathbb{Z}[X]$ et $a, b \in \mathbb{Z}$ (proposition 43 de la page ci-contre). Alors P est dans le noyau, *i.e.* $P(\mathrm{i}) = 0$, si, et seulement si, $a\mathrm{i} + b = 0$, c'est-à-dire $a = b = 0$. Le noyau est donc l'idéal $<X^2 + 1>$ de $\mathbb{Z}[X]$, et, d'après le premier théorème d'isomorphisme : $\mathbb{Z}[X]/<X^2 + 1> \simeq \mathbb{Z}[\mathrm{i}]$.

Exercice 11.

Décrire de manière analogue l'extension $\mathbb{Z}[j]$ de $\mathbb{Z}$ engendrée par $j := \dfrac{-1 + \mathrm{i}\sqrt{3}}{2}$.

Solution. L'application $\sum_{n\geqslant 0} a_n X^n \mapsto \sum_{n\geqslant 0} a_n j^n$ est un morphisme surjectif de l'anneau des polynômes $\mathbb{Z}[X]$ sur l'anneau $\mathbb{Z}[j]$. On sait que $j^2 + j + 1 = 0$. On écrit donc un élément quelconque de $\mathbb{Z}[X]$ sous la forme $P = (X^2 + X + 1)Q + (aX + b)$ avec $Q \in \mathbb{Z}[X]$ et $a, b \in \mathbb{Z}$ (proposition 43 de la page précédente), et l'on a $P(j) = 0$, si, et seulement si, $aj + b = 0$, c'est-à-dire $a = b = 0$. Le noyau est donc l'idéal $<X^2 + X + 1>$ de $\mathbb{Z}[X]$, et, d'après le premier théorème d'isomorphisme : $\mathbb{Z}[X]/<X^2 + X + 1> \simeq \mathbb{Z}[j]$.

Exercice 12.

Soit $d \in \mathbb{Z}$ un entier qui n'est pas un carré. Si $d < 0$, on note $\sqrt{d} := \mathrm{i}\sqrt{-d}$.
Décrire l'extension $\mathbb{Z}[\sqrt{d}]$ de $\mathbb{Z}$ comme un quotient de $\mathbb{Z}[X]$.

Solution.

Raisonnement analogue : tout polynôme de $\mathbb{Z}[X]$ s'écrit $P = (X^2 - d)Q + (aX + b)$ avec $Q \in \mathbb{Z}[X]$ et $a, b \in \mathbb{Z}$, et l'on a $P(\sqrt{d}) = 0$, si, et seulement si, $a\sqrt{d} + b = 0$, c'est-à-dire $a = b = 0$. Ce dernier point mérite d'être détaillé. Si $d < 0$, il est évident ; sinon, on observe que, si d était le carré d'un élément de $\mathbb{Q}$, il serait le carré d'un élément de $\mathbb{Z}$ (cela a été démontré dans le [L1], exercices sur les polynômes : « $\mathbb{Z}$ est intégralement clos »). En fin de compte : $\mathbb{Z}[X]/<X^2 - d> \simeq \mathbb{Z}[\sqrt{d}]$.

Proposition 44. (i) Pour tout idéal I de A, l'idéal $IA[X]$ de $A[X]$ engendré par I est $\{\sum i_k X^k \mid \forall k\ ,\ i_k \in I\}$. C'est le noyau du morphisme surjectif $A[X] \to A/I[X]$, d'où un isomorphisme : $A/I[X] \simeq A[X]/IA[X]$.
(ii) L'idéal $IA[X]$ est premier si, et seulement si, l'idéal I est premier.

Démonstration. (i) La preuve est mécanique et laissée à la patience du lecteur.
(ii) Si $A/I[X] \simeq A[X]/IA[X]$ est intègre, son sous-anneau A/I est évidemment intègre ; la réciproque est vraie en vertu de la proposition 41 de la page 24. ∎

Pratiquement, la *réduction modulo* $IA[X]$ *d'un polynôme* $P \in A[X]$ se calcule en réduisant modulo I les coefficients de P.

Exercice 13.
Décrire de deux manières le quotient $\mathbb{Z}[X]/<2, X^2+1>$. L'idéal $<2>$ de l'anneau $\mathbb{Z}[\mathrm{i}]$ est-il premier ?
Solution.
L'anneau $\mathbb{Z}[X]/<2, X^2+1>$ est isomorphe à $\mathbb{F}_2[X]/<X^2+1> = \mathbb{F}_2[X]/<(X+1)^2>$, et aussi à $\mathbb{Z}[\mathrm{i}]/<2>$. Notons x la classe de X dans $\mathbb{F}_2[X]/<(X+1)^2>$. L'élément $x+1$ est nilpotent non nul, et cet anneau n'est donc pas intègre, donc $\mathbb{Z}[\mathrm{i}]/<2>$ non plus. L'idéal $<2>$ de $\mathbb{Z}[\mathrm{i}]$ n'est donc pas premier.

---

3 Algèbres sur un corps commutatif

Définition 7. On suppose V muni des deux structures suivantes : d'une part, $(V,+,.)$ est un K-espace vectoriel ; d'autre part, $(V,+,\star)$ est un anneau (l'addition $+$ est la même dans les deux cas). On dit que l'on a ainsi muni V d'une structure d'*algèbre sur* K, ou encore de K*-algèbre*, si les deux lois de multiplication sont compatibles :

$$\forall \lambda \in K\ ,\ \forall x,y \in V\ ,\ \begin{cases} (\lambda.x)\star y = \lambda.(x\star y) \\ x\star(\lambda.y) = \lambda.(x\star y). \end{cases}$$

Une *sous-algèbre* de V est une partie W qui est à la fois un sous-espace vectoriel et un sous-anneau. Un *morphisme* de l'algèbre V dans l'algèbre V' est une application $f : V \to V'$ qui est à la fois une application linéaire et un morphisme d'anneaux.

On ne suppose pas l'anneau commutatif. Cependant, on le suppose unitaire, et sa multiplication interne est, bien sûr, associative. Comme de coutume, les symboles de multiplication (interne et externe) seront omis dans l'écriture courante. Pour démontrer que $W \subset V$ est une sous-algèbre, on doit vérifier qu'elle est stable par combinaisons linéaires de deux éléments, qu'elle est stable par multiplication et qu'elle contient l'élément neutre de la multiplication de l'anneau V. On a alors automatiquement une structure d'algèbre sur W.

Exemple.
Soient $\lambda, \mu \in K$ et $x, y \in V$. Calculons dans V le carré de $\lambda x + \mu y$. Le calcul dans l'anneau $(V,+,\star)$ donne d'abord : $(\lambda.x)^2+(\mu.y)^2+(\lambda.x)\star(\mu.y)+(\mu.y)\star(\lambda.x)$. Puis l'on applique les règles de compatibilité, et l'on trouve : $\lambda^2.x^2+\mu^2.y^2+\lambda\mu.(x\star y)+\mu\lambda.(y\star x)$. Noter dans cette dernière formule la présence de carrés et de produits dans K d'une part, dans V d'autre part. Finalement : $(\lambda.x+\mu.y)^2 = \lambda^2.x^2+\mu^2.y^2+\lambda\mu.(x\star y+y\star x)$.

Exercice 14.
Notons e l'unité de l'algèbre non triviale V (donc $e \neq 0$). Montrer que l'application $\lambda \mapsto \lambda.e$ est un morphisme injectif d'algèbres. Que peut-on dire de son image ?
Solution. On a déjà vu que c'est une application linéaire. L'image de l'unité 1 de K est $1.e = e$ (axiomes des espaces vectoriels), donc l'unité de V. La compatibilité des lois interne et externe entraîne : $(\lambda.e) \star (\mu.e) = (\lambda\mu).e$. On a donc un morphisme d'anneaux, donc d'algèbres. Comme tout morphisme d'anneaux dont la source est un corps, il est injectif. Enfin, son image est un sous-anneau de V, et aussi un sous-espace vectoriel (c'est le sous-espace vectoriel Ke), donc une sous-algèbre.

On s'autorisera parfois de ce résultat pour identifier K à la sous-algèbre Ke de V et le scalaire λ à l'élément λe de V. C'est par exemple ce qui se passe lorsque l'on identifie le nombre réel $C \in \mathbb{R}$ à la fonction constante C ou le scalaire $a \in K$ au polynôme constant a (voir les exemples ci-dessous).

Exemples.

1. Les endomorphismes du K-espace vectoriel E forment la K-algèbre $\mathcal{L}_K(E)$. L'image de K dans cette algèbre par l'application de l'exercice est la sous-algèbre des homothéties. De même, les matrices carrées de taille n sur le corps K forment la K-algèbre $M_n(K)$. L'image de K dans cette algèbre par l'application de l'exercice est la sous-algèbre des matrices scalaires.
2. Les polynômes sur le corps K forment la K-algèbre $K[X]$. L'image de K dans cette algèbre par l'application de l'exercice est la sous-algèbre des polynômes constants, que l'on identifiera à K.
3. En définissant le produit fg de deux fonctions, on fait du $\mathbb{R}$-espace vectoriel $\mathcal{F}(I, \mathbb{R})$ une $\mathbb{R}$-algèbre. L'image de $\mathbb{R}$ dans cette algèbre par l'application de l'exercice est la sous-algèbre des fonctions constantes, que l'on identifiera à $\mathbb{R}$. Les fonctions continues forment la sous-algèbre $\mathcal{C}(I, \mathbb{R})$; et l'on a de même la sous-algèbre des fonctions dérivables, des fonctions de classe $\mathcal{C}^k$, etc.
4. Soient I un ensemble d'indices et $(V_i)_{i\in I}$ une famille de K-algèbres indexée par I. On fait du produit $\prod_{i\in I} V_i$ un anneau en posant (avec les notations de la définition) :

$$\forall (x_i)_{i\in I}, (y_i)_{i\in I} \in \prod_{i\in I} V_i \ , \ (x_i)_{i\in I} \star (y_i)_{i\in I} = (x_i \star y_i)_{i\in I}.$$

   La structure de K-espace vectoriel et la structure d'anneau de $\prod_{i\in I} V_i$ sont compatibles et en font une K-algèbre. Si, pour chaque indice i, W_i est une sous-algèbre de V_i, alors $\prod_{i\in I} W_i$ est une sous-algèbre de $\prod_{i\in I} V_i$. Par exemple, K^I est une K-algèbre. Le sous-espace $K^{(I)}$ est stable par multiplication mais ne contient l'unité que si I est fini : sinon, ce n'est donc pas une sous-algèbre.

Exercice 15.
Soit L un corps (que l'on ne suppose pas commutatif) dont K est un sous-corps (commutatif). On a fait de L un K-espace vectoriel. D'autre part, c'est évidemment un anneau. Est-ce une K-algèbre ? Que dire des cas de $\mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{H}$ (où $\mathbb{H}$ désigne le corps des quaternions) ?
Solution. Puisque la loi interne et la loi externe sont toutes deux la multiplication de L, il s'agit de vérifier l'égalité $\lambda xy = x\lambda y$ lorsque $\lambda \in K$ et $x, y \in L$. Cela revient à exiger que tout élément de K commute avec tout élément de L : on dit que K est *central* dans L. Les inclusions $\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$ fournissent autant d'exemples d'algèbres. Le corps $\mathbb{H}$ est une algèbre sur $\mathbb{Q}$ et sur $\mathbb{R}$, mais pas sur $\mathbb{C}$ (car $\mathbb{C}$ n'est pas central dans $\mathbb{H}$).

De l'égalité $(\lambda.e)\star x = \lambda.x$, on déduit que tout idéal I d'une K-algèbre commutative A en est également un sous–espace vectoriel, et l'on obtient donc naturellement une K-algèbre quotient A/I. Par exemple, si P est un polynôme, l'anneau $K[X]/<P>$ est une K-algèbre. Le théorème chinois se transpose tel quel au cas des algèbres : si I et J sont des idéaux étrangers de la K-algèbre A, le morphisme $x \mapsto (x \mod I, x \mod J)$ de A dans $(A/I) \times (A/J)$ est un isomorphisme de K-algèbres.

Proposition et définition 45. Soit x un élément de l'algèbre V. Le sous-espace vectoriel de V engendré par la famille $(x^n)_{n\in\mathbb{N}}$ est une sous-algèbre commutative de V. Cette sous-algèbre est appelée *sous-algèbre engendrée par* x et notée $K[x]$.

On prendra garde à ne pas confondre cette notation avec celle qui désigne l'algèbre $K[X]$ des polynômes. Pour rappeler que, dans cette dernière, la lettre X dénote une « indéterminée », nous essaierons d'utiliser systématiquement une lettre majuscule dans le cas des polynômes.

Démonstration. Par convention, x^0 est l'élément neutre de la multiplication. Il reste à vérifier que $K[x]$ est stable par multiplication. Les éléments de $K[x]$ sont les combinaisons linéaires des x^n, donc de la forme $\sum_{i\in I} a_i x^i$, où I est une partie finie de $\mathbb{N}$ et les $a_i \in K$. Tout élément de $K[x]$ peut donc se mettre sous la forme $a_0x^0 + \cdots + a_mx^m = \sum_{i=0}^{m} a_i x^i$, avec $m \in \mathbb{N}$ et $a_0, \ldots, a_m \in K$. Le calcul de $\left(\sum_{i=0}^{m} a_i x^i\right) \star \left(\sum_{j=0}^{n} b_j x^j\right)$ donne alors :

$$\sum_{i=0,j=0}^{i=m,j=n} (a_i x^i) \star (b_j x^j) = \sum_{i=0,j=0}^{i=m,j=n} (a_i b_j) x^i \star x^j = \sum_{i=0,j=0}^{i=m,j=n} (a_i b_j) x^{i+j} = \sum_{k=0}^{m+n} c_k x^k,$$

où $c_k := \sum_{\substack{i+j=k \\ 0\leqslant i\leqslant m \\ 0\leqslant j\leqslant n}} a_i b_j$. Cette formule entraîne la stabilité pour la multiplication. On en déduit également la commutativité, car, de l'égalité évidente $\sum a_i b_j = \sum b_j a_i$, découle la relation :

$$\left(\sum_{i=0}^{m} a_i x^i\right) \star \left(\sum_{j=0}^{n} b_j x^j\right) = \left(\sum_{j=0}^{n} b_j x^j\right) \star \left(\sum_{i=0}^{m} a_i x^i\right).$$

∎

Théorème 46 (Propriété universelle de l'algèbre des polynômes).
Soit A une K-algèbre. Alors, pour tout $x \in A$, il existe un unique morphisme de K-algèbres $K[X] \to A$ tel que $X \mapsto x$.
L'image de ce morphisme est la sous-algèbre $K[x]$ de A.

Démonstration. Soit φ un tel morphisme ; l'image du polynôme $\sum a_i X^i$ est :

$$\varphi\big(\sum a_i X^i\big) = \sum a_i \varphi(X^i) = \sum a_i \varphi(X)^i = \sum a_i x^i,$$

d'où l'unicité. Réciproquement, en posant $\varphi\big(\sum a_i X^i\big) := \sum a_i X^i$, on prouve facilement que l'on a un morphisme d'algèbres : on étend ainsi les calculs et la définition de « l'évaluation de P en x » ([L1], module sur les polynômes). ∎

Proposition et définition 47. Avec les notations ci-dessus, on a l'alternative :
(i) Soit le morphisme $P \mapsto P(x)$ est injectif et la K-algèbre $K[x]$ est de dimension infinie. Dans ce cas, on dit que x est *transcendant* sur K.
(ii) Soit la K-algèbre $K[x]$ est de dimension finie. Le noyau du morphisme $P \mapsto P(x)$ est un idéal principal non trivial dont les éléments sont appelés *polynômes annulateurs de* x et dont l'unique générateur unitaire est appelé *polynôme minimal de* x ; on le note μ_x. Dans ce cas, on dit que x est *algébrique* sur K. Le degré de μ_x, qui est la dimension de $K[x]$, est appelé *degré de* x (sur K). On a un isomorphisme : $K[x] \simeq \dfrac{K[X]}{<\mu_x>}$.

Démonstration. Si le morphisme est injectif, $K[x]$ est isomorphe à $K[X]$, qui est de dimension infinie. Sinon, son noyau est un idéal $<P>$, où P est unitaire et non constant (il faut en effet avoir $P(x) = 0$). C'est ce polynôme P qui est le polynôme (annulateur) minimal μ_x. Les polynômes annulateurs, *i.e.* les polynômes P tels que $P(x) = 0$, sont les éléments du noyau, donc les multiples de μ_x. D'après le premier théorème d'isomorphisme, on a un isomorphisme $K[x] \simeq \dfrac{K[X]}{<\mu_x>}$, et la dimension de $K[x]$ est donc $\deg \mu_x$. ∎

Corollaire 48. Si x est algébrique et si $\mu_x = P_1^{r_1} \cdots P_k^{r_k}$, (décomposition en facteurs irréductibles dans $K[X]$), on a un isomorphisme : $K[x] \simeq \dfrac{K[X]}{<P_1^{r_1}>} \times \cdots \times \dfrac{K[X]}{<P_k^{r_k}>}$.

Démonstration. Cela découle du théorème chinois. ∎

Exemples.

1. En prenant $K := \mathbb{Q}$ et $A := \mathbb{C}$, on retrouve la notion de nombres algébriques ou transcendants de [L1].
2. Dans une K-algèbre de dimension finie n, tout élément est algébrique et de degré inférieur ou égal à n. En particulier, dans les $\mathbb{R}$-algèbres $\mathbb{C}$ et $\mathbb{H}$, dans les K-algèbres $M_n(K)$, $\mathcal{L}_K(E)$, tout élément est algébrique.

4 Séries formelles

Les séries de fonctions de la forme $\sum_{n\geqslant 0} a_n z^n$ sont étudiées en analyse sous le nom de *séries entières* (modules II.2 et II.6). Les coefficients a_n sont alors, en général, des nombres réels ou complexes. Il y a cependant de bonnes raisons d'étudier de telles séries d'un point purement algébrique (indépendamment des questions de convergence) et pour des coefficients arbitraires. On parle alors de *séries formelles*. Comme pour les polynômes, cette étude est possible sur un *anneau* de coefficients et avec *plusieurs* indéterminées, mais nous n'approfondirons que le cas d'un *corps* de coefficients et d'*une* indéterminée.
Notre but est de donner une signification et des règles de calcul pour des expressions de la forme $\sum_{n\geqslant 0} a_n X^n$, où, pour commencer, les coefficients a_n appartiennent à un anneau commutatif A (nécessairement unitaire par définition). De telles expressions s'apparentent à des « polynômes de degré infini » et c'est cette idée vague qui guide leur construction.

Construction des séries formelles sur un anneau A

Rappelons que l'on a construit $A[X]$ en dotant d'une structure d'anneau le groupe $A^{(\mathbb{N})}$ des suites d'éléments de A *à support fini*. Nous considérons donc ici le groupe $A^{\mathbb{N}}$ de *toutes* les suites $(a_n)_{n\in\mathbb{N}}$ d'éléments de A. Nous définissons un produit sur $A^{\mathbb{N}}$ par la formule (*produit de Cauchy*) :

$$(a_n)_{n\in\mathbb{N}}\,(b_n)_{n\in\mathbb{N}} = (c_n)_{n\in\mathbb{N}},\ \text{où, } \forall n\in\mathbb{N}\ ,\ c_n := \sum_{i+j=n} a_i b_j = \sum_{i=0}^{n} a_i b_{n-i}.$$

Bien que les suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ ne soient pas supposées à support fini, la formule qui définit c_n ne fait intervenir qu'un nombre fini de termes. Comme pour les polynômes, on vérifie que l'on a fait de $A^{\mathbb{N}}$ un anneau commutatif unitaire dont l'élément neutre est la suite $(1,0,\ldots)$. Il est immédiat que $A^{(\mathbb{N})}$ en est un sous-anneau. On note encore X l'élément défini par la suite $(0,1,0,\ldots)$ et $A[[X]]$ l'anneau $A^{\mathbb{N}}$. Ses éléments sont appelés *séries formelles en X à coefficients dans A*. La suite $(a_n)_{n\in\mathbb{N}}$ est alors notée $\sum_{n\geqslant 0} a_n X^n$.
Cette notation ne doit pas être considérée comme une somme ou une combinaison linéaire infinie, ce qui n'aurait aucun sens. Sa signification sera éclaircie à la section 4.2. L'anneau $A[X]$ des polynômes en X à coefficients dans A est donc un sous-anneau de l'anneau $A[[X]]$ des séries formelles en X à coefficients dans A.

Ainsi, l'expression $f := \sum_{n\geqslant 0} X^n Y^n$ désigne évidemment un élément de $\mathbb{Z}[[X]][[Y]]$, et même de $\mathbb{Z}[X][[Y]]$, puisque l'on peut écrire $f = \sum_{n\geqslant 0} a_n Y^n$, où chaque $a_n := X^n$ est élément de $\mathbb{Z}[X]$. Mais elle ne désigne pas un élément de $\mathbb{Z}[[X]][Y]$, car on ne peut mettre f sous la forme d'une somme *finie* $\sum_{n=0}^{d} a_n Y^n$ où chaque a_n est élément de $\mathbb{Z}[[X]]$.

4.1 La K-algèbre des séries formelles

Dorénavant, la lettre K désigne un corps commutatif et $K[[X]] := \Big\{ \sum_{n\geqslant 0} a_n X^n \;\Big|\; (a_n)_{n\in\mathbb{N}} \in K^{\mathbb{N}} \Big\}$ l'anneau des séries formelles en X à coefficients dans K. C'est, de manière évidente, un K-espace vectoriel et même une K-algèbre. Les opérations y sont définies par les formules :

$$\lambda\Big(\sum_{n\geqslant 0} a_n X^n\Big) + \mu\Big(\sum_{n\geqslant 0} b_n X^n\Big) := \sum_{n\geqslant 0} c_n X^n \quad \text{où} \quad \forall n \in \mathbb{N}\ ,\ c_n := \lambda a_n + \mu b_n,$$

$$\Big(\sum_{n\geqslant 0} a_n X^n\Big)\Big(\sum_{n\geqslant 0} b_n X^n\Big) := \sum_{n\geqslant 0} c_n X^n \quad \text{où} \quad \forall n \in \mathbb{N}\ ,\ c_n := \sum_{i+j=n} a_i b_j = \sum_{i=0}^{n} a_i b_{n-i}.$$

La K-algèbre $K[X]$ des polynômes en X à coefficients dans K est donc une sous-algèbre de la K-algèbre $K[[X]]$ des séries formelles en X à coefficients dans K. Nous ferons appel à la commode notation suivante : si $n \in \mathbb{N}$ et si $f := \sum_{n\geqslant 0} a_n X^n$, alors $[X^n]f$ désigne le coefficient a_n (coefficient de X^n dans f).
Ainsi, $[X^n](\lambda f + \mu g) = \lambda[X^n]f + \mu[X^n]g$, etc.

Exercice 16.
La famille des X^n est-elle une base de $K[[X]]$?
Solution. Non : c'est une famille libre, mais le sous–espace vectoriel qu'elle engendre est $K[X]$. On ne connaît aucune base de $K[[X]]$. On peut démontrer que toute base du K-espace vectoriel $K[X]$ est dénombrable ; mais, par exemple, que toute base du $\mathbb{Q}$-espace vectoriel $\mathbb{Q}[[X]]$ a la puissance du continu (pour la signification de ces termes, voir la section sur les cardinaux du module « Fondements » de [L1]).

On appelle *terme constant* de la série formelle $f := \sum_{n\geqslant 0} a_n X^n$ le scalaire $\mathrm{tc}(f) := a_0 = [X^0]f$. On le note également $f(0)$. *Cette notation ne doit pas être considérée comme une évaluation de f en $X := 0$, ce qui n'aurait aucun sens*. Sa signification sera éclaircie à la section 4.2.3. Les règles de calcul (combinaisons linéaires, produits) peuvent se résumer ainsi : $f \mapsto \mathrm{tc}(f)$ est un morphisme de K-algèbres de $K[[X]]$ sur K. Ce morphisme s'identifie d'ailleurs au passage au quotient $K[[X]] \to K[[X]]/{<X>} \simeq K$ par l'idéal principal $<X>$.

Proposition et définition 49.
(i) Soit $f := \sum_{n\geqslant 0} a_n X^n \in K[[X]]$ non nulle. Le plus grand entier i tel que X^i divise f est également le plus petit entier n tel que $a_n \neq 0$. On le note $\omega(f)$ et on l'appelle *ordre* ou *valuation* de f. Par convention, l'ordre de 0 est $+\infty$.
(ii) On a, pour $\lambda \in K^*$ et $f, g \in K[[X]]$, les formules :

$$\omega(\lambda f) = \omega(f),\ \omega(f+g) \geqslant \min\big(\omega(f), \omega(g)\big) \quad \text{et} \quad \omega(fg) = \omega(f) + \omega(g).$$

De plus, si $\omega(f) \neq \omega(g)$, alors $\omega(f+g) = \min\big(\omega(f), \omega(g)\big)$.
Les règles de calcul concernant $+\infty$ sont indiquées dans le module sur les polynômes de [L1].

La démonstration est identique à celle donnée pour les polynômes. En utilisant l'ordre au lieu du degré (qui n'est pas défini ici), on en déduit de même :

Corollaire 50. L'anneau $K[[X]]$ est intègre.

Proposition 51.
Pour que $f \in K[[X]]$ soit inversible, il faut, et il suffit, que $\omega(f) = 0$.

Démonstration. Puisque $fg = 1 \Rightarrow \omega(f) + \omega(g) = 0$, la condition est évidemment nécessaire. On peut aussi bien invoquer l'égalité $\mathrm{tc}(f)\,\mathrm{tc}(g) = 1$.
Supposons réciproquement que $\omega(f) = 0$ et écrivons $f := \sum_{n\geqslant 0} a_n X^n$, avec $a_0 \neq 0$. La série formelle $g := \sum_{n\geqslant 0} b_n X^n$ est inverse de f si, et seulement si, $a_0 b_0 = 1$ et, pour tout $n \geqslant 1$, $\sum_{i+j=n} a_i b_j = 0$. On pose donc $b_0 := a_0^{-1}$ et, pour tout $n \geqslant 1$, la relation de récurrence : $b_n := -a_0^{-1} \sum_{i=0}^{n-1} a_{n-i} b_i$. ∎

Exemple. La série $1 + X + X^2 \in \mathbb{C}[[X]]$ est inversible. Son inverse est $\sum_{n\geqslant 0} b_n X^n$, où $b_0 := 1$, $b_0 + b_1 = 0$, d'où $b_1 := -1$; puis, pour $n \geqslant 2$, la relation de récurrence linéaire à deux pas : $b_n + b_{n-1} + b_{n-2} = 0$. On trouve que la suite des b_n est périodique : $b_{3p} = 1$, $b_{3p+1} = -1$, $b_{3p+2} = 0$.

Exercice 17.
Calculer l'inverse de $1 - X^a$ ($a \in \mathbb{N}^*$).
En déduire un nouveau calcul de l'inverse de $1 + X + X^2 \in \mathbb{C}[[X]]$.
Solution. On trouve que $(1 - X^a)^{-1} = \sum_{n\geqslant 0} b_n X^n$, où $b_n = 1$ si n est multiple de a, et 0 sinon ; autrement dit, $(1 - X^a)^{-1} = \sum_{p\geqslant 0} X^{ap}$. Comme $(1 - X)(1 + X + X^2) = 1 - X^3$, on en déduit que l'inverse de $1 + X + X^2$ est $(1 - X) \sum_{p\geqslant 0} X^{3p} = \sum_{p\geqslant 0} X^{3p} - \sum_{p\geqslant 0} X^{3p+1}$, comme précédemment (voir également l'exercice I.1.48 de la page 55).

Théorème 52. (i) Tout élément non nul de $K[[X]]$ admet une unique écriture uX^k, où u est inversible. L'élément f divise l'élément g si, et seulement si, $\omega(f) \leqslant \omega(g)$.
(ii) L'anneau $K[[X]]$ est principal. Tous ses idéaux non nuls sont de la forme $<X^k>$.

Pour la notion d'anneau princippal, voir le module d'arithmétique de [L1].

Démonstration. (i) Pour la première assertion, il suffit de prendre $k := \omega(f)$. La deuxième assertion en découle.

(ii) Si I est un idéal non nul, soit $f \in I$ d'ordre minimum.
Alors, d'après (i), $I = <f> = <X^k>$, avec $k := \omega(f)$. ∎

Ainsi, étant donnés f et g dans $K[[X]]$, l'un divise l'autre : on dit que $K[[X]]$ est un *anneau de valuation*. L'assertion (ii) se résume en disant que $K[[X]]$ est un *anneau de valuation discrète* (c'est une propriété plus forte). On en déduit qu'il admet un seul idéal maximal, l'idéal $<X>$: on dit que $K[[X]]$ est un *anneau local*. Enfin, ses seuls idéaux premiers sont $\{0\}$ et $<X>$.

4.1.1 Développements d'une fraction rationnelle en série de Laurent

L'anneau $K[[X]]$ étant intègre, on peut construire son corps des fractions, que l'on notera $K((X))$: c'est le corps des *séries de Laurent formelles en X à coefficients dans K*. Une série de Laurent formelle est dite *entière* si elle est élément de $K[[X]]$. De l'assertion (i) du théorème 52 de la page ci-contre, on déduit que toute série de Laurent g qui n'est pas entière est de la forme $\dfrac{1}{f} = \dfrac{1}{uX^k} = \dfrac{v}{X^k}$, où $v \in K[[X]]$ est inversible. On peut donc écrire :

$$g = \sum_{n \geqslant -k} b_n X^n = b_{-k}X^{-k} + \cdots + b_0 + \cdots ,$$

avec $b_{-k} \neq 0$. L'entier $-k$ est *l'ordre $\omega(g)$ de la série de Laurent g*. L'extension de ω à $K((X))$ vérifie les mêmes propriétés qu'à la proposition 49 de la page 31. On a donc les égalités :

$$K((X)) = K[[X]][X^{-1}] = \Big\{ \sum_{n \geqslant n_0} a_n X^n \mid n_0 \in \mathbb{Z} \quad \text{et} \quad \forall n \geqslant n_0 \ , \ a_n \in K \Big\}.$$

Chaque série de Laurent $f := \sum\limits_{n \geqslant n_0} a_n X^n$ admet une unique décomposition en une *partie entière* $\lfloor f \rfloor := \sum\limits_{n \geqslant 0} a_n X^n \in K[[X]]$ et une *partie polaire* :

$$f - \lfloor f \rfloor := \sum_{n_0 \leqslant n < 0} a_n X^n \in X^{-1}K[X^{-1}].$$

Il y correspond une décomposition en somme directe :

$$K((X)) = K[[X]] \oplus X^{-1}K[X^{-1}].$$

Le morphisme d'anneaux intègres : $K[X] \to K[[X]]$ s'étend en un morphisme entre leurs corps de fractions : $K(X) \to K((X))$. On dit que l'on a associé à chaque fraction rationnelle son *développement en série de Laurent*. Par exemple, la fraction rationnelle $\dfrac{X+1}{X^2 - X}$ admet pour développement en série de Laurent le produit de $(X+1)$ par la série de Laurent inverse de la série formelle $X^2 - X$:

$$\frac{X+1}{X^2-X} = -\frac{1}{X}\frac{1+X}{1-X} = -\frac{1}{X}(1+X)(1+X+X^2+\cdots) = -X^{-1}-2-2X-2X^2-\cdots$$

Il y a un rapport entre le développement en série de Laurent et les développements limités du module sur les polynômes de [L1]. Soit $R \in K(X)_0$, autrement dit, R est définie en 0. Son

dénominateur est donc inversible dans $K[[X]]$, et son développement en série de Laurent est une série formelle $\sum_{n\geqslant 0} a_n X^n$ (on parle plutôt dans ce cas de *développement en série formelle*). Alors chaque $DL_m(R)$ (développement limité à l'ordre m) est la troncature à l'ordre m de cette série formelle, c'est-à-dire le polynôme $S := \sum_{n=0}^{m} a_n X^n$. Rappelons que, si $R = \dfrac{P}{Q}$, S s'obtient par *division selon les puissances croissantes de P par Q* : c'est l'unique polynôme de $K_m[X]$ tel que $\omega(P - SQ) \geqslant m+1$.

Exercice 18.

Quel est le développement en série formelle de $\dfrac{1}{(1-X^{a_1})\cdots(1-X^{a_k})}$, où les $a_i \in \mathbb{N}^*$ sont distincts ?

Solution. C'est $\sum_{n\geqslant 0} b_n X^n$, où b_n est le nombre de solutions en entiers naturels de l'équation $a_1x_1 + \cdots + a_kx_k = n$. Le raisonnement est le même que dans le cas de $\dfrac{1}{(1-X^2)(1-X^3)}$ (exercices du module sur les polynômes de [L1]).

Proposition 53. Pour que la série formelle $f := \sum_{n\geqslant 0} a_n X^n$ soit le développement en série formelle d'une fraction rationnelle, il faut, et il suffit, qu'il existe $d \in \mathbb{N}$ et $\lambda_0, \ldots, \lambda_q \in K$ non tous nuls tels que :

$$\forall n \geqslant d\ ,\ a_n\lambda_0 + \cdots + a_{n+q}\lambda_q = 0.$$

Démonstration. Notons $P := \lambda_0 X^q + \cdots + \lambda_q$. Alors $a_n\lambda_0 + \cdots + a_{n+q}\lambda_q$ est le coefficient $[X^{n+q}](Pf)$ de X^{n+q} dans la série formelle Pf. La condition de l'énoncé équivaut donc à dire que Pf est un polynôme de degré strictement inférieur à $d+q$, c'est-à-dire que f est le développement en série formelle d'une fraction rationnelle (de dénominateur P). ∎

Comme le montre l'exercice suivant, d'autres critères de rationalité sont basés sur les *déterminants de Hankel* :

$$H_n^{(k)} := \det A_n^{(k)},\ \text{où } A_n^{(k)} := \begin{pmatrix} a_n & a_{n+1} & \cdots & a_{n+k-1} \\ a_{n+1} & a_{n+2} & \cdots & a_{n+k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n+k-1} & a_{n+k} & \cdots & a_{n+2k-2} \end{pmatrix}.$$

Exercice 19.

Soit $f := \sum_{n\geqslant 0} a_n X^n \in K[[X]]$. Montrer que si $H_{d+j}^{(q+1)} = 0$ et $H_{d+j}^{(q)} \neq 0$ pour tout entier naturel j, alors f est rationnelle.

Solution. Puisque $H_d^{(q)} \neq 0$, le système $A_d^{(q)} \begin{pmatrix} \lambda_0 \\ \vdots \\ \lambda_{q-1} \end{pmatrix} = - \begin{pmatrix} a_{d+q} \\ \vdots \\ a_{d+2q-1} \end{pmatrix}$ admet une unique solution non triviale. Puisque $H_d^{(q+1)} = 0$, la dernière ligne de $A_d^{(q)}$ est combinaison linéaire des précédentes (qui ne sont pas liées parce que $H_d^{(q)} \neq 0$) et l'on tire de ce système l'égalité $a_{d+q}\lambda_0 + \cdots + a_{d+2q-1}\lambda_{q-1} = -a_{d+2q}$, d'où $A_{d+1}^{(q)} \begin{pmatrix} \lambda_0 \\ \vdots \\ \lambda_{q-1} \end{pmatrix} = - \begin{pmatrix} a_{d+1+q} \\ \vdots \\ a_{d+2q} \end{pmatrix}$. En itérant, on trouve une relation analogue pour tout $n \geqslant d$, et l'on peut appliquer la proposition. (Ce critère est complété dans l'exercice I.1.51 de la page 55.)

Le développement en série de Laurent décrit ci-dessus s'apparente aux développements limités *à l'origine* d'une fonction numérique. L'analogue des développements limités *en d'autres points* s'obtient comme suit. Soit $a \in K$ et soit $R \in K(X)$. En posant $X = a+Y$ et $R(X) = R(a+Y) = S(Y)$, on définit une fraction rationnelle $S \in K(Y)$. Celle-ci admet un développement en série de Laurent $S = \sum_{n \geqslant n_0} a_n Y^n \in K[[Y]]$. On remplace alors Y par $X - a$, et l'on écrit :

$$R(X) = \sum_{n \geqslant n_0} a_n (X-a)^n.$$

Ainsi, le développement en série de Laurent de $\dfrac{1}{X}$ au voisinage de 1 s'écrit-il :

$$\frac{1}{X} = \frac{1}{1+(X-1)} = \sum_{n \geqslant 0} (-1)^n (X-1)^n.$$

Il ne faut toutefois pas attacher un sens littéral à ces écritures. En particulier, $X - a$ doit être traité comme une indéterminée et $(X-a)^n$ ne doit pas être développé.

Il y a enfin un analogue pour les développements limités à l'infini, le *développement en série de Laurent à l'infini* d'une fraction rationnelle $R \in K(X)$. De tels développements apparaîtront lors de l'étude des fractions continues dans les modules « Compléments d'arithmétique » et « Polynômes orthogonaux » de [L2-L3]. On applique à R l'isomorphisme de $K(X)$ sur $K(\xi)$ défini par $X \mapsto \dfrac{1}{\xi}$, que l'on compose avec le plongement de $K(\xi)$ dans $K((\xi))$ (c'est-à-dire le morphisme de corps qui a été défini au début de 4.1.1, page 33). Autrement dit, on pose $S(\xi) := R(1/\xi)$ et l'on développe S en série de Laurent.

Exercice 20.
Développer en série de Laurent à l'infini la fraction rationnelle $R := \dfrac{X^3}{X^2+1}$.

Solution. On note $\xi := \dfrac{1}{X}$.

$$R = \frac{\xi^{-3}}{\xi^{-2}+1} = \frac{1}{\xi+\xi^3} = \frac{1}{\xi}\frac{1}{1+\xi^2} = \frac{1}{\xi}(1-\xi^2+\xi^4-\cdots) = \xi^{-1} - \xi + \xi^3 - \cdots$$

Proposition 54. Notons $\varphi : K(X) \to K((\xi))$ le morphisme qui associe à R son développement en série de Laurent à l'infini. On a alors :

$$\omega\big(\varphi(R)\big) = -\deg R.$$

De plus, la partie polaire de $\varphi(R)$ est $\varphi(P) - \mathrm{tc}(P)$, où P est la partie entière de R.

Démonstration. Si $R = a_0 + \cdots + a_n X^n$ est un polynôme de degré n, alors $P = R$, $\mathrm{tc}(P) = a_0$ et $\varphi(R) = a_n\xi^{-n} + \cdots + a_1\xi^{-1} + a_0$ a pour ordre $-n = -\deg P$ et pour partie polaire $a_n\xi^{-n} + \cdots + a_1\xi^{-1} = \varphi(P) - \mathrm{tc}(P)$.

Dans le cas général, écrivons $R = P + \dfrac{b_0 + \cdots + b_p X^p}{c_0 + \cdots + c_q X^q}$, où $P \in K[X]$, $b_p c_q \neq 0$ et $p < q$ et conservons les mêmes notations pour P. Alors :

$$\varphi(R - P) = \xi^{q-p}\frac{b_0\xi^p + \cdots + b_p}{c_0\xi^q + \cdots + c_q}$$

est entière et d'ordre $q - p > 0$ puisque $b_p c_q \neq 0$. La conclusion découle de même. ∎

4.1.2 Dérivation dans $K[[X]]$

Appelons *dérivation* de la K-algèbre A un endomorphisme D du K-espace vectoriel A qui vérifie la *règle de Leibniz* : $\forall f, g \in A\ ,\ D(fg) = fD(g) + gD(f)$. La démonstration de la proposition suivante est similaire à celle donnée pour les polynômes et les fractions rationnelles (module sur les polynômes de [L1]). Cette définition étend d'ailleurs celle donnée pour $K(X)$.

Proposition et définition 55.
Pour toute série de Laurent $f := \sum_{n \geqslant n_0} a_n X^n \in K((X))$, on appelle *dérivée de* f la série de Laurent : $f' := \sum_{n \geqslant n_0} n a_n X^{n-1}$. L'application $f \mapsto f'$ est une dérivation de la K-algèbre $K((X))$. La dérivée d'une série formelle est une série formelle.

On en déduit les règles usuelles de calcul. Par exemple, introduisant la *dérivée logarithmique* $\dfrac{f'}{f}$, on constate que la dérivée logarithmique de fg (resp. de f/g) est égale à la somme (resp. à la différence) des dérivées logarithmiques de f et de g, etc.

La preuve de la proposition suivante est facile et laissée au lecteur.

Proposition 56. On suppose le corps K de caractéristique nulle. Alors, si $\omega(f) \neq 0$, on a : $\omega(f') = \omega(f) - 1$. De plus, $f' = 0$ si, et seulement si, f est constante (*i.e.* $f \in K$). Les séries de Laurent admettant une primitive (*i.e.* une série dont elles sont la dérivée) sont celles dont le coefficient de X^{-1} est nul. Les primitives d'une série donnée diffèrent entre elles d'une constante.

Ces hypothèses ne sont pas gratuites. Par exemple $1+X$ et $1+X^2$ ont même ordre, mais pas leurs dérivées ; et, sur un corps de caractéristique 2, la dérivée de X^{-2} est 0 (contredisant les deux premières règles données), et X^{-3} n'a pas de primitive.

Exemple. On suppose K de caractéristique nulle.

L'unique primitive de la série $\dfrac{1}{1+X} = \sum_{n\geqslant 0}(-1)^n X^n$ dont le terme constant est 0 est la série formelle : $L(X) := \sum_{n\geqslant 1} \dfrac{(-1)^{n-1}}{n} X^n$. Nous verrons (section 4.2.4) que c'est la série logarithme $\ln(1+X)$.

Exercice 21.

On suppose K de caractéristique nulle. Calculer la dérivée $k^{ème}$ de $\dfrac{1}{1-X}$ et en déduire le développement en série formelle de $\dfrac{1}{(1-X)^{k+1}}$.

Solution. La dérivée $k^{ème}$ de $\dfrac{1}{1-X} = \sum_{n\geqslant 0} X^n$ est $\dfrac{k!}{(1-X)^{k+1}} = \sum_{n\geqslant 0} \dfrac{(n+k)!}{n!} X^n$ (dériver k fois les deux membres de l'égalité), d'où la formule :

$$\frac{1}{(1-X)^{k+1}} = \sum_{n\geqslant 0} \binom{n+k}{k} X^n.$$

Équations différentielles

Nous examinerons deux cas : équation différentielle linéaire d'ordre d, équation différentielle non linéaire d'ordre 1. *Nous supposons le corps K de caractéristique nulle.*

Proposition 57. Soient $A_0, \dots, A_d \in K[[X]]$. On suppose $\omega(A_0) = 0$. Soient de plus $c_0, \dots, c_{d-1} \in K$. Alors l'équation différentielle $A_0 f^{(d)} + \cdots + A_d f = 0$ admet une unique solution $f \in K[[X]]$ telle que $f^{(i)}(0) = c_i$ pour $i = 0, \dots, d-1$.

Démonstration.

Puisque A_0 est inversible, on peut écrire l'équation sous la forme $f^{(d)} = \sum_{i=0}^{d-1} B_{d-i} f^{(i)}$, où $B_i := -\dfrac{A_i}{A_0} = \sum_{n\geqslant 0} b_{i,n} X^n$. Notons $f := \sum_{n\geqslant 0} a_n X^n$. On a donc $[X^j] B_{d-i} = b_{d-i,j}$ et $[X^k] f^{(i)} = \dfrac{(k+i)!}{k!} a_{k+i}$. L'équation différentielle donne, par identification des termes en X^n :

$$\frac{(n+d)!}{n!} a_{n+d} = \sum_{i=0}^{d-1} \sum_{j+k=n} b_{d-i,j} \frac{(k+i)!}{k!} a_{k+i},$$

ce qui est une relation de récurrence, puisque le membre droit ne fait intervenir que des a_m d'indices $m \leqslant n+d-1$ et que l'on peut diviser par $\dfrac{(n+d)!}{n!}$. La récurrence est initialisée par les « conditions initiales » $f^{(i)}(0) = c_i$, c'est-à-dire $a_i := \dfrac{c_i}{i!}$ pour $i = 0, \dots, d-1$. ∎

Exemple. L'équation $f' - f = 0$ admet une unique solution telle que $f(0) = 1$. C'est la série formelle $\sum_{n\geqslant 0} a_n X^n$, avec $a_0 = 1$ et $(n+1)a_{n+1} = a_n$, d'où $a_n = \dfrac{1}{n!}$.

On a donc $E(X) := \sum_{n\geqslant 0} \dfrac{1}{n!} X^n$. Nous verrons (section 4.2.4) que c'est la série exponentielle $\exp(X)$.

Exercice 22.

Résoudre l'équation $(1+X)f' - \alpha f = 0$ (où $\alpha \in K$) avec la condition initiale $f(0) = 1$.

Solution.

L'unique solution est $f := \sum_{n\geqslant 0} a_n X^n$, avec $a_0 = 1$ et $(n+1)a_{n+1} = \alpha \sum_{k=0}^{n} (-1)^{n-k} a_k$. Cette récurrence est malaisée (récurrence forte) et il vaut mieux identifier les coefficients en X^n de $(1+X)f'$ et de αf, ce qui donne : $(n+1)a_{n+1} + na_n = \alpha a_n$, d'où $a_{n+1} = \dfrac{\alpha - n}{n+1} a_n$. En itérant, on trouve : $a_n = \dbinom{\alpha}{n}$. Par exemple, si α est entier, c'est $(1+X)^\alpha$ (formule du binôme, qui sera généralisée à la section 4.2.5).

Exercice 23.

Résoudre l'équation $Xf' - \alpha f = 0$ (où $\alpha \in K$).

Solution. Cette équation ne vérifie pas l'hypothèse $\omega(A_0) = 0$ du théorème. On dit qu'elle n'est pas *régulière*. Si $\alpha \in \mathbb{Z} \subset K$, les solutions dans $K((X))$ sont les cX^α ($c \in K$). Sinon, il n'y a aucune solution non triviale dans $K((X))$.

Proposition 58. Soient $A_i := \sum_{n\geqslant 0} \alpha_{i,n} X^n \in K[[X]]$ ($i \in \mathbb{N}$) des séries formelles et $\varphi(X,Y) := \sum_{i\geqslant 0} A_i(X) Y^i \in K[[X]][[Y]]$. Il existe une unique solution $f \in K[[X]]$ de l'équation différentielle $f' = \varphi(X, f)$ telle que $f(0) = 0$.

Démonstration. Il s'agit de trouver $f := \sum_{n\geqslant 0} a_n X^n \in K[[X]]$ telle que $a_0 = 0$ et $f' = \sum_{i\geqslant 0} A_i(X) f^i$. Il découlera de la section 4.2 que cette somme infinie a bien un sens, parce que $f(0) = 0$, mais nous allons vérifier pratiquement que le calcul des coefficients ne pose pas de problème.

Pour $i, k \in \mathbb{N}$, le coefficient $[X^k] f^i$ ne fait intervenir que les coefficients $a_1, \ldots, a_k$ et il est polynomial. Notons-le $P_{i,k}(a_1, \ldots, a_k)$. Les $P_{i,k}$ sont des polynômes « universels » à coefficients entiers naturels, que l'on pourrait expliciter (exercice I.1.57 de la page 56). Comme $\omega(f^i) \geqslant i$, le polynôme $P_{i,k}$ est nul si $k < i$. L'équation différentielle donne, par identification des termes en X^n, la relation de récurrence : $(n+1)a_{n+1} = \sum_{\substack{i,j,k\geqslant 0 \\ j+k=n}} \alpha_{i,j} P_{i,k}(a_1, \ldots, a_k)$. Les seuls termes non nuls sont ceux tels que $i \leqslant k$, la somme est donc finie. ∎

4.2 Convergence dans $K[[X]]$

Le corps K étant quelconque, il n'est pas question de parler de convergence de suites (ou de séries) d'éléments de K. Cependant, il est possible de donner forme à l'idée que la suite des polynômes $\sum_{n=0}^{N} a_n X^n$ « converge » vers la série formelle $\sum_{n\geqslant 0} a_n X^n$, et que les termes $a_n X^n$ sont « de plus en plus petits ».

Définition 8. On dit qu'une suite $(f_p)_{p\in\mathbb{N}}$ d'éléments de $K((X))$ *converge* vers $f \in K((X))$ si $\lim_{p\to+\infty} \omega(f-f_p) = +\infty$.

On écrit alors $f = \lim_{p\to+\infty} f_p$ et l'on dit que f est la *limite* de la suite $(f_p)_{p\in\mathbb{N}}$.

Du fait que $\omega(f-g) \geqslant \min\big(\omega(f-f_p), \omega(g-f_p)\big)$, si la suite $(f_p)_{p\in\mathbb{N}}$ converge vers f et vers g, on a $\omega(f-g) = +\infty$ et $f = g$. La limite, si elle existe, est donc unique et notre terminologie est cohérente. On parle parfois de *convergence X-adique*.

Point Méthode

Dire que $\omega(f-f_p) \geqslant N$, c'est dire que les coefficients de f et de f_p d'indice $n < N$ sont les mêmes. Lorsque la suite $(f_p)_{p\in\mathbb{N}}$ converge vers f, les coefficients se stabilisent donc : pour tout n, il existe un rang p_0 tel que, pour $p \geqslant p_0$, tous les f_p ont même coefficient d'indice n ; et ce coefficient est également le coefficient d'indice n de f. Autrement dit, $[X^n]f$ est la valeur en laquelle stationne la suite des $[X^n]f_p$. Il en résulte que $\omega(f_p)$ stationne à la valeur $\omega(f)$. Une limite d'éléments de $K[[X]]$ est donc un élément de $K[[X]]$.

Proposition 59. Si les suites $(f_p)_{p\in\mathbb{N}}$ et $(g_p)_{p\in\mathbb{N}}$ convergent, alors, quels que soient $\alpha, \beta \in K$, les suites $(\alpha f_p + \beta g_p)_{p\in\mathbb{N}}$ et $(f_p g_p)_{p\in\mathbb{N}}$ convergent et :

$$\lim_{p\to+\infty} (\alpha f_p + \beta g_p) = \alpha \lim_{p\to+\infty} f_p + \beta \lim_{p\to+\infty} g_p,$$

$$\lim_{p\to+\infty} f_p g_p = \big(\lim_{p\to+\infty} f_p\big)\big(\lim_{p\to+\infty} g_p\big).$$

Si de plus $\lim_{p\to+\infty} f_p$ est non nul, alors il existe un rang p_0 tel que, pour $p \geqslant p_0$, tous les f_p sont non nuls et :

$$\lim_{p\to+\infty} f_p^{-1} = \big(\lim_{p\to+\infty} f_p\big)^{-1}.$$

Démonstration. (i) La première assertion repose sur la relation :

$$\omega\big((\alpha f + \beta g) - (\alpha f_p + \beta g_p)\big) \geqslant \min\big(\omega(f-f_p), \omega(g-g_p)\big).$$

Pour la deuxième, on remarque que, pour p assez grand, $\omega(g_p) = \omega(g)$, et :

$$\omega(fg - f_p g_p) = \omega\big(f(g-g_p) + (f-f_p)g_p\big) \geqslant \min\big(\omega(g) + \omega(f-f_p), \omega(f) + \omega(g-g_p)\big).$$

Pour la troisième, on invoque un argument similaire et la relation :

$$\omega(f^{-1} - f_p^{-1}) = \omega(f_p - f) - \omega(f f_p) = \omega(f-f_p) - 2\omega(f).$$

■

Interprétation topologique.
Fixons $a \in]1, +\infty[$ et posons $|f| := a^{-\omega(f)}$ et en particulier $|0| = 0$.
L'application $f \mapsto |f|$ de $K((X))$ dans $\mathbb{R}_+$ vérifie :
(i) $|f| = 0 \Leftrightarrow f = 0$; (ii) $|fg| = |f|\,|g|$; (iii) $|f+g| \leqslant |f| + |g|$.
On dit que c'est une *valeur absolue* sur $K((X))$. Il en découle immédiatement que $d(f,g) := |f-g|$ définit une *distance* sur $K((X))$ (module II.1) et la notion de limite ci-dessus est celle associée à l'espace métrique ainsi construit. Si l'on utilise un autre réel $a' > 1$ à la place de a, la distance d' obtenue vérifie : $\dfrac{\ln d'(f,g)}{\ln d(f,g)} = \dfrac{a'}{a}$, et les *topologies* correspondantes sont les mêmes.

Le lecteur attentif aura remarqué que l'inégalité du triangle (iii) est ici remplacée par la relation plus forte : $|f+g| \leqslant \max\big(|f|, |g|\big)$. La distance vérifie de même : $d(f,h) \leqslant \max\big(d(f,g), d(g,h)\big)$. On dit que la valeur absolue et la distance sont *ultramétriques* ou *non archimédiennes* (voir également l'exercice II.1.5 de la page 469).

Toute série formelle est limite de polynômes : $\lim\limits_{N\to+\infty} \sum\limits_{n=0}^{N} a_n X^n = \sum\limits_{n\geqslant 0} a_n X^n$. On dit que $K[X]$ est *dense* dans $K[[X]]$. Certaines constructions pourront donc être effectuées sur les polynômes, avant d'être étendues aux séries formelles par *continuité*.

De même que l'on avait *complété* $\mathbb{Q}$ en $\mathbb{R}$, on a ici complété $K(X)$ (et $K[X]$), mais il y a plusieurs complétions possibles : $K[[X]]$, les $K[[X-a]]$ et $K[[\xi]]$.

4.2.1 Critères de convergence

Le théorème suivant donne l'analogue du critère de Cauchy dans $K[[X]]$. L'extrême simplicité de la condition vient de ce que la distance est ultramétrique.

Théorème 60 (Critère de Cauchy pour les séries formelles).
Dans $K[[X]]$, pour qu'une suite $(f_p)_{p\in\mathbb{N}}$ converge, il faut, et il suffit, que $\lim\limits_{p\to+\infty}(f_{p+1} - f_p) = 0$; pour qu'une série $\sum\limits_{p\geqslant 0} g_p$ converge, il faut, et il suffit, que $\lim\limits_{p\to+\infty} g_p = 0$.

Démonstration.
Les deux affirmations sont manifestement équivalentes (poser $g_p := f_{p+1} - f_p$). Prouvons la première. Si $\lim\limits_{p\to+\infty} f_p = f$, alors $\lim\limits_{p\to+\infty} f_{p+1} = f$, d'où $\lim\limits_{p\to+\infty}(f_{p+1} - f_p) = f - f = 0$. Supposons réciproquement que $\lim\limits_{p\to+\infty}(f_{p+1} - f_p) = 0$. Pour tout indice $n \in \mathbb{N}$, on a $\omega(f_{p+1} - f_p) > n$ à partir d'un certain rang $p := p_n$ (dépendant de n). Tous les f_p pour $p \geqslant p_n$ ont donc même coefficient d'indice n : notons a_n ce coefficient et posons $f := \sum\limits_{n\geqslant 0} a_n X^n$. Alors, pour $p \geqslant \max(p_0, \ldots, p_n)$, f_p et f ont mêmes coefficients d'indices $0, \ldots, n$, et l'on a donc $\omega(f - f_p) > n$. On en déduit que $\lim\limits_{p\to+\infty} f_p = f$. ∎

Corollaire 61. Soit $(h_p)_{p\in\mathbb{N}}$ une suite d'éléments non nuls de $K[[X]]$. Pour que le produit infini $\prod_{p\geqslant 0} h_p$ converge, il faut, et il suffit, que $\sum_{p\geqslant 0} \omega(h_p) = +\infty$ (et dans ce cas, le produit est nul) ; ou bien que $\lim_{p\to+\infty} h_p = 1$ (et dans ce cas, le produit est non nul).

Démonstration. La valeur du produit infini est, par définition, la limite (si elle existe) des produits partiels $\prod_{p=0}^{N} h_p$. Puisque $\omega\left(\prod_{p=0}^{N} h_p\right) = \sum_{p=0}^{N} \omega(h_p)$, il est clair que $\prod_{p\geqslant 0} h_p = 0 \Leftrightarrow \sum_{p\geqslant 0} \omega(h_p) = +\infty$.

Si $\prod_{p\geqslant 0} h_p = h \neq 0$, alors $\lim_{p\to+\infty} h_p = \frac{h}{h} = 1$. Si $\lim_{p\to+\infty} h_p = 1$, alors, en posant $f_p := \prod_{i=0}^{p} h_i$, on voit que $\omega(f_p - f_{p-1}) \geqslant \omega(h_p - 1)$, qui tend vers l'infini, et l'on peut appliquer le théorème. ∎

Exemple. Le produit $\prod_{p\geqslant 0}(1 + X^{2^p})$ converge d'après le critère ci-dessus. Comme $1 + X^{2^p} = \frac{1 - X^{2^{p+1}}}{1 - X^{2^p}}$, ce produit infini vaut $\frac{1}{1-X} = \sum_{n\geqslant 0} X^n$. En le développant, on voit que cette égalité revient à affirmer l'existence et l'unicité de l'écriture en base 2 de tout entier naturel n.

L'algorithme de Babylone. On suppose ici le corps K de caractéristique différente de 2. Nous voulons calculer les racines carrées d'une série de Laurent. Puisque $K((X))$ est de caractéristique différente de 2, une série de Laurent qui admet une racine carrée en admet exactement deux, et elles sont opposées. Si $f := \sum_{n\geqslant n_0} a_n X^n \in K((X))$ est le carré de $g := \sum_{n\geqslant p_0} b_n X^n \in K((X))$, alors $\omega(f) = 2\omega(g)$ et le terme constant de $X^{-\omega(f)} f$ est le carré du terme constant de $X^{-\omega(g)} g$. Il est donc *nécessaire* que $\omega(f)$ soit pair et que le coefficient $\left[X^{\omega(f)}\right] f$ soit un carré. On est donc ramené à mettre f sous la forme $b^2 X^{2m}(1 + c_1 X + \cdots)$ et à chercher si ce dernier facteur est un carré.
Nous supposerons donc d'emblée que $f = 1 + a_1 X + \cdots$ et nous prouverons l'existence d'une unique racine carrée $g = 1 + b_1 X + \cdots$ On définit une suite $(g_p)_{p\in\mathbb{N}}$ de séries formelles par $g_0 := 1$ et $g_{p+1} := \frac{1}{2}\left(g_p + \frac{f}{g_p}\right)$. On vérifie d'abord par récurrence que tous les g_p ont pour terme constant 1 (en particulier, ils sont inversibles et la suite est bien définie dans $K[[X]]$). Un petit calcul donne $g_{p+1}^2 - f = \frac{1}{4g_p^2}(g_p^2 - f)^2$, d'où, puisque $\omega(g_p) = 0$, $\omega(g_{p+1}^2 - f) = 2\omega(g_p^2 - f)$. Comme $\omega(g_0^2 - f) \geqslant 1$, on a $\omega(g_p^2 - f) \geqslant 2^p$ (donc $g_p^2 \to f$). On en tire d'abord $\omega(g_{p+1}^2 - g_p^2) \geqslant 2^p$, puis $\omega(g_{p+1} - g_p) \geqslant 2^p$ (puisque le terme constant

de $g_{p+1}+g_p$ est 2). La suite g_p converge donc vers g et la suite g_p^2 vers $g^2=f$. La convergence est d'ailleurs rapide, puisque $\omega(g_p-g)\geqslant 2^p$.

Lorsque $\mathrm{tc}(f)=1$, nous appellerons *racine carrée de* f , et nous noterons $\sqrt{f}$, l'unique série formelle g telle que $\mathrm{tc}(g)=1$ et $g^2=f$.

Revenant au cas général $f=aX^{2m}+\cdots$, avec $a=b^2\neq 0$, on peut appliquer ce qui précède à $\dfrac{f}{bX^{2m}}$ et $\dfrac{g}{bX^m}$. On posera donc $g_0:=bX^m$, puis $g_{p+1}:=\dfrac{1}{2}\left(g_p+\dfrac{f}{g_p}\right)$ (même relation de récurrence). On en déduit que g_p converge vers l'unique série $g=bX^m+\cdots$ telle que $g^2=f$, et que $\omega(g-g_p)\geqslant m+2^p$.

Exercice 24.

À l'aide d'une équation différentielle, calculer $\sqrt{1-4X}$.

Solution.

Si $g^2=f$, alors $2\dfrac{g'}{g}=\dfrac{f'}{f}=\dfrac{-4}{1-4X}$, d'où l'équation différentielle $(1-4X)g'=-2g$, et la condition initiale $g(0)=1$. On écrit $g=b_0+b_1X+\cdots$, où $b_0=1$. Par identification des termes en X^n, on trouve $(n+1)b_{n+1}-4nb_n=-2b_n$, d'où la relation de récurrence : $b_{n+1}:=2\dfrac{2n-1}{n+1}b_n$.

En itérant, on en déduit : $b_1=-2$ et, pour $n\geqslant 2$: $b_n=-2^n\dfrac{1.3\cdots(2n-3)}{n!}$.

Le numérateur est égal à $\dfrac{(2n-2)!}{2.4\cdots(2n-2)}=\dfrac{(2n-2)!}{2^{n-1}(n-1)!}$, d'où, en fin de compte :

$$\sqrt{1-4X}=1-2\sum_{n\geqslant 1}\frac{1}{n}\binom{2n-2}{n-1}X^n.$$

Application aux nombres de Catalan.

Ces nombres ont été définis dans le module « Fondements » de [L1]. On a $c_0=1$ et, pour $n\geqslant 1$, $c_n=\sum_{i+j=n-1}c_ic_j$.

On introduit leur *série génératrice* $C:=\sum_{n=0}c_nX^n\in\mathbb{Q}[[X]]$. Alors :

$$C=1+\sum_{n\geqslant 1}\Big(\sum_{i+j=n-1}c_ic_j\Big)X^n=1+X\sum_{n\geqslant 0}\Big(\sum_{i+j=n}c_ic_j\Big)X^n=1+XC^2.$$

On en déduit que $(1-2XC)^2=1-4X$ (c'est la tactique usuelle de mise sous forme canonique d'un trinôme du second degré), d'où $1-2XC=\sqrt{1-4X}$ (car son coefficient constant est 1), d'où $C=\dfrac{1-\sqrt{1-4X}}{2X}$. On peut alors appliquer l'exercice 24 et en tirer : $C=\sum_{n\geqslant 0}\dfrac{1}{n+1}\binom{2n}{n}X^n$, d'où enfin : $\forall n\geqslant 0\ ,\ c_n=\dfrac{1}{n+1}\binom{2n}{n}$.

4.2.2 La méthode du point fixe

Nous dirons qu'une application $\Phi : K[[X]] \to K[[X]]$ est *contractante* si, quels que soient $f, g \in K[[X]]$, on a : $\omega\big(\Phi(f) - \Phi(g)\big) \geqslant \omega(f-g)+1$. En termes métriques, cela équivaut, pour $f \neq g$, à $d\big(\Phi(f), \Phi(g)\big) < d(f,g)$. Nous dirons que f est un *point fixe* de Φ si $\Phi(f) = f$.

Théorème 62 (Théorème du point fixe pour les séries formelles).
Si Φ est contractante, elle admet un unique point fixe $f \in K[[X]]$. De plus, quel que soit $f_0 \in K[[X]]$, la suite définie par $f_{p+1} := \Phi(f_p)$ converge vers f.

Démonstration.
Si f et g sont des points fixes, on a $\omega(f-g) = \omega\big(\Phi(f) - \Phi(g)\big) \geqslant \omega(f-g)+1$, ce qui n'est possible que si $\omega(f-g) = +\infty$, *i.e.* $f = g$.
Soit $f_0 \in K[[X]]$ arbitraire. On va montrer que la suite définie par $f_{p+1} := \Phi(f_p)$ converge et que sa limite est un point fixe. On a $\omega(f_{p+2} - f_{p+1}) \geqslant \omega(f_{p+1} - f_p) + 1$, d'où, en itérant, $\omega(f_{p+1} - f_p) \geqslant \omega(f_1 - f_0) + p$. Le critère de Cauchy s'applique donc et la suite converge vers $f \in K[[X]]$. Alors $\omega(\Phi(f) - f_{p+1}) = \omega\big(\Phi(f) - \Phi(f_p)\big) \geqslant \omega(f - f_p) + 1$, ce qui montre que la même suite converge vers $\Phi(f)$, qui est donc égal à f. ∎

Exemples. (i) Soit $h := 1 - Xg$. Pour inverser h, on résout $(1 - Xg)f = 1$, soit $f = \Phi(f) := 1 + Xgf$. Comme $\Phi(f_1 - f_2) = Xg(f_1 - f_2)$, il est clair que Φ est contractante. Si l'on itère à partir de $f_0 := 0$, on trouve $f_p = 1 + Xg + \cdots + X^{p-1}g^{p-1}$.
(ii) L'équation différentielle $X^2 f' - f + X = 0$ a une unique solution, qui est le point fixe de l'opérateur $\Phi(f) := X + X^2 f'$. Celui-ci est contractant, car $\Phi(f_1 - f_2) = X^2(f_1' - f_2')$ et l'on applique la proposition 56 de la page 36. (Voir l'exercice I.1.54 de la page 56.)
(iii) L'équation algébrique $Xf^2 - f + 1$ a une unique solution, qui est le point fixe de l'opérateur $\Phi(f) := 1 + Xf^2$, dont le lecteur vérifiera qu'il est bien contractant. C'est la série génératrice des nombres de Catalan (voir page 42).

Équations algébriques

Nous allons démontrer un analogue formel du théorème des fonctions implicites.

Théorème 63. Soient $A_i := \sum_{n \geqslant 0} \alpha_{i,n} X^n \in K[[X]]$ $(0 \leqslant i \leqslant d)$ des séries formelles et $\varphi(X,Y) := \sum_{i=0}^{d} A_i(X) Y^i \in K[[X]][Y]$. Soit α une racine simple du polynôme $\varphi(0,Y) := \sum_{i=0}^{d} A_i(0) Y^i \in K[Y]$. Il existe une unique solution $f \in K[[X]]$ de l'équation algébrique $\varphi(X,f) = 0$ telle que $f(0) = \alpha$.

Démonstration. L'équation $\varphi(X,f) = 0$ signifie, bien entendu, $\sum_{i=0}^{d} A_i(X) f^i = 0$. Commençons par le cas particulier $\alpha = 0$. L'hypothèse sur $\varphi(0,Y)$ signifie que $A_0(0) = 0$ et $A_1(0) \neq 0$.

La série A_1 est donc inversible. De plus, on cherche f tel que $f(0) = 0$, ce qui revient à poser $f = Xg$ et à chercher $g \in K[[X]]$ tel que $g = \Phi(g) := B_0 + B_2 g^2 + \cdots + B_d g^d$, où $B_0 = -X^{-1}\dfrac{A_0}{A_1} \in K[[X]]$ et où , pour $i \geqslant 2$, $B_i = -X^{i-1}\dfrac{A_i}{A_1} \in K[[X]]$ est d'ordre strictement positif. Quels que soient $g_1, g_2 \in K[[X]]$, $\Phi(g_1) - \Phi(g_2)$ est donc multiple de $X(g_1 - g_2)$: l'opérateur Φ est contractant. La conclusion découle alors du théorème 62 de la page précédente.

Dans le cas général, on pose $Y := \alpha + Z$, et $\varphi(X, Y) := \sum_{i=0}^{d} B_i(X) Z^i = \psi(X, Z)$, où $B_i := \sum_{j=i}^{d} \binom{j}{i} \alpha^{j-i} A_j$, et l'on cherche f sous la forme $\alpha + g$. On veut donc $g(0) = 0$, $\psi(X, g) = 0$ et l'on sait que 0 est racine simple de $\psi(0, Z)$: on est ramené au cas particulier précédent. ∎

Notons que cela ne marche pas avec $\varphi(X, Y) := Y^2 - X$: il n'existe pas de $f \in K[[X]]$ tel que $f^2 = X$. Mais il est vrai que le polynôme $\varphi(0, Y)$ n'a pas de racine simple !

4.2.3 Composition des séries formelles

Soient $f := \sum_{n \geqslant 0} a_n X^n$ et $g := \sum_{n \geqslant 0} b_n X^n$ des séries formelles. Nous voulons donner un sens à la composition $f \circ g = \sum_{n \geqslant 0} a_n g^n$, comme nous l'avons fait pour les polynômes. Comme la somme est *a priori* infinie, le terme constant $(f \circ g)(0) = \sum_{n \geqslant 0} a_n b_0^n$ n'est pas défini, sauf si l'on suppose $b_0 = 0$.

Théorème et définition 64. On suppose que $\omega(g) \geqslant 1$. Le morphisme de K-algèbres $P \mapsto P(g)$ de $K[X]$ dans $K[[X]]$ s'étend en un unique morphisme de $K[[X]]$ dans $K[[X]]$. L'image de la série formelle $f := \sum_{n \geqslant 0} a_n X^n$ est la somme de la série $\sum_{n \geqslant 0} a_n g^n$, qui est convergente. On la note $f \circ g$ et on l'appelle *composée de f et de g* ou *substitution de g dans f*. On écrit aussi abusivement $f\big(g(X)\big)$ pour $f \circ g$.

Démonstration. Nous laissons au lecteur la preuve (calculatoire mais facile) que la série $\sum_{n \geqslant 0} a_n g^n$ est convergente et que l'application $\sum_{n \geqslant 0} a_n X^n \mapsto \sum_{n \geqslant 0} a_n g^n$ est bien un morphisme. Montrons que c'est le seul possible. Soit F l'image de f par un tel morphisme φ. Notons $P_N := \sum_{n=0}^{N} a_n X^n$. Alors X^{N+1} divise $f - P_N$, donc $\varphi(X^{N+1}) = g^{N+1}$ divise $\varphi(f - P_N) = F - P_N(g)$, qui est donc d'ordre supérieur ou égal à $\omega(g^{N+1}) \geqslant N + 1$. On a donc nécessairement $F = \lim_{N \to +\infty} P_N(g) = \sum_{n \geqslant 0} a_n g^n$ (c'est donc un raisonnement par continuité). ∎

Corollaire 65. L'application $f \mapsto f \circ g$ est l'unique endomorphisme de la K-algèbre $K[[X]]$ tel que X ait pour image g.

Proposition 66. (i) L'ordre de la composée est $\omega(f\circ g)=\omega(f)\omega(g)$.
(ii) On a $X\circ g=g$ et $f\circ X=f$.
(iii) Si $\omega(g),\omega(h)\geqslant 1$, on a $f\circ(g\circ h)=(f\circ g)\circ h$.
(iv) On a $(f\circ g)'=(f'\circ g)g'$.

Démonstration. La preuve de (i) et (ii) est laissée au lecteur.
(iii) Si l'on compose les morphismes $f\mapsto f\circ g$ et $f\mapsto f\circ h$, on obtient le morphisme $f\mapsto(f\circ g)\circ h$. Comme l'image de X par ce dernier est $g\circ h$, c'est le morphisme $f\mapsto f\circ(g\circ h)$ (corollaire 65 de la page précédente).
(iv) La formule à démontrer est vraie lorsque f est un polynôme (règles de calcul usuelles sur la dérivée). On reprend les notations du théorème 64 de la page ci-contre. Alors $(f\circ g)'-(f'\circ g)g'$ diffère de $(P_N\circ g)'-(P_N'\circ g)g'=0$ par un multiple de X^N quel que soit N, donc est nul (encore un raisonnement par continuité). ∎

Si $g=0$, le morphisme $f\mapsto f\circ g$ est le morphisme $f\mapsto \mathrm{tc}(f)=f(0)$ déjà rencontré. Si $g\neq 0$, la formule $\omega(f\circ g)=\omega(f)\omega(g)$ montre que le morphisme $f\mapsto f\circ g$ est injectif. Si de plus $\omega(g)\geqslant 2$, il n'est pas surjectif, puisque l'ordre d'un élément de l'image est multiple de $\omega(g)$. Reste le cas où $\omega(g)=1$, ce qui équivaut à $g'(0)\neq 0$.

Proposition et définition 67. Pour que l'endomorphisme $f\mapsto f\circ g$ de la K-algèbre $K[[X]]$ soit un automorphisme, il faut, et il suffit, que $g'(0)$ soit non nul, autrement dit, $\omega(g)=1$. L'automorphisme réciproque est alors de la forme $f\mapsto f\circ h$, et $h\circ g=g\circ h=X$. La série h est appelée *réciproque* de la série g.

Démonstration. On vient de voir que la condition était nécessaire.
Supposons que $g=b_1X+\sum_{n\geqslant 2}b_nX^n$, avec $b_1\neq 0$. L'équation $g\circ h=X$ équivaut à $h=b_1^{-1}\big(X-\sum_{n\geqslant 2}b_nh^n\big)$.

En posant $h=Xf$, cela équivaut à $f=\Phi(f):=b_1^{-1}\big(1-\sum_{n\geqslant 2}b_nX^{n-1}f^n\big)$. Comme Φ est contractante, il y a bien une solution.
Cela entraîne que pour tout f, on a $f\circ g\circ h=f$. En particulier, en l'appliquant à $f\circ h$, on trouve $f\circ h\circ g\circ h=f\circ h$, donc $f\circ h\circ g=f$ par injectivité de $f\mapsto f\circ h$ (car $h\neq 0$). Toutes les affirmations s'ensuivent. ∎

4.2.4 Logarithme et exponentielle

On suppose K de caractéristique nulle. Dans toutes les formules qui suivent, nous noterons :
$XK[[X]]:=\{g\in K[[X]]\mid \mathrm{tc}(g)=0\}$ et $1+XK[[X]]:=\{h\in K[[X]]\mid \mathrm{tc}(h)=1\}$.

On a introduit dans l'exemple de la page 37 la série : $L(X):=\sum_{n\geqslant 1}\dfrac{(-1)^{n-1}}{n}X^n$, qui est l'unique primitive de $\dfrac{1}{1+X}$ élément de $XK[[X]]$; et, dans l'exemple de la page 38 la série : $E(X):=\sum_{n\geqslant 0}\dfrac{1}{n!}X^n$, qui est l'unique élément de $1+XK[[X]]$ égal à sa dérivée.

Théorème et définition 68. (i) Pour toute série $h \in 1 + XK[[X]]$, la série $L \circ (h-1)$ est l'unique élément de $XK[[X]]$ de dérivée $\dfrac{h'}{h}$; on la note $\ln h$ et on l'appelle *logarithme de* h. On a donc :

$$\ln h := \sum_{n \geqslant 1} \frac{(-1)^{n-1}}{n}(h-1)^n.$$

Pour toute série $g \in XK[[X]]$, la série $E \circ g$ est l'unique élément $f \in 1 + XK[[X]]$ tel que $\dfrac{f'}{f} = g$; on la note $\exp h$ et on l'appelle *exponentielle de* g. On a donc :

$$\exp g := \sum_{n \geqslant 0} \frac{1}{n!} g^n.$$

(ii) Les applications $g \mapsto \exp(g)$ de $XK[X]$ dans $1 + XK[[X]]$ et $h \mapsto \ln(h)$ de $1 + XK[X]$ dans $XK[[X]]$ sont réciproques l'une de l'autre.
(iii) On a les égalités :

$$\forall g_1, g_2 \in XK[[X]]\ ,\ \exp(g_1 + g_2) = \exp(g_1)\exp(g_2) \quad \text{et}$$
$$\forall h_1, h_2 \in 1 + XK[[X]]\ ,\ \ln(f_1 f_2) = \ln f_1 + \ln f_2.$$

Démonstration. (i) On applique la proposition 66 de la page précédente.
(ii) La dérivée de $\ln(\exp g)$ est $\dfrac{g' \exp(g)}{\exp(g)} = g'$. Les séries $\ln(\exp g)$ et g ont même dérivée et même terme constant 0, elles donc sont égales. La dérivée de $f := \exp(\ln h)$ est $\dfrac{h'}{h} f$. On a donc $\left(\dfrac{h}{f}\right)' = 0$. Comme f et h ont même terme constant 1, elles sont égales.
(iii) Les deux égalités se prouvent par dérivation et comparaison des termes constants. Les preuves sont laissées en exercice. ∎

On voit donc en particulier que $L = \ln(1+X)$ et $E = \exp(X)$.

Corollaire 69.
Les applications ci-dessus induisent un isomorphisme entre le groupe $XK[[X]]$ (muni de l'addition) et le groupe $1 + XK[[X]]$ (muni de la multiplication).

4.2.5 La formule du binôme généralisée

Proposition et définition 70.
Soient $h \in 1 + XK[[X]]$ et $\alpha \in K$. La série formelle $h^\alpha := \exp\big(\alpha \ln(h)\big)$ est l'unique série de terme constant 1 dont la dérivée logarithmique est $\alpha \dfrac{h'}{h}$.

Démonstration. Le fait que h^α vérifie les conditions découle par un calcul simple du théo-

rème 68 de la page ci-contre. Pour toute série H telle que $\dfrac{H'}{H} = \alpha\dfrac{h'}{h}$, on vérifie facilement que la dérivée logarithmique de $\dfrac{H}{h^\alpha}$ est nulle, donc que $H = H(0)h^\alpha$, d'où l'unicité. ∎

Corollaire 71. Soient $h, k \in 1 + XK[[X]]$ et $\alpha, \beta \in K$. On a alors les égalités :

$$(h^\alpha)' = \alpha h^{\alpha-1}h', \quad h^{\alpha+\beta} = h^\alpha h^\beta, \quad (hk)^\alpha = h^\alpha k^\alpha \quad \text{et} \quad (h^\alpha)^\beta = h^{\alpha\beta}.$$

Démonstration. La première formule découle immédiatement des règles de calcul. Les autres s'obtiennent en comparant dérivées logarithmiques et termes constants des deux membres. ∎

On en déduit par exemple que $h^{1/2} = \sqrt{h}$, que $h^{1/q}$ est l'unique série de terme constant 1 dont la puissance $q^{ème}$ est égale à h, etc.

Proposition 72 (Formule du binôme généralisé). On a l'égalité :

$$(1+X)^\alpha = \sum_{n\geqslant 0} \binom{\alpha}{n} X^n.$$

Démonstration. La série $h := (1+X)^\alpha$ vérifie $h(0) = 1$ et $\dfrac{h'}{h} = \dfrac{\alpha}{1+X}$, et l'on peut appliquer l'exercice 22 de la page 38. ∎

Exemples. Du calcul fait dans l'exercice 24 de la page 42, on tire :

$$\forall n \geqslant 1\ ,\ \binom{1/2}{n} = -2\left(\frac{-1}{4}\right)^n \frac{1}{n}\binom{2n-2}{n-1},$$

d'où

$$(1+X)^{1/2} = 1 - 2\sum_{n\geqslant 1}\left(\frac{-1}{4}\right)^n \frac{1}{n}\binom{2n-2}{n-1}X^n.$$

De l'égalité $\binom{\alpha}{n} = \dfrac{\alpha}{\alpha-n}\binom{\alpha-1}{n}$, on déduit alors :

$$\forall n \geqslant 0\ ,\ \binom{-1/2}{n} = \left(\frac{-1}{4}\right)^n \binom{2n}{n} \quad \text{et} \quad (1+X)^{-1/2} = \sum_{n\geqslant 0}\left(\frac{-1}{4}\right)^n \binom{2n}{n} X^n$$

(il y a un calcul de développement limité similaire dans le module « Approximation » de [L1]). De même, $\binom{3/2}{n} = \left(\dfrac{-1}{4}\right)^n \dfrac{12}{n(n-1)}\binom{2n-4}{n-2}$ (pour $n \geqslant 2$), etc. (exercice I.1.62 de la page 56).

Les *coefficients binomiaux* $\binom{\alpha}{n}$ sont d'un usage fréquent dans toutes les branches des mathématiques. Leur étude se base sur les propriétés de la fonction Gamma qui seront établies

en [MPA] (« Fonctions spéciales »). Voici quelques formules dont nous aurons l'usage dans ce volume.

Lemme 73. Pour tout $\alpha \in \mathbb{C} \setminus \mathbb{N}$, il existe une constante $C_\alpha \in \mathbb{C}^*$ telle que :

$$\binom{\alpha}{n} \sim \frac{(-1)^n C_\alpha}{n^{\alpha+1}} \quad \text{lorsque } n \to +\infty.$$

On a de plus $C_\alpha = -\alpha C_{\alpha-1}$.

Démonstration. Soit $u_n := (-1)^n n^{\alpha+1}\binom{\alpha}{n} \in \mathbb{C}^*$. Alors :

$$\frac{u_{n+1}}{u_n} = \frac{n-\alpha}{n+1}\left(1+\frac{1}{n}\right)^{\alpha+1} = 1 + \mathrm{O}\left(\frac{1}{n^2}\right),$$

et le produit infini $\displaystyle\prod_{n\geqslant 0} \frac{u_{n+1}}{u_n}$ converge dans $\mathbb{C}^*$ (voir l'exercice I.1.61 de la page 56). La deuxième formule vient alors des relations $\displaystyle\binom{\alpha}{n} = \frac{\alpha}{\alpha-n}\binom{\alpha-1}{n} \sim \frac{-\alpha}{n}\binom{\alpha-1}{n}$. ∎

Une expression exacte de C_α découlera de l'étude de la fonction Gamma (*cf.* le module « Suites et séries de fonctions », théorème 48 de la page 524). En voici des cas particuliers.

Proposition 74. On a les équivalents suivants lorsque $n \to +\infty$:

$$\binom{-1/2}{n} \sim \frac{(-1)^n}{\sqrt{\pi n}} \quad \left(\textit{i.e. } C_{-1/2} = \frac{1}{\sqrt{\pi}}\right),$$

$$\binom{1/2}{n} \sim \frac{(-1)^{n-1}}{2\sqrt{\pi n^3}} \quad \left(\textit{i.e. } C_{-1/2} = \frac{-1}{2\sqrt{\pi}}\right),$$

$$c_n \sim \frac{4^n}{\sqrt{\pi n^3}}.$$

Démonstration. Lors de l'étude de la *formule de Stirling* (module II.8, section 1.4), nous démontrerons la relation : $\displaystyle\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n}}$ $(n \to +\infty)$. Il suffit alors d'invoquer les exemples de la page 47. ∎

La conséquence suivante nous sera utile plus tard (lemme 16 de la page 911).

Corollaire 75. On a l'égalité et l'équivalence :

$$\binom{1/2}{n+1} = (-1)^n 2^{-2n-1}\frac{1}{n+1}\binom{2n}{n} \sim \frac{1}{2\sqrt{\pi n^3}}.$$

Exercice type corrigé

I.1.0 Le but de cet exercice est de déterminer les idéaux premiers de $\mathbb{Z}[X]$.
(i) Soit $\mathcal{P}$ un tel idéal. Vérifier que $\mathcal{P} \cap \mathbb{Z}$ est un idéal premier de $\mathbb{Z}$. Il est donc nul ou de la forme $p\mathbb{Z}$, $p \in \mathbb{N}$ premier : nous étudierons ici le deuxième cas seulement (le premier cas sera traité dans le chapitre d'arithmétique de [L2-L3]).
(ii) On suppose donc que $\mathcal{P} \cap \mathbb{Z} = p\mathbb{Z}$, $p \in \mathbb{N}^*$ premier. Montrer que les idéaux $\mathcal{P}$ de ce type sont en bijection avec les idéaux premiers de $\mathbb{F}_p[X]$. On rappelle que $\mathbb{F}_p$ désigne le corps fini $\mathbb{Z}/p\mathbb{Z}$.
(iii) En déduire que, soit $\mathcal{P} = p\mathbb{Z}[X]$, soit $\mathcal{P} = <p, f>$, où $f \in \mathbb{Z}[X]$ est un polynôme unitaire dont la réduction modulo p est un polynôme irréductible $\overline{f} \in \mathbb{F}_p[X]$.
(iv) Que peut-on dire de l'anneau quotient $\mathbb{Z}[X]/\mathcal{P}$ dans chacun de ces deux cas ?
(v) Expliciter le cas où $p = 2$ et $\deg f = 2$.
Solution. (i) Cela découle des exemples de la page 21.
(ii) Les idéaux $\mathcal{P}$ tels que $\mathcal{P} \cap \mathbb{Z} = p\mathbb{Z}$ sont ceux qui contiennent l'élément p, donc également ceux qui contiennent l'idéal $p\mathbb{Z}[X]$. La bijection invoquée (corollaire 34 de la page 21) fait se correspondre les idéaux premiers de $\mathbb{Z}[X]$ qui contiennent l'idéal $p\mathbb{Z}[X]$ avec les idéaux premiers de $\mathbb{Z}[X]/p\mathbb{Z}[X]$: $\mathcal{P}$ est l'image réciproque d'un tel idéal $\overline{\mathcal{P}}$ par la projection canonique. Or, $\mathbb{Z}[X]/p\mathbb{Z}[X] \simeq \mathbb{F}_p[X]$.
(iii) L'anneau $\mathbb{F}_p[X]$ est principal ; d'après le lemme d'Euclide ([L1], module « Groupes, anneaux, corps »), ses idéaux premiers sont 0 et les idéaux principaux $<\overline{f}>$, où $\overline{f}$ est irréductible dans $\mathbb{F}_p[X]$. Si $\overline{\mathcal{P}} = 0$, on a $\mathcal{P} = p\mathbb{Z}[X]$. Si $\overline{\mathcal{P}} = <\overline{f}>$, $\mathcal{P} = <p, f>$, où $f \in \mathbb{Z}[X]$ est un polynôme unitaire dont la réduction modulo p est $\overline{f}$.
(iv) Dans le premier cas, l'anneau quotient est l'anneau $\mathbb{Z}[X]/p\mathbb{Z}[X] \simeq \mathbb{F}_p[X]$, qui est intègre mais qui n'est pas un corps.
Dans le deuxième cas, l'anneau quotient est l'anneau $\mathbb{F}_p[X]/<\overline{f}>$. C'est un corps (quotient de l'anneau principal $\mathbb{F}_p[X]$ par un idéal premier non nul, donc maximal), et l'idéal $\mathcal{P}$ est donc maximal dans ce cas. Ce corps contient $\mathbb{F}_p$, et c'est une $\mathbb{F}_p$-algèbre de dimension $\deg \overline{f} = \deg f$. Comme $\mathbb{F}_p$-espace vectoriel , il est isomorphe à $\mathbb{F}_p^{\deg f}$, donc en bijection avec $\mathbb{F}_p^{\deg f}$: c'est donc un corps fini de cardinal $p^{\deg f}$.
(v) Les polynômes de degré 2 sur $\mathbb{F}_2$ sont X^2, $X^2 + X$, $X^2 + 1 = (X + 1)^2$ et $X^2 + X + 1$ qui n'a aucune racine, donc est irréductible (et c'est le seul). L'idéal premier correspondant de $\mathbb{Z}[X]$ est $<2, X^2+X+1>$. Le corps $\mathbb{Z}[X]/<2, X^2+X+1>$ est isomorphe à $\mathbb{F}_2[X]/<X^2+X+1>$, qui est une extension de degré 2 de $\mathbb{F}_2$, donc un corps à 4 éléments ; il est décrit concrètement à l'exercice I.1.26 de la page 53. C'est (à isomorphisme près) l'unique corps à 4 éléments (*cf.* [MPA1]) : il est noté $\mathbb{F}_4$.

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 Rappelons que, si a est un entier non nul et p un nombre premier, $v_p(a)$ désigne l'unique $k \in \mathbb{N}$ tel que $p^{-k}a$ est un entier non divisible par p. Pourquoi peut-on définir l'application $\dfrac{a}{b} \mapsto v_p(a) - v_p(b)$ de $\mathbb{Q}^*$ dans $\mathbb{Z}$?

Peut-on définir une application $\dfrac{a}{b} \mapsto \dfrac{a^2+1}{b^2+1}$ de $\mathbb{Q}$ dans lui-même ?

I.1.2 $*$ Soit a un réel. On considère l'ensemble E des couples (f, J) formés d'un intervalle ouvert $J \subset \mathbb{R}$ contenant a et d'une fonction f continue de J dans $\mathbb{R}$. Sur l'ensemble E, on définit deux lois de composition interne et une relation d'équivalence $\mathcal{R}$ (abrégée $\sim$) :

$$(f_1, J_1) + (f_2, J_2) := (g, K) \text{ où } K := J_1 \cap J_2 \quad \text{et} \quad g := (f_1)_{|K} + (f_2)_{|K}$$
$$(f_1, J_1)(f_2, J_2) := (h, K) \text{ où } K := J_1 \cap J_2 \quad \text{et} \quad h := (f_1)_{|K}(f_2)_{|K}$$
$$(f_1, J_1) \sim (f_2, J_2) \Longleftrightarrow \exists K \subset J_1 \cap J_2 \text{ (intervalle ouvert contenant } a) \ : \ (f_1)_{|K} = (f_2)_{|K}.$$

Vérifier que $\mathcal{R}$ est bien une relation d'équivalence. Démontrer qu'elle est compatible avec l'addition et la multiplication et que les lois quotients font de $E/\mathcal{R}$ un anneau commutatif unitaire (*anneau des germes de fonctions continues en* a). Démontrer que l'application $(f, J) \mapsto f(a)$ de E dans $\mathbb{R}$ est compatible avec la relation $\mathcal{R}$: on peut donc parler de la valeur en a d'un germe. Démontrer que les germes de fonctions de classe $\mathcal{C}^k$ au voisinage de a forment un sous-anneau.

I.1.3 L'élément $3 \in \mathbb{N}$ est régulier (*i.e.* simplifiable) pour la multiplication. Est-ce encore le cas après passage au quotient par $\equiv_9$?

I.1.4 $*$ (i) Soit G le groupe des fractions d'un groupe M (paragraphe 1.2.1 de la page 8). Montrer que le morphisme $M \to G$ est un isomorphisme.
(ii) Soit A un anneau commutatif intègre. Vérifier que la multiplication fait de $M = A\backslash\{0\}$ un monoïde commutatif. Vérifier que le groupe des fractions de M s'identifie au groupe multiplicatif K^* du corps des fractions K de A.

I.1.5 $*$ Soit $f : M \to G'$ un morphisme d'un monoïde commutatif M dans un groupe G' et soit $i : M \to G$ le morphisme de M dans son groupe des fractions. Vérifier qu'il existe un unique morphisme $g : G \to G'$ tel que $f = g \circ i$.

I.1.6 Soit S une partie multiplicative de l'anneau A et soit $i : A \to S^{-1}A$ le morphisme de A dans son anneau de fractions (paragraphe 1.2.2 de la page 9). Quels sont les éléments $a \in A$ tels que $i(a)$ est inversible dans $S^{-1}A$? Qu'obtient-on lorsque $S := A^*$? Pourquoi, sous les hypothèses du texte, la partie S ne contient-elle pas d'éléments nilpotents ? Qu'obtiendrait-on si l'on avait permis que $0 \in S$?

I.1.7 $*$ (i) Soit S une partie multiplicative de l'anneau A et soit $f : A \to A'$ un morphisme d'anneaux commutatifs tel que tous les éléments de $f(S)$ sont inversibles dans A'. Démontrer qu'il existe un unique morphisme d'anneaux $g : S^{-1}A \to A'$ tel que $f = g \circ i$.
(ii) Démontrer que, si $S \subset T$ sont deux parties multiplicatives, le morphisme d'anneaux $A \to T^{-1}A$ est le composé du morphisme d'anneaux $A \to S^{-1}A$ par un unique morphisme d'anneaux $S^{-1}A \to T^{-1}A$.
(iii) On suppose que T est le *saturé* de S, c'est à dire l'ensemble des diviseurs des éléments de S. Vérifier que c'est bien une partie multiplicative et que le morphisme $S^{-1}A \to T^{-1}A$ est un isomorphisme.
(iv) On suppose que A est intègre et que $T = A \setminus \{0\}$. Démontrer que l'on a ainsi défini une injection de $S^{-1}A$ dans le corps des fractions de A.

I.1.8 Justifier l'équivalence $a^{-1}b \in H \Leftrightarrow aH = bH$ dans la preuve du lemme 8.

I.1.9 Notons ι_a l'automorphisme $x \mapsto axa^{-1}$ de G. Montrer que $a \mapsto \iota_a$ est un morphisme de groupes de G dans le groupe $\mathrm{Aut}(G)$ des automorphismes de G. Quel est le noyau de ce morphisme ? Vérifier que le sous-groupe H de G est distingué si, et seulement s'il est stable par tous les automorphismes intérieurs.

I.1.10 Soit $f : G \to G'$ un morphisme de groupes, et soient $H \lhd G$, $H' \lhd G'$. Montrer que $f^{-1}(H') \lhd G$. Est-il exact que $f(H) \lhd G'$? Même question en supposant f surjectif.

I.1.11 Soient $m, p \in \mathbb{N}^*$ et $n = mp$. Montrer que l'application $x \mapsto mx \pmod n$ définit un morphisme de groupes surjectif de $\mathbb{Z}$ sur $m\mathbb{Z}/n\mathbb{Z}$, dont le noyau est $p\mathbb{Z}$. En déduire que le groupe $m\mathbb{Z}/n\mathbb{Z}$ est isomorphe au groupe $\mathbb{Z}/p\mathbb{Z}$.

I.1.12 $*$ Soient H et G' des sous-groupes de G tels que $H \lhd G$. Montrer que HG' est un sous-groupe de G et que $H \lhd HG'$. Montrer que le morphisme composé $G' \to HG' \to HG'/H$ est surjectif et de noyau $G' \cap H$. En déduire un isomorphisme : $\dfrac{G'}{G' \cap H} \simeq \dfrac{HG'}{H}$ (troisième théorème d'isomorphisme pour les groupes). Peut-on affaiblir l'hypothèse que H est distingué dans G ?

I.1.13 $*$ Montrer que tout morphisme de $\mathbb{C}/\mathbb{Z}$ dans $\mathbb{C}$ est trivial sur le sous-groupe $\mathbb{Q}/\mathbb{Z}$ de $\mathbb{C}/\mathbb{Z}$. On pourra montrer que $\mathbb{Q}/\mathbb{Z}$ est l'ensemble des éléments de torsion de $\mathbb{C}/\mathbb{Z}$ et que le groupe $\mathbb{C}$ est sans torsion (pour la terminologie, voir [L1], exercices du module d'algèbre linéaire générale). En déduire une nouvelle preuve qu'il n'existe pas de morphisme logarithme de $\mathbb{C}^*$ dans $\mathbb{C}$.

I.1.14 Soient $m, p \in \mathbb{N}^*$ et $n = mp$. Montrer que le morphisme d'anneaux $\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$ passe au quotient en un morphisme surjectif d'anneaux $\mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z}$, dont le noyau est l'idéal $m\mathbb{Z}/n\mathbb{Z}$ de $\mathbb{Z}/n\mathbb{Z}$.

I.1.15 En prenant $A = \mathbb{Z}[X]$ (sous-anneau de $\mathbb{R}[X]$ formé des polynômes à coefficients entiers), $B = \mathbb{R}[X]$ et $J = (X^2+1)\mathbb{R}[X]$, montrer que l'on a $A \cap J = (X^2+1)\mathbb{Z}[X]$. On en déduit donc un morphisme injectif : $\dfrac{\mathbb{Z}[X]}{(X^2+1)\mathbb{Z}[X]} \to \dfrac{\mathbb{R}[X]}{(X^2+1)\mathbb{R}[X]}$.
On a identifié dans le texte ce dernier anneau à $\mathbb{C}$. Quel sous-anneau obtient-on ainsi ?

I.1.16 Soit A un anneau non nécessairement commutatif. On appelle *idéal bilatère* de A un sous-groupe I de $(A,+)$ tel que : $\forall a \in I, \forall c \in A\ ,\ ac \in I$ et $ca \in I$. Démontrer que la relation $a - b \in I$ est alors compatible avec l'addition et la multiplication et permet de définir un anneau quotient. Réciproque ?

I.1.17 $*$ Démontrer que les seuls idéaux bilatères de $M_n(K)$ sont $\{0\}$ et $M_n(K)$. Démontrer que, si E est de dimension infinie, l'ensemble des endomorphismes de E de rang fini est un idéal bilatère de l'anneau $\mathcal{L}_K(E)$ qui n'est ni nul ni égal à $\mathcal{L}_K(E)$.

I.1.18 Soient E et F deux sous-espaces vectoriels de V.
Montrer que l'application $E \to \dfrac{E+F}{F}$ a pour noyau $E \cap F$ et induit par passage au quotient un isomorphisme de $\dfrac{E}{E \cap F}$ sur $\dfrac{E+F}{F}$ (troisième théorème d'isomorphisme pour les espaces vectoriels).

I.1.19 Soit P un polynôme unitaire de degré n sur le corps K. Démontrer que le K-espace vectoriel $K[X]/<P>$ est isomorphe à $K_{n-1}[X]$. En déduire que tous les idéaux non nuls de $K[X]$ sont des sous-espaces vectoriels de codimension finie.

I.1.20 Déduire de la proposition 29 de la page 18 que, pour que l'intersection de k hyperplans soit un sous–espace vectoriel de codimension k, il faut, et il suffit, que des formes linéaires définissant ces hyperplans soient linéairement indépendantes.

I.1.21 $*$ Soit F un sous–espace vectoriel de l'espace vectoriel de dimension finie E. Soient $\mathcal{C}$ une base de F et $\overline{\mathcal{D}}$ une base de E/F.
(i) On relève $\overline{\mathcal{D}}$ en une famille $\mathcal{D}$: autrement dit, on choisit un antécédent arbitraire dans E pour chaque élément de $\overline{\mathcal{D}}$. Montrer que $\mathcal{D}$ est libre et engendre un supplémentaire G de F. On obtient donc une base $\mathcal{B}$ en juxtaposant les bases $\mathcal{C}$ et $\mathcal{D}$.
(ii) Soit f un endomorphisme de E et soit M sa matrice dans la base $\mathcal{B}$. À quelle condition sur M le sous–espace vectoriel F est-il stable par f ? Décrire alors la matrice de la restriction $f_{|F}$ dans la base $\mathcal{C}$ et la matrice de l'endomorphisme $\overline{f}$ de E/F obtenu par passage au quotient dans la base $\mathcal{D}$.

Exercices sur la section 2

I.1.22 Dans le sous-anneau $A := \mathbb{Z}[X]$ (polynômes à coefficients entiers) de $B := \mathbb{Q}[X]$, les idéaux principaux $I := 2A$ (polynômes dont tous les coefficients sont pairs) et $J := XA$ (polynômes sans terme constant) ont pour somme l'idéal non principal $I + J = 2A + XA$ (polynômes P tels que $P(0)$ est pair) ; pour intersection et pour produit un idéal principal $IJ = I \cap J = 2XA$. Les idéaux de B qu'ils engendrent sont $IB = B$ et $JB = XB$. Les quotients sont respectivement isomorphes à $A/I \simeq \mathbb{F}_2[X]$, à $A/J \simeq \mathbb{Z}$ et à $A/(I+J) \simeq \mathbb{F}_2$ (on note $\mathbb{F}_2$ le corps $\mathbb{Z}/2\mathbb{Z}$). Vérifier (en les calculant) que $(I+J)^2 = I^2 + IJ + J^2$.

I.1.23 Dans l'anneau A, soient I l'idéal engendré par les x_i et J l'idéal engendré par les y_j. Montrer que IJ est engendré par les $x_i y_j$. Vérifier que le carré de l'idéal $<a, b>$ est $<a^2, ab, b^2>$. Que vaut I^n ? Montrer que l'idéal $I := <1 + \sqrt{-5}, 1 - \sqrt{-5}>$ de l'anneau $A = \mathbb{Z}[\sqrt{-5}]$ n'est pas principal, mais que $I^2 = <2>$ l'est.

I.1.24 $*$ (i) Soient A_1 et A_2 deux anneaux commutatifs et soient I_1 et I_2 des idéaux respectifs de A_1 et de A_2. Montrer que $I_1 \times I_2$ est un idéal de l'anneau produit $A_1 \times A_2$, et que le quotient correspondant s'identifie à l'anneau produit $A_1/I_1 \times A_2/I_2$. Démontrer que tout idéal de $A_1 \times A_2$ est de cette forme.
(ii) On suppose que I_1 et I_2 sont respectivement des idéaux maximaux de A_1 et de A_2. Montrer que $I_1 \times A_2$ et $A_1 \times I_2$ sont des idéaux maximaux de l'anneau produit $A_1 \times A_2$, de quotients respectifs A_1/I_1 et A_2/I_2. Réciproque ?
(iii) Mêmes questions pour des idéaux premiers.

I.1.25 $*$ Soient K un corps et $P \in K[X]$ un polynôme non nul. Démontrer que l'anneau $K[X]/ < P >$ est un produit de corps si, et seulement si, P est un produit de polynômes irréductibles distincts.

I.1.26 (i) Décrire concrètement le corps $\mathbb{F}_2[X]/<X^2 + X + 1>$.
(ii) Décrire de même un corps à 8 éléments et un corps à 9 éléments.

I.1.27 $**$ (i) Soit $A = \mathcal{C}([a, b], \mathbb{R})$ l'anneau des fonctions continues du segment $[a, b]$ dans $\mathbb{R}$. Pour tout $x \in [a, b]$, on définit l'idéal $\mathfrak{M}_x$ des fonctions continues nulles en x. Démontrer que $\mathfrak{M}_x$ est un idéal maximal de A.
(ii) Soit I un idéal propre de A. Démontrer qu'alors :

- soit toute fonction $f \in I$ s'annule en un point au moins du segment $\left[a, \dfrac{a+b}{2}\right]$;
- soit toute fonction $f \in I$ s'annule en un point au moins du segment $\left[\dfrac{a+b}{2}, b\right]$.

En itérant cet argument, construire des segments emboîtés $[a_n, b_n]$ de longueurs $2^{-n}(b - a)$ et tels que toute fonction $f \in I$ s'annule en un point au moins du segment $[a_n, b_n]$. En déduire l'existence d'un $x \in [a, b]$ tel que toute fonction $f \in I$ s'annule en x, puis conclure que les $\mathfrak{M}_x$ sont les seuls idéaux maximaux de A.
(iii) Soit $B = \mathcal{C}(]a, b[, \mathbb{R})$ et soit I l'idéal des fonctions de B nulles en dehors d'un segment $[c, d] \subset]a, b[$. Démontrer que I n'est inclus dans aucun idéal de la forme $\mathfrak{M}_x$.

I.1.28 Soient A un anneau commutatif et $a \in A$. Montrer que a est inversible si, et seulement s'il n'appartient à aucun idéal maximal de A. On dit que A est *local* s'il admet un unique idéal maximal (par exemple, un corps). Montrer que cela équivaut à la propriété suivante : la somme de deux éléments non inversibles est non inversible. Montrer que, si p est premier et $k \in \mathbb{N}^*$, les anneaux $\mathbb{Z}/p^k\mathbb{Z}$ et $K[X]/<X^k>$ sont locaux.

I.1.29 Soit S une partie multiplicative de l'anneau commutatif A. On note i le morphisme $A \to S^{-1}A$. Démontrer que, pour tout idéal I de A, l'idéal de $S^{-1}A$ engendré par $i(I)$ est égal à $S^{-1}I := \{a/s \mid a \in I \quad \text{et} \quad s \in S\}$.

I.1.30 $*$ Démontrer que, pour tout idéal premier $\mathcal{P}$ de A qui ne rencontre pas S, l'idéal $\mathcal{Q} := S^{-1}\mathcal{P}$ est premier dans $S^{-1}A$. Démontrer que, pour tout idéal premier $\mathcal{Q}$ de $S^{-1}A$, l'idéal premier $\mathcal{P} := i^{-1}(\mathcal{Q})$ de A ne rencontre pas S. Vérifier que l'on obtient ainsi deux applications réciproques l'une de l'autre entre l'ensemble des idéaux premiers de A qui ne rencontrent pas S et l'ensemble des idéaux premiers de $S^{-1}A$.

I.1.31 $*$ Soit $x \in A$ non nilpotent. Montrer qu'il existe un idéal premier ne contenant pas x. En conclure que le radical de A est l'intersection de ses idéaux premiers.

I.1.32 On appelle *racine* d'un idéal I de A, et l'on note $\sqrt{I}$, l'ensemble des $x \in A$ tels qu'il existe $n \in \mathbb{N}$ tel que $x^n \in I$ (le nilradical de A est donc la racine de l'idéal nul).
Déduire de l'exercice I.1.31 que $\sqrt{I}$ est un idéal égal à l'intersection des idéaux premiers contenant I.

I.1.33 Prouver la réciproque de la première assertion de la proposition 36 de la page 22.

I.1.34 (i) Soient A_1 et A_2 deux anneaux commutatifs. Démontrer que le groupe des unités de l'anneau produit $A_1 \times A_2$ est : $(A_1 \times A_2)^* = A_1^* \times A_2^*$.
(ii) À l'aide du théorème chinois, en déduire que, si $m, n \in \mathbb{N}$ sont premiers entre eux, le groupe $(\mathbb{Z}/mn\mathbb{Z})^*$ est isomorphe au groupe $(\mathbb{Z}/m\mathbb{Z})^* \times (\mathbb{Z}/m\mathbb{Z})^*$. Retrouver ainsi l'égalité $\varphi(mn) = \varphi(m)\varphi(n)$ (*multiplicativité* de l'indicatrice d'Euler).

I.1.35 Quels sont les facteurs irréductibles de X^4-1, de X^4+1, de X^4+X^2+1 dans $\mathbb{Z}[X]$? Soit $a \in \mathbb{Z} \setminus \{0\}$. Montrer que $X^4 + aX^2 - 1$ est irréductible dans $\mathbb{Z}[X]$.

I.1.36 Soit $f : A \to B$ un morphisme d'anneaux. Quel est le noyau du morphisme induit de $A[X]$ dans $B[X]$?

I.1.37 $*$ Montrer que le polynôme $1 + aX$ est inversible dans $A[X]$ si, et seulement si, a est nilpotent. Si $1 + aX$ n'est pas inversible, soit $\mathfrak{M}$ un idéal maximal qui le contient (on expliquera pourquoi il en existe). Montrer que $\mathcal{P} := A \cap \mathfrak{M}$ est un idéal premier de A qui ne contient pas a. En déduire une nouvelle démonstration du fait que le radical est l'intersection des idéaux premiers.

I.1.38 Soit $x \in \mathbb{Q}$. Quel est le noyau du morphisme $P \mapsto P(x)$ de $\mathbb{Z}[X]$ dans $\mathbb{Q}$?

Exercices sur la section 3

I.1.39 Traduire la définition 7 de la page 26 en termes de bilinéarité de la multiplication interne.

I.1.40 Soient K un corps, A un anneau non trivial et $\varphi : K \to A$ un morphismes d'anneaux. Démontrer que φ est injectif : on peut donc identifier K à un sous-anneau de A. On suppose que K est central dans A. Montrer que A est muni d'une structure naturelle de K-algèbre.

I.1.41 Décrire les K-algèbres $K[X]/<X^2-1>$, $K[X]/<X^2>$ et $K[X]/<X^2+1>$: bases, règles de calcul, intégrité, idéaux, idéaux premiers, idéaux maximaux, radical.

I.1.42 Démontrer que toute algèbre de dimension finie qui est un anneau intègre est un corps.

I.1.43 $*$ Soit $E = K[X]$. Dans l'algèbre $\mathcal{L}_K(E)$, la dérivation $D : P \mapsto P'$ est-elle un élément algébrique ?

I.1.44 $*$ Soit $A = M_2(\mathbb{C})$. Utiliser le théorème chinois pour décrire la sous-algèbre $\mathbb{C}[x]$ dans les cas suivants : x est une matrice scalaire ; x n'est pas scalaire mais admet une seule valeur propre ; x admet deux valeurs propres distinctes.

I.1.45 $**$ (i) Soit A une K-algèbre de dimension finie. Pour tout $x \in A$, on définit l'application $\lambda_x : y \mapsto xy$ de A dans lui-même. Démontrer que l'application $x \mapsto \lambda_x$ est un morphisme injectif d'algèbres de A dans $\mathcal{L}_K(A)$, et que le polynôme minimal de λ_x est égal à celui de x.
(ii) On définit la *norme* de x par la formule : $N(x) := \det \lambda_x$. Démontrer que $N(xy) = N(x)N(y)$ et que x est inversible si, et seulement si, $N(x)$ est non nul.
(iii) Calculer $N(x)$ dans les cas suivants : $A = K[X]/<X^2 - d>$; $K = \mathbb{Q}$, $A = \mathbb{Q}[\sqrt{d}]$; $K = \mathbb{R}$, $A = \mathbb{C}$ ou $\mathbb{H}$.

I.1.46 $**$ (i) Soient x, y deux éléments algébriques d'une K-algèbre A. On suppose que x et y commutent. Démontrer que $x + y$ et xy sont algébriques.
(ii) Dans la $\mathbb{Q}$ algèbre $\mathbb{H}$, démontrer que $\alpha\mathbf{i} + \beta\mathbf{j} + \gamma\mathbf{k}$ est algébrique si, et seulement si, le réel $\alpha^2 + \beta^2 + \gamma^2$ est un nombre algébrique. En déduire des éléments x, y algébriques tels que $x + y$ n'est pas algébrique.

Exercices sur la section 4

I.1.47 Montrer que $f \in A[[X]]$ est inversible si, et seulement si, son terme constant est inversible dans l'anneau A. Montrer que, si A est intègre, alors $A[[X]]$ est intègre.

I.1.48 Calculer $(1 + X + X^2)^{-1} \in \mathbb{C}[[X]]$ par décomposition en éléments simples.

I.1.49 $*$ Pour tout $k \in \mathbb{N}^*$, montrer que $G_k := \{f \in K[[X]] \mid \omega(f-1) \geqslant k\}$ est un sous-groupe de $K[[X]]^*$ (inversibles de $K[[X]]$). Montrer que $G_k/G_{k+1} \simeq K$.

I.1.50 $*$ Comment utiliser la décomposition en éléments simples de $R \in \mathbb{C}(X)$ pour obtenir son développement en série de Laurent ?

I.1.51 $**$ On reprend les notations de l'exercice 19 de la page 34. En *admettant* l'identité (due à Hadamard) $H_n^{(k)} H_{n+2}^{(k)} - H_n^{(k+1)} H_{n+2}^{(k-1)} = \left(H_{n+1}^{(k)}\right)^2$, démontrer que, si l'on a $H_{m+j}^{(k+1)} = 0$ pour $0 \leqslant j \leqslant r-1$, les r déterminants $H_{m+j}^{(k)}$ pour $1 \leqslant j \leqslant r$ sont tous nuls ou tous non nuls. En déduire que, pour que f soit rationnelle, il faut, et il suffit, qu'il existe $d, q \in \mathbb{N}$ tels que $H_{d+j}^{(q+1)} = 0$ pour tout $j \in \mathbb{N}$.

I.1.52 Soient $K \subset L$ deux corps. Si le développement en série formelle de $R \in L(X)$ est à coefficients dans K, alors $R \in K(X)$.

I.1.53 $*$ Soit $h \in K[[X]]$. L'unique dérivation D de la K-algèbre $K[[X]]$ telle que $D(X) = h$ est $f \mapsto f'h$.

I.1.54 Résoudre l'équation différentielle $X^2 f' - f = -X$.

I.1.55 $**$ Démontrer que $K[[X]]$ est compact si, et seulement si, K est fini.

I.1.56 $*$ Montrer que le produit infini $\prod_{n \geqslant 1} \frac{1}{1 - X^n}$ converge dans $K[[X]]$ et décrire ses coefficients.

I.1.57 $*$ Expliciter les polynômes universels $P_{i,k}$ de la proposition 58 de la page 38.

I.1.58 Justifier rigoureusement la notation $f(0)$.

I.1.59 Définir $\sin$ et $\cos$ comme des séries formelles et prouver que $\sin^2 + \cos^2 = 1$.

I.1.60 Démontrer la formule : $\dfrac{1}{\sqrt{1 - 4X}} = \displaystyle\sum_{n \geqslant 0} \binom{2n}{n} X^n$.

I.1.61 Soit $(v_n)_{n \in \mathbb{N}}$ une suite de complexes non nuls telle que la série de terme général $w_n := v_n - 1$ est absolument convergente. On pose $V_N := \prod_{n=0}^{N} v_n$. Vérifier les inégalités :

$$\begin{aligned} \forall p \in \mathbb{N}\ ,\ |V_{N+p} - V_N| &\leqslant \prod_{n=0}^{N} (1 + |w_n|) \left(\prod_{n=N+1}^{N+p} (1 + |w_n|) - 1 \right) \\ &\leqslant \exp\left(\sum_{n=0}^{N} |w_n| \right) \left(\exp\left(\sum_{n=N+1}^{N+p} |w_n| \right) - 1 \right) \end{aligned}$$

En déduire l'existence du *produit infini* : $\prod_{n \geqslant 0} v_n := \lim_{N \to +\infty} \prod_{n=0}^{N} v_n$.

Vérifier que celui-ci est non nul.

I.1.62 $*$ Pour tout $p \in \mathbb{N}$ et tout $n \geqslant p$, démontrer la formule :

$$\binom{p - 1/2}{n} = (-1)^p \frac{\binom{2p}{p}}{\binom{n}{p}} \left(\frac{-1}{4} \right)^n \binom{2(n-p)}{n-p}.$$

En déduire un équivalent simple de $\binom{p - 1/2}{n}$.

Actions de groupes

I.2

Nous savons ce qu'est un groupe : c'est un ensemble muni d'une loi de composition interne vérifiant certaines propriétés simples. Historiquement, ce n'est pas sous cette forme « abstraite » que les groupes sont apparus, mais sous la forme de « groupes de permutations ». Un groupe de permutations d'un ensemble X est, avec le vocabulaire dont nous disposons, un sous-groupe du groupe symétrique $\mathfrak{S}(X)$. Ces deux points de vue se rejoignent dans la notion d'action de groupe, dont l'étude est l'objet du présent module.

Une action de groupe met en jeu un groupe G, un ensemble X, et une loi externe $G \times X \to X$ vérifiant deux propriétés simples. Les éléments de X, appelés points, sont « transformés » par les éléments de G, et chaque élément de G « agit » sur les points de X. Le vocabulaire que nous utiliserons est d'ailleurs d'origine géométrique. La raison est historique. Au XIX$^{\text{e}}$ siècle, des géométries « non euclidiennes » (*i.e.* ne vérifiant pas le « postulat » d'Euclide sur les parallèles) ont été découvertes. Il est apparu (avec F. Klein) que chaque type de géométrie pouvait être considéré comme l'étude des propriétés invariantes par l'action d'un certain groupe.

Lorsque X n'est pas seulement un ensemble mais qu'il a une structure, qui peut être algébrique, géométrique ou « topologique », on peut exiger que l'action considérée « respecte » cette structure. Si par exemple X est un espace vectoriel, on peut exiger que les éléments de G agissent « linéairement » sur X ; cela conduit aux « représentations linéaires » des groupes, dont l'étude est importante, en physique notamment. Elle est abordée dans [MPA1]. En adaptant ainsi l'action considérée à la nature de X, on obtient un outil très puissant, ce que nous montrerons à travers quelques applications simples.

1 Généralités

1.1 Définitions et exemples

Dans ce qui suit, les groupes seront notés multiplicativement, sauf mention explicite du contraire. L'élément neutre d'un groupe sera noté 1.

Définition 1. Soient G un groupe et X un ensemble. On appelle *action* de G sur X toute application de $G\times X$ dans X, notée $(g,x)\longmapsto g\cdot x$ et vérifiant les deux propriétés suivantes :

1. Pour tous $g,h\in G$ et tout $x\in X$, $g\cdot(h\cdot x)=(gh)\cdot x$.
2. Pour tout $x\in X$, $1\cdot x=x$.

Une action de G sur X est ainsi une *loi externe*, la propriété 1 est appelée « associativité mixte », on notera l'analogie avec la loi externe d'un espace vectoriel. Si $g\in G$ et $x\in X$, nous dirons aussi que $g\cdot x$ est le *transformé* de x par g, pour l'action donnée. Comme pour la loi externe d'un espace vectoriel, nous écrirons souvent gx au lieu de $g\cdot x$, si cela ne prête pas à confusion. Enfin la locution « le groupe G agit sur X » signifiera que l'on s'est donné une action du groupe G sur l'ensemble X.
Si la loi de G était notée additivement, par exemple, les propriétés 1 et 2 devraient être écrites ainsi : $g\cdot(h\cdot x)=(g+h)\cdot x$ et $0\cdot x=x$.
Voici une notation utile. Soit $(g,x)\mapsto g\cdot x$ une action de G sur X. Étant données une partie S de G et une partie Y de X, nous poserons :

$$S\cdot Y\ :=\ \big\{s\cdot y\ \big|\ (s,y)\in S\times Y\big\}. \tag{1}$$

Si S,T sont deux parties de G et si Y est une partie de X, on a $S\cdot(T\cdot Y)=(ST)\cdot Y$, en se rappelant que, par définition, ST est la partie de G formée des produits st, (s,t) décrivant $S\times T$.
Si $S:=\{s\}$ est un singleton, on écrit $s\cdot Y$ au lieu de $\{s\}\cdot Y$. De même, si $Y=\{y\}$ est un singleton, on écrit $S\cdot y$ au lieu de $S\cdot\{y\}$.
Voici une propriété résultant immédiatement de la définition 1 :

PROPRIÉTÉ. *Considérons une action du groupe G sur l'ensemble X. Étant donnés des éléments $x,y\in X$ et $g\in G$, les égalités $y=g\cdot x$ et $x=g^{-1}\cdot y$ sont équivalentes.*

En effet, supposons que $y=g\cdot x$. Alors :

$$g^{-1}\cdot y\ =\ g^{-1}\cdot(g\cdot x)\ =\ (g^{-1}g)\cdot x\ =\ 1\cdot x\ =\ x.$$

Pour la réciproque, on échange les rôles de x et y (resp. de g et g^{-1}).

Proposition 1. Soient G un groupe et X un ensemble. Il revient au même de se donner une action $(g,x)\mapsto g\cdot x$ de G sur X ou un morphisme φ de G dans le groupe $(\mathfrak{S}(X),\circ)$. La traduction se fait au moyen de la formule suivante, valable pour tous $g\in G$ et $x\in X$:

$$g\cdot x\ =\ \varphi(g)(x). \tag{2}$$

Nous dirons que φ est le morphisme *associé* à l'action donnée.

Démonstration. Partons d'une action $(g,x)\mapsto g\cdot x$ de G sur X. Pour tout $g\in G$, notons $\varphi(g)$ l'application $x\mapsto g\cdot x$ de X dans X. Les propriétés 1 et 2 de la définition se traduisent ainsi : d'une part $\varphi(gh)=\varphi(g)\circ\varphi(h)$ pour tous $g,h\in G$, et d'autre part $\varphi(1)=\mathrm{Id}_X$. Si $g\in G$, on en déduit :

$$\varphi(g)\circ\varphi(g^{-1})=\varphi(gg^{-1})=\varphi(1)=\mathrm{Id}_X\,,$$

et de même $\varphi(g^{-1}) \circ \varphi(g) = \mathrm{Id}_X$. Ainsi, $\varphi(g)$ est une permutation de X (bijection de X sur X), dont l'inverse (bijection réciproque) est $\varphi(g^{-1})$. L'égalité $\varphi(gh) = \varphi(g) \circ \varphi(h)$, valable pour tous $g, h \in G$, montre alors que φ est un morphisme de G dans $\mathfrak{S}(X)$.

Soit inversement $\varphi : G \to \mathfrak{S}(X)$ un morphisme. Si $(g, x) \in G \times X$, posons $g \cdot x := \varphi(g)(x)$. L'élément neutre du groupe $(\mathfrak{S}(X), \circ)$ est Id_X. Puisque φ est un morphisme, $\varphi(1) = \mathrm{Id}_X$, *i.e.* $1 \cdot x = x$ pour tout $x \in X$. De même, si $g, h \in G$, l'égalité $\varphi(gh) = \varphi(g) \circ \varphi(h)$ se traduit ainsi : $g \cdot (h \cdot x) = (gh) \cdot x$ pour tout $x \in X$. Les propriétés de la définition 1 de la page précédente sont donc vérifiées. ∎

Il y a ainsi deux langages. Le premier est celui des lois externes, il est plus « géométrique », nous le verrons dans la section suivante, notamment par le vocabulaire employé. Le second langage, plus « algébrique », est celui des morphismes. La considération du noyau et de l'image de φ permet, par exemple, d'obtenir des résultats de nature combinatoire (voir l'exercice 14 de la page 84).

Lorsque φ est injectif, *i.e.* de noyau trivial, on dit que l'action est *fidèle*. Cette terminologie peut s'expliquer en disant que, si l'on remplace un élément g de G par la permutation $\varphi(g)$ de X, « on ne perd pas d'information ». En général, le noyau de φ est aussi appelé *noyau de l'action* ; il est formé des $g \in G$ tels que $g \cdot x = x$ pour tout $x \in X$.

> ***Remarque***. Au lieu d'« action », on dit aussi « opération » de G sur X. Pour être précis, il faudrait parler « d'action à gauche ». En effet, il existe aussi une notion *d'action à droite* de G sur X : la propriété 1 de la définition 1 de la page précédente est remplacée par la suivante : $g \cdot (h \cdot x) = (hg) \cdot x$ pour tous g, h, x. Il est dans ce cas préférable d'écrire $x \cdot g$ au lieu de $g \cdot x$, la propriété ci-dessus s'écrivant alors $(x \cdot g) \cdot h = x \cdot (gh)$ pour tous g, h, x. L'inconvénient de ces actions à droite est que la proposition 1 de la page ci-contre n'est plus vraie : pour tous $g, h \in G$, on a $\varphi(g) \circ \varphi(h) = \varphi(hg)$, de sorte que φ n'est pas un morphisme, en général. Bien entendu, si G est *abélien* (commutatif), les notions d'action à gauche et d'action à droite de G sur un ensemble coïncident.
>
> Dans cet ouvrage, nous parlerons le plus souvent d'actions à gauche, en supprimant la mention « à gauche ».

Exemples.

1) Soient G un groupe, X un ensemble. En posant $g \cdot x = x$ pour tout $(g, x) \in G \times X$, on obtient une action, dite *triviale*, de G sur X. Le morphisme associé est le morphisme *trivial* de G dans $\mathfrak{S}(X)$.

2) Soit E un espace vectoriel sur un corps K. L'application $(\lambda, x) \mapsto \lambda \cdot x$ est une action du groupe $(K^*, \times)$ sur E.

3) Soit X un ensemble. Pour tous $s \in \mathfrak{S}(X)$ et $x \in X$, posons $s \cdot x := s(x)$. Nous obtenons une action, dite *naturelle*, du groupe $(\mathfrak{S}(X), \circ)$ sur X. Cette action est *fidèle* : le morphisme associé est l'application identité de $\mathfrak{S}(X)$.

4) Partons d'une action $G \times X \to X$. Si H est un sous-groupe de G, la restriction de cette action à $H \times X$ est une action de H sur X. Le morphisme $H \to \mathfrak{S}(X)$ associé est la restriction à H du morphisme $G \to \mathfrak{S}(X)$ donné.

5) Soit E un espace vectoriel. Le groupe linéaire $\mathcal{GL}(E)$ est un sous-groupe de $\mathfrak{S}(E)$. Il résulte donc des deux exemples précédents que $(g,x) \mapsto g(x)$ est une action, dite encore *naturelle*, de $\mathcal{GL}(E)$ sur E.

6) Partons d'une action $G \times X \to X$. À cette action est associée, de façon naturelle, une action de G sur $\mathcal{P}(X)$, ensemble des parties de X. Il s'agit de l'application $(g,Y) \mapsto g\cdot Y$ de $G \times \mathcal{P}(X)$ dans $\mathcal{P}(X)$, *cf.* la formule (1). Le lecteur peut vérifier que l'on obtient bien une action de G sur $\mathcal{P}(X)$.

7) Soit G un groupe agissant sur plusieurs ensembles $X_1, \dots, X_n$. Alors G agit de façon « diagonale » sur le produit $P := X_1 \times X_2 \times \cdots \times X_n$. Pour tous $g \in G$ et $(x_1, \dots, x_n) \in P$, il suffit de poser :

$$g \cdot (x_1, \dots, x_n) := (g \cdot x_1, \dots, g \cdot x_n). \tag{3}$$

Ainsi, à toute action de G sur X et tout $n \in \mathbb{N}^*$ correspond une action « diagonale » de G sur X^n.

8) Voici un exemple théorique important. Soit G un groupe (multiplicatif). La multiplication $G \times G \to G$ est une action du groupe G sur l'*ensemble* G. En effet, posons ici $g \cdot x := gx$ pour tous $g, x \in G$. Dans la définition 1 de la page 58, la propriété 1 est vérifiée (c'est l'associativité de la loi de G), la propriété 2 aussi (elle traduit le fait que 1 est élément neutre de G). Dans cet exemple, G joue un double rôle : celui du groupe qui agit et celui de l'ensemble sur lequel ce groupe agit.
Cette action est appelée *action par translations à gauche* du groupe G sur lui-même. Cette terminologie est justifiée : pour tout $g \in G$ fixé, la permutation $\varphi(g)$ de G associée à g est la *translation à gauche* $x \mapsto gx$ de G (*cf.* [L1]).
Cette action est *fidèle* : pour tout $g \in G$ distinct de 1, on a $\varphi(g)(1) = g \cdot 1 = g1 = g \neq 1$, en particulier $\varphi(g) \neq \mathrm{Id}_G$. En fait, cette action a une propriété plus forte que la fidélité : pour tous $g \in G$ et $x \in G$, l'égalité $g \cdot x = x$, c'est-à-dire $gx = x$, est vraie si, et seulement si, $g = 1$ (régularité dans un groupe). En général, une action d'un groupe G sur un ensemble X est dite *simple* si elle possède cette propriété :

$$\forall g \in G,\ \forall x \in X,\ (g \cdot x = x) \Longrightarrow (g = 1).$$

On dit alors aussi que G agit *simplement* sur X.

9) Soit G un groupe multiplicatif. Pour tous $g, x \in G$, posons $g \cdot x := gxg^{-1}$. Nous obtenons encore une action du groupe G sur l'ensemble G. En effet, si $g, h, x \in G$, il vient :

$$g \cdot (h \cdot x) = g \cdot (hxh^{-1}) = g(hxh^{-1})g^{-1} = (gh)x(gh)^{-1} = (gh) \cdot x,$$

d'où la propriété 1 de la définition. La propriété 2 est clairement vérifiée.
L'action obtenue est appelée *conjugaison*, on dit que G agit sur lui-même *par conjugaison*. Notons $\varphi : G \to \mathfrak{S}(G)$ le morphisme associé à cette action. Soit $g \in G$. Par définition, on a donc :

$$\varphi(g)(x) := gxg^{-1} \quad \text{pour tout} \quad x \in G. \tag{4}$$

Nous savons que $\varphi(g)$ n'est pas seulement une permutation de l'ensemble G, c'est un *automorphisme* du groupe G. Ainsi φ est à valeurs dans le groupe $\mathrm{Aut}(G)$ des automorphismes de G, *i.e.* φ est un morphisme de G dans $\mathrm{Aut}(G)$.

Le noyau de φ est le *centre* $Z(G)$ de G, formé des $g \in G$ commutant avec tout élément x de G (*cf.* [L1]). En particulier, l'action par conjugaison de G sur lui-même est triviale si, et seulement si, G est commutatif.

10) Soit G un groupe agissant sur deux ensembles X et Y. On définit une action de G sur l'ensemble $\mathcal{F}(X,Y)$ des applications de X dans Y par la formule suivante :

$$(g \cdot f)(x) \;:=\; g \cdot f\big(g^{-1} \cdot x\big), \quad \text{où} \quad f \in \mathcal{F}(X,Y),\ g \in G,\ x \in X.$$

Si $\varphi : G \to \mathfrak{S}(X)$ et $\psi : G \to \mathfrak{S}(Y)$ sont les deux morphismes associés aux actions données, la formule ci-dessus s'écrit : $g \cdot f = \psi(g) \circ f \circ \varphi(g)^{-1}$ pour tous f, g. Alors $1 \cdot f = f$ pour toute f. Si $g, h \in G$ et $f \in \mathcal{F}(X,Y)$, $g \cdot (h \cdot f)$ vaut par définition $\psi(g) \circ \big[\psi(h) \circ f \circ \varphi(h)^{-1}\big] \circ \varphi(g)^{-1}$, c'est-à-dire, puisque φ et ψ sont des morphismes, $\psi(gh) \circ f \circ \varphi(gh)^{-1}$, soit $(gh) \cdot f$. Nous avons donc obtenu une action de G sur $\mathcal{F}(X,Y)$.

Exercice 1.

Soit X un ensemble. Montrer qu'il revient au même de se donner une action du groupe $(\mathbb{Z}, +)$ sur X ou une permutation de X.

Solution. Considérons une action $(n, x) \mapsto n \cdot x$ de $(\mathbb{Z}, +)$ sur X, et soit $\varphi : \mathbb{Z} \to \mathfrak{S}(X)$ le morphisme associé. Alors $s := \varphi(1)$ est une permutation de X. Puisque φ est un morphisme du groupe $(\mathbb{Z}, +)$ dans le groupe $(\mathfrak{S}(X), \circ)$, on a $\varphi(n) = s^n$ pour tout $n \in \mathbb{Z}$. Autrement dit :

$$n \cdot x \;=\; s^n(x) \quad \text{pour tout couple} \quad (n, x) \in \mathbb{Z} \times X. \tag{5}$$

Soit inversement s une permutation de X. Pour tout couple $(n, x) \in \mathbb{Z} \times X$, *définissons* $n \cdot x \in X$ par la formule précédente. Nous obtenons une action de $(\mathbb{Z}, +)$ sur X. En effet, $s^0 = \mathrm{Id}_X$, donc $0 \cdot x = x$ pour tout $x \in X$ $\big($attention, l'élément neutre du groupe $(\mathbb{Z}, +)$ est 0, et non $1\big)$. Soient ensuite $m, n \in \mathbb{Z}$. L'égalité $s^m \circ s^n = s^{m+n}$ donne, pour tout $x \in X$:

$$m \cdot (n \cdot x) \;=\; s^m\big(s^n(x)\big) \;=\; (s^m \circ s^n)(x) \;=\; s^{m+n}(x) \;=\; (m+n) \cdot x,$$

d'où la propriété 1 de la définition 1 de la page 58. Enfin, notant $\varphi : \mathbb{Z} \to \mathfrak{S}(X)$ le morphisme associé à l'action que nous venons de définir, il vient :

$$\varphi(1)(x) \;=\; 1 \cdot x \;=\; s^1(x) \;=\; s(x) \quad \text{pour tout} \quad x \in X,$$

d'où $\varphi(1) = s$. La conclusion en résulte.

Voici un résultat général sur les groupes finis :

Théorème 2 (Cayley). Tout groupe fini d'ordre n est isomorphe à un sous-groupe du groupe symétrique $\mathfrak{S}_n$.

Démonstration. Soit donc G un groupe fini d'ordre n. Nous avons vu que le morphisme $\varphi : G \to \mathfrak{S}(G)$ associé à l'action de G sur lui-même par translations à gauche est injectif. Par ailleurs, puisque $\mathrm{card}\,(G) = n$, il existe un isomorphisme θ de $\mathfrak{S}(G)$ sur $\mathfrak{S}_n$ (*cf.* exercice II.2.12 de [L1]). Alors $\theta \circ \varphi : G \to \mathfrak{S}_n$ est un morphisme injectif, il induit donc un isomorphisme de G sur $\mathrm{Im}(\theta \circ \varphi)$, qui est un sous-groupe de $\mathfrak{S}_n$. ∎

À titre d'illustration, prenons $n := 4$. D'après l'exercice II.2.14 de [L1], il existe, à isomorphisme près, deux groupes d'ordre 4, à savoir les groupes additifs $\mathbb{Z}/4\mathbb{Z}$ et $(\mathbb{Z}/2\mathbb{Z})^2$. Le premier est isomorphe au sous-groupe $\big\langle(1\,2\,3\,4)\big\rangle$ de $\mathfrak{S}_4$ engendré par le 4-cycle $(1\,2\,3\,4)$, le second est isomorphe au sous-groupe V_4 de $\mathfrak{S}_4$ défini dans l'exercice précité.

> ***Remarque***. Soit n un entier strictement positif. D'après le théorème de Cayley, pour « classifier » les groupes d'ordres n, *i.e.* pour déterminer à isomorphisme près tous les groupes d'ordre n, il « suffirait » de se limiter aux sous-groupes de $\mathfrak{S}_n$. En réalité, on ne peut procéder ainsi que pour des petites valeurs de n (comme $n := 4$ ci-dessus). *Classifier tous* les groupes finis est en fait une tâche hors de portée des mathématiques actuelles. Notons cependant que la classification de tous les groupes *abéliens* finis est assez simple, elle est faite dans [MPA1].
>
> La portée du théorème de Cayley est plutôt historique. Ce théorème montre que la notion « abstraite » de groupe (en tous cas de groupe fini) introduite dans [L1] est essentiellement la même que celle de *groupe de permutations*, dont nous avons parlé dans l'introduction.

1.2 Espaces affines et actions de groupes

Revenons sur la notion d'espace affine, introduite dans [L1].

Exercice 2.

Soit K un corps. La notion d'*espace affine* sur K a été définie dans [L1] (module III.1). Soient donc X un espace affine sur K et E sa direction (c'est un K-espace vectoriel). À tout couple $(\vec{u}, A) \in E \times X$ associons le point $A + \vec{u}$ de X (« translaté » de A par $\vec{u}$). Montrer que l'on obtient ainsi une action du groupe additif $(E, +)$ sur X, vérifiant la propriété $(*)$ suivante : pour tous points $A, B \in X$, il existe un unique vecteur $\vec{u} \in E$ tel que $B = A + \vec{u}$. Étudier la réciproque.

Solution. Pour tout couple $(\vec{u}, A) \in E \times X$, posons ici $\vec{u} \cdot A := A + \vec{u}$ (cette notation n'est que provisoire). Rappelons que $A + \vec{u}$ est l'unique point $M \in X$ tel que $\overrightarrow{AM} = \vec{u}$. Montrons que $(\vec{u}, A) \mapsto A + \vec{u}$ est une action de $(E, +)$ sur X. Pour cela, appliquons librement le module III.1 de [L1].

Soit d'abord $A \in X$. L'égalité $\overrightarrow{AA} = \vec{0}$ se traduit par $A + \vec{0} = A$, soit $\vec{0} \cdot A = A$. La propriété 1 de la définition 1 de la page 58 est donc vérifiée.

Soient ensuite $A \in X$ et $\vec{u}, \vec{v} \in E$. Posons $B := A + \vec{u}$ et $C := B + \vec{v}$. Ainsi $\overrightarrow{AB} = \vec{u}$ et $\overrightarrow{BC} = \vec{v}$. D'après la relation de Chasles, $\overrightarrow{AC} = \vec{u} + \vec{v}$, c'est-à-dire $A + (\vec{u} + \vec{v}) = C$. En d'autres termes, puisque $B = \vec{u} \cdot A$ et $C = \vec{v} \cdot B$, nous avons prouvé l'égalité $\vec{v} \cdot (\vec{u} \cdot A) = (\vec{v} + \vec{u}) \cdot A$ (vu la commutativité de l'addition de E). La propriété 2 de la définition 1 de la page 58 est donc vérifiée.

Soient $A, B \in X$. Pour tout $\vec{u} \in E$, les conditions $\overrightarrow{AB} = \vec{u}$, $B = A + \vec{u}$ et $B = \vec{u} \cdot A$ sont équivalentes. Il y a donc un unique $\vec{u} \in E$ tel que $B = \vec{u} \cdot A$, à savoir $\vec{u} = \overrightarrow{AB}$. La propriété $(*)$ est donc vérifiée.

Réciproquement, soient E un espace vectoriel sur K, X un ensemble non vide et $(\vec{u}, A) \mapsto \vec{u} \cdot A$ une action de $(E, +)$ sur X satisfaisant la propriété $(*)$. Pour tous $A, B \in X$, il existe donc un unique vecteur $\vec{u} \in E$ tel que $B = \vec{u} \cdot A$, notons ce vecteur $\overrightarrow{AB}$. Il s'agit de voir que l'application $(A, B) \mapsto \overrightarrow{AB}$ de $X \times X$ dans E fait de X un espace affine de direction E.
Vérifions d'abord la relation de Chasles. Soient $A, B, C \in X$. Posons $\vec{u} = \overrightarrow{AB}$ et $\vec{v} = \overrightarrow{BC}$, de sorte que $B = \vec{u} \cdot A$ et $C = \vec{v} \cdot B$. Puisque nous sommes partis d'une action du groupe commutatif $(E, +)$ sur X, $(\vec{u}+\vec{v}) \cdot A = \vec{v} \cdot (\vec{u} \cdot A)$, soit $(\vec{u}+\vec{v}) \cdot A = C$. Alors, par définition, $\overrightarrow{AC} = \vec{u} + \vec{v}$, *i.e.* $\overrightarrow{AC} = \overrightarrow{AB} + \overrightarrow{BC}$.
Soient ensuite $A \in X$ et $\vec{u} \in E$. Nous devons vérifier qu'il existe un unique point $B \in X$ tel que $\overrightarrow{AB} = \vec{u}$. C'est évident : le seul point B ayant la propriété requise est $B := \vec{u} \cdot A$.
En conclusion, étant donnés un espace vectoriel E et un ensemble non vide X, se donner sur X une structure d'espace affine de direction E revient à se donner une action de $(E, +)$ sur X satisfaisant la condition $(*)$.

À propos d'espaces affines, donnons un autre exemple d'action de groupe. Soient X un espace affine de direction E sur un corps K. Si $f : X \to X$ est une application affine, sa partie linéaire sera notée $\vec{f}$, c'est le seul endomorphisme de E tel que $\vec{f}(\overrightarrow{AB}) = \overrightarrow{f(A)f(B)}$ pour tous $A, B \in X$. Si $f, g : X \to X$ sont deux applications affines, il est facile de vérifier que $g \circ f$ est aussi affine, et que sa partie linéaire est $\vec{g} \circ \vec{f}$. Ainsi $\overrightarrow{g \circ f} = \vec{g} \circ \vec{f}$. Voici un résultat déjà connu si l'on remplace « affine » par « linéaire » :

Proposition 3. Soit X un espace affine. Une application affine $f : X \to X$ est bijective si, et seulement si, sa partie linéaire $\vec{f}$ l'est, auquel cas f^{-1} est une application affine, de partie linéaire $\vec{f}^{-1}$. De plus, les applications affines bijectives de X sur lui-même forment un sous-groupe de $\big(\mathfrak{S}(X), \circ\big)$.

Démonstration. Soit $f : X \to X$ une application affine. Considérons un point $O \in X$. Soient ensuite $P \in X$ et $\vec{v} := \overrightarrow{f(O)P}$.
Pour tout point $M \in X$, les égalités $f(M) = P$, $\overrightarrow{f(O)f(M)} = \vec{v}$ et $\vec{f}\big(\overrightarrow{OM}\big) = \vec{v}$ sont équivalentes. Puisque les applications $M \mapsto \overrightarrow{OM}$ et $P \mapsto \overrightarrow{f(O)P}$ de X dans E sont bijectives, il en résulte que M est l'unique antécédent de P par f si, et seulement si, $\overrightarrow{OM}$ est l'unique antécédent de $\vec{v}$ par $\vec{f}$. On en déduit que f est bijective si, et seulement si, $\vec{f}$ l'est.
Supposons f bijective. Alors $\vec{f}^{-1}$ est linéaire. Soient $\Omega, P \in X$, posons $O := f^{-1}(\Omega)$ et $M := f^{-1}(P)$. Les égalités $\overrightarrow{\Omega P} = \overrightarrow{f(O)f(M)} = \vec{f}\big(\overrightarrow{OM}\big)$ montrent que $\overrightarrow{OM} = \vec{f}^{-1}\big(\overrightarrow{\Omega P}\big)$. Ainsi $\overrightarrow{f^{-1}(\Omega)f^{-1}(P)} = \vec{f}^{-1}\big(\overrightarrow{\Omega P}\big)$, ce qui montre que f^{-1} est affine et que sa partie linéaire est $\vec{f}^{-1}$.
La dernière assertion de l'énoncé est évidente, elle est laissée au lecteur. ∎

Définition 2. On appelle *transformation affine* d'un espace affine X toute bijection *affine* de X sur lui-même. Ces transformations affines forment un sous-groupe de $\big(\mathfrak{S}(X),\circ\big)$, appelé *groupe affine* de X et noté $\mathcal{GA}(X)$.

Comme exemples de transformations affines, nous connaissons déjà les translations et les homothéties. Soit E la direction de X. L'action de $(E,+)$ sur X (*cf.* l'exercice 2 de la page 62) définit un morphisme de $(E,+)$ dans $\mathfrak{S}(X)$, associant à tout $\vec{u}\in E$ la *translation* $t_{\vec{u}}$ de vecteur $\vec{u}$: $A\mapsto a+\vec{u}$. C'est une transformation affine de X, de partie linéaire Id_E. En effet, si $A,B\in X$, et si $M:=A+\vec{u}=t_{\vec{u}}(A)$ et $N:=B+\vec{u}=t_{\vec{u}}(B)$, on a $\overrightarrow{AM}=\vec{u}=\overrightarrow{BN}$, d'où $\overrightarrow{MN}=\overrightarrow{AB}$ en vertu de la relation de Chasles.

Soient $O\in X$, $\lambda\in K^*$ et $h_{O,\lambda}$ l'*homothétie* (affine) de centre O et de rapport λ, associant à tout point $M\in X$ le point M' défini par $\overrightarrow{OM'}:=\lambda\overrightarrow{OM}$. C'est une transformation affine de X, de partie linéaire $\lambda\,\mathrm{Id}_E$ car, si l'on pose $M:=h_{O,\lambda}(A)$ et $N:=h_{O,\lambda}(B)$, on a :

$$\overrightarrow{OM}=\lambda\overrightarrow{OA}\quad\text{et}\quad\overrightarrow{ON}=\lambda\overrightarrow{OB}\quad\text{d'où}\quad\overrightarrow{MN}=\overrightarrow{MO}+\overrightarrow{ON}=-\lambda\overrightarrow{OA}+\lambda\overrightarrow{OB}=\lambda\overrightarrow{AB}.$$

Proposition 4. L'application $\theta : f\mapsto\vec{f}$ est un morphisme surjectif du groupe affine $\mathcal{GA}(X)$ dans le groupe linéaire $\mathcal{GL}(E)$.
Le noyau de θ est le groupe $\mathcal{T}(X):=\{t_{\vec{u}}\mid\vec{u}\in E\}$ des translations de X, c'est un sous-groupe distingué de $\mathcal{GA}(X)$. Pour tous $\vec{u}\in E$ et $f\in\mathcal{GA}(X)$, on a :

$$f\circ t_{\vec{u}}\circ f^{-1}\;=\;t_{\vec{f}(\vec{u})}. \tag{6}$$

Démonstration. Nous savons que $\overrightarrow{g\circ f}=\vec{g}\circ\vec{f}$ pour tous $f,g\in\mathcal{GA}(X)$, donc l'application θ est un morphisme de $\mathcal{GA}(X)$ dans $\mathcal{GL}(E)$. L'inclusion $\mathcal{T}(X)\subset\operatorname{Ker}\theta$ a été prouvée ci-dessus. Soit inversement $f\in\operatorname{Ker}\theta$. Soit $A\in X$, et posons $\vec{u}:=\overrightarrow{Af(A)}$. Si $M\in X$, on a $\overrightarrow{f(A)f(M)}=\overrightarrow{AM}$, d'où par Chasles $\overrightarrow{Mf(M)}=\overrightarrow{Af(A)}=\vec{u}$, soit $f(M)=M+\vec{u}=t_{\vec{u}}(M)$. Ainsi $f=t_{\vec{u}}$, d'où $\operatorname{Ker}\theta=\mathcal{T}(X)$.
Soient $f\in\mathcal{GA}(X)$ et $\vec{u}\in E$. Soient $A\in X$ et $M:=A+\vec{u}=t_{\vec{u}}(A)$.
Alors $\overrightarrow{AM}=\vec{u}$, d'où $\vec{f}(\vec{u})=\overrightarrow{f(A)f(M)}$, soit $f(M)=f(A)+\vec{f}(\vec{u})$. En d'autres termes, $\big(f\circ t_{\vec{u}}\big)(A)=\big(t_{\vec{f}(\vec{u})}\circ f\big)(A)$, ce qui prouve l'égalité (6).
Prouvons la surjectivité de θ. Soient $\varphi\in\mathcal{GL}(E)$ et $A\in X$. Définissons $f:X\to X$ par la formule $f(M):=A+\varphi(\overrightarrow{AM})$. Si $M,N\in X$, il vient :

$$\overrightarrow{f(M)f(N)}=\overrightarrow{Af(N)}-\overrightarrow{Af(M)}=\varphi(\overrightarrow{AN})-\varphi(\overrightarrow{AM})=\varphi(\overrightarrow{AN}-\overrightarrow{AM})=\varphi(\overrightarrow{MN}).$$

Ainsi f est affine, de partie linéaire φ, donc $f\in\mathcal{GA}(X)$ et $\theta(f)=\varphi$. ∎

Puisque $\mathcal{GA}(X)$ est un sous-groupe de $\mathfrak{S}(X)$, nous disposons d'une action naturelle de $\mathcal{GA}(X)$ sur X $\big($donc sur $\mathcal{P}(X)$, etc$\big)$, à savoir $(g,x)\mapsto g(x)$. Cette action nous servira d'illustration dans les sections suivantes.

2 Orbites, stabilisateurs

Dans ce qui suit, nous considérons une action d'un groupe G, noté en général multiplicativement, sur un ensemble X. Les trois notions centrales que nous allons définir et illustrer sont celles « d'orbite d'un point », de « point fixe » et de « stabilisateur » (d'un point ou d'une partie). Commençons par la notion d'orbite.

2.1 Orbites

Proposition 5. Soit $\mathcal{R}$ la relation binaire définie ainsi sur X : $x\mathcal{R}y$ s'il existe un élément g de G tel que $g \cdot x = y$. Alors $\mathcal{R}$ est une *relation d'équivalence*.

Démonstration. La réflexivité vient de la formule $\forall x \in X,\ 1 \cdot x = x$. La symétrie est une conséquence immédiate de la propriété de la page 58. Pour la transitivité, supposons que l'on ait $x\mathcal{R}y$ et $y\mathcal{R}z$. Il existe donc $g, h \in G$ tels que $y = g \cdot x$ et $z = h \cdot y$. Alors $z = h \cdot (g \cdot x) = (hg) \cdot x$, d'où $x\mathcal{R}z$. ∎

Définition 3. Les classes d'équivalence pour la relation $\mathcal{R}$ ci-dessus sont appelées *orbites* de X sous l'action de G. Plus précisément, si $a \in X$, la classe de a est appelée *orbite de* a sous l'action de G (ou orbite de a sous G).

L'ensemble des orbites sous l'action de G est donc le quotient de X par la relation $\mathcal{R}$, on le note X/G (ou parfois $G \setminus X$).

Proposition 6. Pour tout $a \in X$, l'orbite de a sous l'action de G est :

$$G \cdot a \ := \ \big\{ g \cdot a \ \big|\ g \in G \big\}.$$

Démonstration. C'est une conséquence triviale de la définition de $\mathcal{R}$. ∎

Nous savons que, pour toute relation d'équivalence sur un ensemble X, les classes d'équivalence forment une *partition* de X, *i.e.* ce sont des parties non vides deux à deux disjointes et dont la réunion est X. Ainsi :

Proposition 7. Les orbites de X sous l'action de G forment une *partition* de X.

Rappelons (*cf.* le module I.1 de [L1]) qu'un *ensemble de représentants* pour la relation d'équivalence $\mathcal{R}$ est, par définition, une partie T de X contenant un et un seul élément de chaque classe d'équivalence. Le résultat ci-après est alors une conséquence immédiate de la proposition 6 et de la proposition 7.

Proposition 8. Soient G un groupe agissant sur un ensemble *fini* X et T un ensemble de représentants des orbites de X sous l'action de G. Alors :

$$\operatorname{card}(X) \ = \ \sum_{a \in T} \operatorname{card}(G \cdot a).$$

Cette égalité est parfois appelée *équation aux classes*. Le second membre est aussi noté $\displaystyle\sum_{a \bmod G} \operatorname{card}(G \cdot a)$.

Exercice 3.
Soit E un espace vectoriel de dimension finie $n \geqslant 1$ sur un corps K. Quelles sont les orbites de E sous l'action naturelle de $\mathcal{GL}(E)$?
Solution. L'une de ces orbites est évidemment $\{0\}$. Prouvons que $E_* = E \setminus \{0\}$ est aussi une orbite, ce qui montrera qu'il y a deux orbites. Soient $x, x' \in E_*$. Nous devons montrer qu'il existe $g \in \mathcal{GL}(E)$ tel que $x' = g(x)$. D'après le théorème de la base incomplète, et puisque $x \neq 0$, il existe une base $(e_1, \ldots, e_n)$ de E telle que $e_1 = x$. De même, il existe une base $(e'_1, \ldots, e'_n)$ de E telle que $e'_1 = x'$. Puisque $(e_1, \ldots, e_n)$ et $(e'_1, \ldots, e'_n)$ sont deux bases de E, nous savons qu'il existe un (unique) automorphisme $g \in \mathcal{GL}(E)$ tel que $g(e_i) = e'_i$ pour $i = 1, \ldots, n$. Alors $x' = e'_1 = g(e_1) = g(x)$.

Le cas de l'action par conjugaison d'un groupe sur lui-même (exemple 9 de la page 60) mérite une attention et une terminologie particulières.

Définition 4. Considérons l'action d'un groupe G sur lui-même par conjugaison.

1. La relation d'équivalence $\mathcal{R}$ associée est appelée *conjugaison*, les orbites correspondantes sont appelées *classes de conjugaison*.
2. Deux éléments x, y de G sont dits *conjugués* s'ils ont même classe de conjugaison, *i.e.* s'il existe $g \in G$ tel que $y = gxg^{-1}$.

La classe de conjugaison d'un élément a de G est donc formée des conjugués gag^{-1} de a, g décrivant G. Noter que cette classe est *ponctuelle*, *i.e.* réduite à $\{a\}$, si, et seulement si, $gag^{-1} = a$, soit $ga = ag$, ceci pour tout $g \in G$, autrement dit si, et seulement si, a appartient au centre de G.

Exercice 4.
Montrer que dans un groupe, deux éléments conjugués ont même ordre (fini ou non).
Solution. Soient a, b deux éléments conjugués du groupe G. Il existe donc $g \in G$ tel que $b = gag^{-1}$. L'automorphisme $\alpha : x \mapsto gxg^{-1}$ de G applique a sur b. On en déduit que $\alpha\big(\langle a \rangle\big) = \langle b \rangle$, donc α induit un isomorphisme (en particulier une bijection) de $\langle a \rangle$ sur $\langle b \rangle$. Par définition, l'ordre de a (par exemple) est le cardinal, fini ou non, de $\langle a \rangle$, d'où la conclusion.

Remarque. La description des orbites d'une action donnée correspond souvent à un problème de *classification* (voir la section 3.1). Dans le cas d'un groupe agissant par conjugaison sur lui-même, la description des classes de conjugaison n'est pas toujours évidente. Nous étudierons deux cas significatifs à cet égard : le cas du groupe symétrique $\mathfrak{S}(X)$ d'un ensemble fini (section 3.3), et le cas des groupes $GL_n(\mathbb{C})$ (dans le module I.5).

Le terme « orbite » est d'origine cinématique (ou géométrique). Les deux exercices ci-après illustrent ce fait.

Exercice 5.

Soit G le groupe des rotations vectorielles de $\mathbb{R}^2$. Quelles sont les orbites de $\mathbb{R}^2$ sous l'action naturelle de G ?

Solution. Pour tout $\theta \in \mathbb{R}$, soit R_θ la rotation de centre O et d'angle θ, de sorte que $G := \{R_\theta \mid \theta \in \mathbb{R}\}$. Nous savons que $R_\theta \circ R_\eta = R_{\theta+\eta}$ pour tous $\theta, \eta \in \mathbb{R}$. Identifions chaque R_θ avec sa matrice dans la base canonique de $\mathbb{R}^2$: $R_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

Tout d'abord, l'orbite de O sous l'action de G est évidemment le singleton $\{O\}$.

Soit A un point de $\mathbb{R}^2$ distinct de O, posons $r := OA$. Montrons que l'orbite de A sous l'action de G est le cercle $C(O,r)$ de centre O et de rayon r. Soit $\Omega := (r,0)$. Lorsque θ décrit $\mathbb{R}$, le point $M_\theta := (r\cos\theta, r\sin\theta)$ décrit $C(O,r)$. Mais $M_\theta = R_\theta(\Omega)$, donc M_θ décrit $G\cdot\Omega$, orbite de Ω sous l'action de G. Ainsi $G\cdot\Omega = C(O,r)$. Comme $A \in C(O,r)$, l'orbite de A est la même que celle de Ω, d'où notre assertion.

En conclusion, les orbites de $\mathbb{R}^2$ sous l'action de G sont d'une part le singleton $\{O\}$ et d'autre part les cercles de centre O.

Exercice 6.

Soit $\mathbf{A} \in M_n(\mathbb{R})$ ($n \geqslant 1$). Pour tous $t \in \mathbb{R}$ et $\mathbf{y} \in \mathbb{R}^n$, posons $t\cdot\mathbf{y} := e^{t\mathbf{A}}\cdot\mathbf{y}$. Montrer que l'on définit ainsi une action de $(\mathbb{R},+)$ sur $\mathbb{R}^n$. Interpréter les orbites de cette action à l'aide du système différentiel linéaire $\mathbf{y}' = \mathbf{A}\cdot\mathbf{y}$. Comment retrouver l'exercice précédent ?

Solution. Le lecteur se reportera au module II.2 pour ce qui concerne l'exponentielle d'une matrice carrée réelle ou complexe. D'abord $e^{\mathbf{0}} = I_n$, donc $0\cdot\mathbf{y} = \mathbf{y}$ pour tout $\mathbf{y} \in \mathbb{R}^n$. Soient ensuite $s,t \in \mathbb{R}$. Les matrices $s\mathbf{A}$ et $t\mathbf{A}$ commutent, donc $e^{s\mathbf{A}}e^{t\mathbf{A}} = e^{(s+t)\mathbf{A}}$. L'associativité mixte en résulte, de sorte qu'il s'agit bien d'une action de $(\mathbb{R},+)$ sur $\mathbb{R}^n$.

Soit $\mathbf{y_0} \in \mathbb{R}^n$. L'orbite de $\mathbf{y_0}$ sous l'action de $\mathbb{R}$ est $\{e^{t\mathbf{A}}\cdot\mathbf{y}_0 \mid t \in \mathbb{R}\}$. C'est donc la trajectoire correspondant à l'unique solution maximale du système différentiel $\mathbf{y}' = \mathbf{A}\cdot\mathbf{y}$ prenant à l'instant $t := 0$ la valeur $\mathbf{y_0}$. En observant les figures IV.6.1, IV.6.2 ou IV.6.3 de [L1], le lecteur pourra visualiser, dans le cas $n := 2$, le fait que les orbites forment une partition de $\mathbb{R}^2$.

Prenons $n := 2$ et $\mathbf{A} := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Le système différentiel correspondant s'écrit : $x' = -y,\ y' = x$. Soit $\mathbf{y_0} = (a,b)$ un vecteur arbitraire de $\mathbb{R}^2$. La solution du système prenant à l'instant 0 la valeur $\mathbf{y_0}$ est $t \longmapsto (a\cos t - b\sin t, a\sin t + b\cos t)$. Si $\mathbf{y_0} = \mathbf{0}$, la trajectoire correspondante est $\{O\}$. Si $\mathbf{y_0} \neq \mathbf{0}$ et si $r := ||\mathbf{y_0}|| = \sqrt{a^2+b^2}$, la trajectoire correspondante est $C(O,r)$, avec les notations de l'exercice précédent.

Notons que, pour tout $t \in \mathbb{R}$, $(a\cos t - b\sin t, a\sin t + b\cos t) = R_t(\mathbf{y_0})$. D'après le module IV.6 de [L1], on en déduit l'égalité :

$$e^{t\mathbf{A}} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} = R_t \quad \text{pour tout} \quad t \in \mathbb{R}\,, \qquad (*)$$

ce que l'on peut d'ailleurs vérifier par un calcul direct.

Voyons le lien avec l'exercice précédent. D'abord $f : t \mapsto R_t$ est un morphisme *surjectif* de $(\mathbb{R}, +)$ sur G, de noyau $2\pi\mathbb{Z}$. Ensuite, le morphisme $j : G \to \mathfrak{S}(\mathbb{R}^2)$ associé à l'action naturelle de G sur $\mathbb{R}^2$ est le morphisme d'inclusion. Si $t \in \mathbb{R}$ et $\mathbf{y} \in \mathbb{R}^2$, l'égalité $(*)$ montre que $t \cdot \mathbf{y} = R_t(\mathbf{y})$. Il en résulte que le morphisme $\varphi : \mathbb{R} \to \mathfrak{S}(\mathbb{R}^2)$ associé à l'action de $(\mathbb{R}, +)$ sur $\mathbb{R}^2$ définie ci-dessus est le composé $j \circ f$. On en déduit, vu la surjectivité de f, que les orbites de $\mathbb{R}^2$ sous l'action de $(\mathbb{R}, +)$ sont les mêmes que les orbites de $\mathbb{R}^2$ sous l'action de G. Ces dernières orbites sont donc $\{O\}$ et les cercles de centre O. Nous retrouvons ainsi le résultat de l'exercice précédent.

Voici une définition liée à la notion d'orbite :

Définition 5. On dit qu'une action d'un groupe G sur un ensemble X est *transitive* (ou que G agit *transitivement* sur X) s'il y a une unique orbite de X sous l'action de G, à savoir X.

Si l'action donnée est transitive, X n'est pas vide. Supposons inversement X non vide. Il y a deux façons, évidemment équivalentes, d'exprimer la transitivité de l'action donnée. La première est de dire qu'il existe un point $a \in X$ tel que $G{\cdot}a = X$. La deuxième façon est de dire que deux points x, y quelconques de X ont la même orbite sous G, *i.e.* $G \cdot x = G \cdot y$; cela signifie, rappelons-le, qu'il existe $g \in G$ tel que $y = g \cdot x$.

Exemple. L'action par translations à gauche d'un groupe G sur lui-même (exemple 8 de la page 60) est *transitive* : étant donnés $x, y \in G$, il existe un élément $g \in G$ tel que $g{\cdot}x = y$, soit $y = gx$. En fait un tel g est unique (et $g := yx^{-1}$). Cela donne un exemple d'action *simplement transitive*, c'est-à-dire à la fois simple et transitive. En général, une action de G sur un ensemble non vide X est simplement transitive, par définition, si, pour tous $x, y \in X$, il existe un *unique* $g \in G$ transformant x en y. Autrement dit, pour tout $a \in X$, $g \mapsto g \cdot a$ est une bijection de G sur X. On dit aussi que G agit *simplement transitivement* sur X.
Avec cette terminologie, l'exercice 2 de la page 62 se traduit comme suit. Se donner un espace affine sur un corps K revient à se donner trois ingrédients : un ensemble non vide X, un K-espace vectoriel E, et une action *simplement transitive* du groupe $(E, +)$ sur X.

2.2 Points fixes. Stabilisateur

Définition 6. Soient G un groupe agissant sur un ensemble X et $x \in X$.

1. Lorsque $g \in G$ et $g \cdot x = x$, on dit que x est un *point fixe* de g, ou que g laisse fixe x. Si x est point fixe de *tout* élément de G, on dit que x est un *point fixe de* G.
2. L'ensemble des $g \in G$ laissant fixe x est appelé *stabilisateur* ou *groupe d'isotropie* de x, il est noté G_x. Ainsi :

$$G_x := \big\{ g \in G \;\big|\; g \cdot x = x \big\}.$$

Proposition 9. Sous les hypothèses de la définition, G_x est un sous-groupe de G.

Démonstration. D'abord $1 \cdot x = x$, donc $1 \in G_x$.
Si $g, h \in G_x$, l'associativité mixte donne $(gh) \cdot x = g \cdot (h \cdot x) = g \cdot x = x$, d'où $gh \in G_x$.
Enfin, pour tout $g \in G$, $g \in G_x \iff g^{-1} \in G_x$, en vertu de la propriété de la page 58. ∎

Remarque. L'action de G sur X est *simple* si, et seulement si, $G_x = \{1\}$ pour tout $x \in X$. C'est le cas, par exemple, si G agit sur lui-même par translations à gauche. Par contre, si G agit sur lui-même par conjugaison et si $a \in G$, le stabilisateur G_a est :

$$G_a = \{g \in G \mid gag^{-1} = a\} = \{g \in G \mid ga = ag\}.$$

Ainsi G_a est formé des éléments de G commutant avec a, il contient toujours a. Cette action n'est donc simple que si $G = \{1\}$.

Exercice 7.
Identifions tout automorphisme (linéaire) de $\mathbb{R}^2$ avec sa matrice dans la base canonique (e_1, e_2) de $\mathbb{R}^2$. Dans l'action naturelle de $GL_2(\mathbb{R})$ sur $\mathbb{R}^2$, quel est le stabilisateur de e_1 ?

Solution. Soit $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2(\mathbb{R})$. Puisque $A(e_1) = ae_1 + ce_2$, A appartient au stabilisateur $GL_2(\mathbb{R})_{e_1}$ de e_1 si, et seulement si, $a = 1$ et $c = 0$. Ainsi :

$$GL_2(\mathbb{R})_{e_1} = \left\{ \begin{pmatrix} 1 & b \\ 0 & d \end{pmatrix} \;\middle|\; (b, d) \in \mathbb{R} \times \mathbb{R}^* \right\}.$$

Ce stabilisateur est donc formé des automorphismes $(x, y) \longmapsto (x + by, dy)$, où (b, d) parcourt $\mathbb{R} \times \mathbb{R}^*$.

Lorsqu'un groupe G agit sur un ensemble X, il agit *ipso facto* sur $\mathcal{P}(X)$, *cf.* l'exemple 6 de la page 60. Pour cette action de G sur $\mathcal{P}(X)$, nous pouvons encore parler de points fixes et de stabilisateurs. Ce point de vue va s'avérer très fructueux. Dans cette optique, voici une définition importante :

Définition 7. Soient G un groupe agissant sur un ensemble X et $A \subset X$.

1. On dit que A est *stable* (ou *globalement invariante*) sous G si $g \cdot A = A$ pour tout $g \in G$, c'est-à-dire si $(g, a) \in G \times A \Longrightarrow g \cdot a \in A$.
2. L'ensemble des $g \in G$ tels que $g \cdot A = A$ est appelé *stabilisateur* de A sous G, il est noté $\mathrm{Stab}(A)$. Ainsi :

$$\mathrm{Stab}(A) := \{g \in G \mid g \cdot A = A\}.$$

Il convient d'observer que le stabilisateur de A est un sous-groupe de G : cela vient de la proposition précédente, appliquée à l'action de G sur $\mathcal{P}(X)$. Par ailleurs, le « c'est-à-dire » de l'énoncé demande une justification. L'implication $(g, a) \in G \times A \Longrightarrow g \cdot a \in A$

signifie ceci : $(\forall g \in G,\ g \cdot A \subset A)$. Comment alors en déduire $(\forall g \in G,\ g \cdot A = A)$? Voici l'idée : par hypothèse, l'inclusion $g \cdot A \subset A$ est vraie *pour tout* $g \in G$. Soit maintenant $g \in G$; nous avons *en particulier* $g^{-1} \cdot A \subset A$, et l'on en déduit $A \subset g \cdot A$ par associativité mixte, comme dans la preuve de la propriété de la page 58.

Soit $g \in G$. Si tout point de A est fixé par g, on a $g \cdot A = A$, mais la réciproque est fausse. Par exemple, $g \cdot X = X$, mais l'action donnée n'est pas forcément triviale ! Pour une partie A de X, on dit donc « A est stable sous G » plutôt que « A est fixe sous G ».

La preuve du résultat suivant est évidente et sera laissée au lecteur :

Proposition 10. Soient G un groupe agissant sur un ensemble X et A une partie de X. Alors $(g, a) \mapsto g \cdot a$ est une action de $\mathrm{Stab}(A)$ sur A, dite *induite* par l'action donnée. Si A est stable sous G, cette action induite est une action de G sur A.

Exemple. Soit G un groupe, agissant sur lui-même par conjugaison. Soit $\varphi : G \to \mathrm{Aut}(G)$ le morphisme correspondant (exemple 9 de la page 60). Soit ensuite $\mathcal{G}$ l'ensemble des sous-groupes de G. Faisons agir G (par conjugaison) sur $\mathcal{P}(G)$: pour tous $g \in G$ et $A \subset G$, $g \cdot A := gAg^{-1}$. Dans cette nouvelle action, $\mathcal{G}$ est une partie de $\mathcal{P}(G)$ stable sous G : si H est un sous-groupe de G et $g \in G$, gHg^{-1} est encore un sous-groupe de G, comme image de H par l'automorphisme $\varphi(g) : x \mapsto gxg^{-1}$ de G.
Vu la proposition 10, nous disposons donc d'une action *par conjugaison* de G sur $\mathcal{G}$, donnée par $(g, H) \mapsto gHg^{-1}$. Nous dirons que deux sous-groupes H, H' de G sont *conjugués* s'il existe $g \in G$ tel que $H' = gHg^{-1}$. Dans ce cas, H et H' sont en particulier *isomorphes* $\big($penser à $\varphi(g)\big)$.
Observons également qu'un sous-groupe A de G est *distingué* (*cf.* le module I.1) si, et seulement si, A est stable par conjugaison.

Exercice 8.

Soient G un groupe agissant sur un ensemble X et $A \subset X$. Montrer que A est stable sous G si, et seulement si, A est réunion de certaines des orbites de X sous G.

Solution. Remarquons en passant que toute réunion (resp. intersection) de parties de X stables sous G est stable sous G. Cela étant, soit $\mathcal{R}$ la relation d'équivalence sur X correspondant à l'action donnée. La stabilité de A sous G signifie ceci : pour tous $a \in A$ et $x \in X$, $(a\mathcal{R}x) \Longrightarrow (x \in A)$, ce que l'on peut encore traduire en disant que A contient chaque classe d'équivalence (orbite) qu'elle rencontre. La conclusion est alors claire.

Retenons en particulier de cet exercice que, si A est une orbite de l'action d'un groupe G sur un ensemble X, $(g, a) \mapsto g \cdot a$ est une action de G sur A. Cette action est *transitive*, par définition d'une orbite.

Exemple. Soient G le groupe des rotations vectorielles de $\mathbb{R}^2$ et r un réel strictement positif. Dans l'exercice 5 de la page 67, nous avons vu que le cercle $C(O, r)$ de centre O et de rayon r est l'une des orbites de l'action naturelle de G sur $\mathbb{R}^2$. Ainsi G agit *transitivement*

sur $C(O,r)$. En fait, cette action de G sur $C(O,r)$ est *simplement* transitive. En effet, une rotation autre que l'identité n'a qu'un point fixe, à savoir O.

Remarque. La notion de *repère* peut s'interpréter en termes d'actions de groupes. Ainsi, soit E un espace vectoriel de dimension finie $n \geqslant 1$. Le groupe $\mathcal{GL}(E)$ agit naturellement sur E, puis sur E^n, et donc sur $\mathcal{P}(E^n)$. Si $u \in \mathcal{GL}(E)$ et si $(e_1, \ldots, e_n)$ est une base (ou repère) de E, $\big(u(e_1), \ldots, u(e_n)\big)$ est aussi une base de E. Ainsi l'ensemble des bases de E est une partie de $\mathcal{P}(E^n)$ stable sous $\mathcal{GL}(E)$. Mieux : l'action de $\mathcal{GL}(E)$ sur l'"ensemble des bases de E est *simplement transitive*. En effet, si $(e_1, \ldots, e_n)$ et $(e'_1, \ldots, e'_n)$ sont deux bases de E, il existe un *unique* $u \in \mathcal{GL}(E)$ tel que $u(e_i) = e'_i$ pour $i = 1, \ldots, n$, c'est-à-dire $u \cdot (e_1, \ldots, e_n) = (e'_1, \ldots, e'_n)$.
Anticipons sur les modules I.4 et I.6. Soit E un espace vectoriel euclidien de dimension $n \geqslant 1$. Comme ci-dessus, l'action du *groupe orthogonal* $\mathcal{O}(E)$ sur l'ensemble des bases orthonormées de E est simplement transitive : si $(e_1, \ldots, e_n)$ et $(e'_1, \ldots, e'_n)$ sont deux bases orthonormées de E, il existe un *unique* $u \in \mathcal{O}(E)$ tel que $u(e_i) = e'_i$ pour $i = 1, \ldots, n$, c'est-à-dire $u \cdot (e_1, \ldots, e_n) = (e'_1, \ldots, e'_n)$.

Proposition 11. Soit G un groupe agissant sur un ensemble X. Lorsque deux points de X appartiennent à une même orbite sous G, leurs stabilisateurs sont deux sous-groupes *conjugués*, en particulier *isomorphes*, de G. En fait :

$$G_{g\cdot x} \;=\; gG_xg^{-1} \quad \text{pour tout} \quad (g,x) \in G \times X. \tag{7}$$

Démonstration. Soient $x \in X$, $\gamma \in G_x$ et $\delta := g\gamma g^{-1}$. Alors :

$$\delta \cdot (g \cdot x) = (\delta g) \cdot x = \big[(g\gamma g^{-1})g\big] \cdot x = (g\gamma) \cdot x = g \cdot (\gamma \cdot x) = g \cdot x,$$

d'où $\delta \in G_{g\cdot x}$. On en déduit l'inclusion $gG_xg^{-1} \subset G_{g\cdot x}$. Cela étant vrai pour tous g, x, remplaçons x par $g \cdot x$ et g par g^{-1}. Nous obtenons l'inclusion $g^{-1}G_{g\cdot x}\, g \subset G_x$, soit $G_{g\cdot x} \subset gG_xg^{-1}$. D'où l'égalité (7).
Soient alors x, y deux points de X appartenant à une même orbite sous G : il existe $g \in G$ tel que $y = g \cdot x$. D'après ce qui précède, $G_y = G_{g\cdot x} = gG_xg^{-1}$, donc G_x et G_y sont conjugués. ∎

2.3 Classes modulo un sous-groupe

Dans le début de cette section, nous considérons un groupe multiplicatif G et un sous-groupe H de G. Dans le module I.1, nous avons vu que la relation binaire $\mathcal{R}$ sur G définie par $g\,\mathcal{R}\,g'$ si $g^{-1}g' \in H$ est une relation d'équivalence. Les *classes à gauche de G modulo H* sont, par définition, les classes d'équivalence de $\mathcal{R}$. Chacune de ces classes est une partie de G. De même, soit $\mathcal{R}'$ la relation d'équivalence sur G définie par $g\,\mathcal{R}'\,g'$ si $g'g^{-1} \in H$. Les classes à droite de G modulo H sont, par définition, les classes d'équivalence de $\mathcal{R}'$. Si G est commutatif, il n'y a pas à distinguer les classes à gauche des classes à droite, car $\mathcal{R} = \mathcal{R}'$.

Proposition 12. Soit $g \in G$. La classe à gauche (resp. à droite) de g modulo H est :

$$gH := \{gh \mid h \in H\}, \quad \text{resp.} \quad Hg := \{hg \mid h \in H\}.$$

Démonstration. Soit $\gamma \in G$. Par définition, γ appartient à la classe à gauche de g modulo H, *i.e.* $g \mathcal{R} \gamma$, si, et seulement si, $g^{-1}\gamma \in H$, ce qui équivaut à $\gamma \in gH$, d'où le premier résultat. Le second est analogue. ∎

Définition 8. L'ensemble des classes à gauche de G modulo H est noté G/H. Le cardinal de G/H, fini ou non, est appelé *indice* de G dans H et on le note $(G : H)$. Ainsi $(G : H) := \operatorname{card}(G/H)$.

A priori, G/H n'est qu'un ensemble : nous n'avons pas supposé que H était un sous-groupe *distingué* de G, nous ne disposons donc pas de la loi de groupe sur G/H correspondante (*cf.* le module I.1).

L'ensemble des classes à droite de G modulo H est noté $H \setminus G$. Il a le même cardinal que G/H. Plus précisément, $C \longmapsto C^{-1} := \{g^{-1} \mid g \in C\}$ est une bijection de G/H sur $H \setminus G$ (le vérifier).
Comme conséquence du résultat suivant, nous retrouvons le théorème de Lagrange (*cf.* [L1]) : dans un groupe fini, l'ordre de tout sous-groupe divise l'ordre du groupe.

Proposition 13. Si G est un groupe fini et H est un sous-groupe de G, on a :

$$\operatorname{card}(G) = \operatorname{card}(H) \times (G : H) \quad \text{soit} \quad (G : H) = \frac{\operatorname{card}(G)}{\operatorname{card}(H)}.$$

Démonstration. Si $g \in G$, $h \mapsto gh$ est une bijection de H sur gH, donc chaque classe à gauche de G modulo H a même cardinal que H. La conclusion vient alors du fait que ces classes forment une partition de G. ∎

Théorème 14. Soient G un groupe agissant sur un ensemble X et a un point de X.

1. L'application $g \mapsto g \cdot a$ de G dans X induit une bijection de G/G_a (ensemble des classes à gauche de G modulo G_a) sur l'orbite $G \cdot a$.
2. L'orbite $G \cdot a$ est finie si, et seulement si, l'indice $(G : G_a)$ de G_a dans G est fini, auquel cas $\operatorname{card}(G \cdot a) = (G : G_a)$.
3. Si G est fini, l'orbite $G \cdot a$ est finie, et l'on a :

$$\operatorname{card}(G \cdot a) = \frac{\operatorname{card}(G)}{\operatorname{card}(G_a)}. \tag{8}$$

Démonstration. L'application $f : g \mapsto g \cdot a$ est, par définition, une application surjective de G sur $G \cdot a$. Soient $\mathcal{R}$ la relation d'équivalence sur G associée à f et $G/\mathcal{R}$ l'ensemble

quotient correspondant. D'après le module I.1, f induit une bijection F de $G/\mathcal{R}$ sur $G\cdot a$:

$$F\big(\mathrm{Cl}(g)\big) \;:=\; f(g) \;=\; g\cdot a \quad \text{pour tout} \quad g\in G,$$

en notant $\mathrm{Cl}(g)$ la classe modulo $\mathcal{R}$ de $g \in G$. Explicitons la relation $\mathcal{R}$. Pour tous $g, g' \in G$, nous avons les équivalences :

$$(g\,\mathcal{R}\,g') \iff (g\cdot a = g'\cdot a) \iff \big((g^{-1}g')\cdot a = a\big) \iff (g^{-1}g' \in G_a).$$

Cela montre que $\mathcal{R}$ n'est autre que la relation d'équivalence dont les classes sont les classes à gauche de G modulo G_a. Ainsi $G/\mathcal{R} = G/G_a$, et l'assertion 1 en résulte.

Puisque F est bijective, G/G_a et $G\cdot a$ ont même cardinal. Or $\mathrm{card}\,(G/G_a) = (G : G_a)$, ce qui prouve l'assertion 2. L'assertion 3 est une conséquence de 2 et de la proposition 13 de la page ci-contre. ∎

Voici une application « géométrique » très simple de ce théorème, illustrant les notions d'orbite et de point fixe.

Exercice 9.

Soient X un espace affine réel, G le barycentre d'une famille finie $\big((A_i,\lambda_i)\big)_{i\in I}$ (les A_i sont des points de X et les λ_i sont des réels de somme non nulle), et f une application affine de X dans un espace affine Y. Montrer que $f(G)$ est le barycentre de la famille $\big((f(A_i),\lambda_i)\big)_{i\in I}$. En déduire que, si Γ est un sous-groupe *fini* du groupe affine $\mathcal{GA}(X)$, il existe un point de X fixe sous Γ.

Solution. Le point G est caractérisé par l'égalité $\sum\limits_{i\in I}\lambda_i\overrightarrow{GA_i} = 0$ (*cf.* [L1]). Or la partie linéaire $\vec{f}$ de f est linéaire, donc :

$$\vec{0} \;=\; \sum_{i\in I}\lambda_i\vec{f}\big(\overrightarrow{GA_i}\big) \;=\; \sum_{i\in I}\lambda_i\overrightarrow{f(G)f(A_i)},$$

ce qui prouve la première assertion.

Soient maintenant M un point arbitraire de X et $\Omega := \Gamma\cdot M$ son orbite sous Γ. Le théorème précédent montre que Ω est finie. L'idée est alors de considérer l'*isobarycentre* (ou *centre de gravité*) G de Ω, c'est-à-dire le barycentre de la famille $\big((x,1)\big)_{x\in\Omega}$: tous les points de Ω sont affectés de la « masse » 1. Nous allons montrer que G est fixe sous Γ.

Soit $\gamma\in\Gamma$. Puisque γ est affine, $\gamma(G)$ est l'isobarycentre de la famille $\Big(\big(\gamma(x),1\big)\Big)_{x\in\Omega}$, nous venons de le voir. Puisque Ω est une orbite sous Γ, $\gamma(\Omega)=\Omega$, *i.e.* $x\mapsto\gamma(x)$ est une permutation de Ω. Or l'isobarycentre d'une famille finie de points ne dépend pas de l'ordre dans lequel ces points sont pris, puisque tous les coefficients valent 1. Il en résulte que $\gamma(G)=G$. Cela étant vrai pour tout $\gamma\in\Gamma$, G est un point fixe sous G.

Terminons par une remarque. Soient A, B, M trois points de X tels que M soit le milieu de $[AB]$, *i.e.* le barycentre de $\big((A,1),(B,1)\big)$. Cela revient à dire que $B = s_M(A)$, en notant $s_M \in \mathcal{GA}(X)$ la symétrie centrale de centre M. Soit $f\in GA(X)$. Le début de l'exercice montre que $f(M)$ est milieu de $[f(A)f(B)]$, *i.e.* $f(B) = s_{f(M)}\big(f(A)\big)$. Ainsi $(f\circ s_M)(A) = (s_{f(M)}\circ f)(A)$. Cette égalité est vraie pour tout A, d'où la formule de conjugaison suivante, valable pour tous $M\in X$ et $f\in\mathcal{GA}(X)$:

$$f\circ s_M\circ f^{-1} \;=\; s_{f(M)}. \tag{9}$$

Une application algébrique typique du théorème précédent est donnée dans la proposition ci-après (voir aussi la proposition 22 de la page 83). Donnons d'abord une définition :

Définition 9. Soit p un nombre premier. On appelle *p-groupe* tout groupe fini dont l'ordre est une puissance de p.

Proposition 15. Soit p un nombre premier. Si un p-groupe G agit sur un ensemble fini X, le nombre de points fixes de X sous G est congru modulo p au cardinal de X.

Démonstration. Soit X^G l'ensemble des points fixes de X sous G, formé des $x \in X$ tels que $G \cdot x = \{x\}$. Soit ensuite T un ensemble de représentants des orbites de X sous G. Il contient X^G, ce qui nous permet de poser $T' := T \setminus X^G$. L'égalité de la proposition 8 de la page 65 peut être écrite ainsi :

$$\operatorname{card}(X) \;=\; \operatorname{card}(X^G) + \sum_{a \in T'} \operatorname{card}(G \cdot a).$$

Il nous suffit de montrer que, pour tout $a \in T'$, $\operatorname{card}(G \cdot a)$ est multiple de p. D'après l'égalité (8), $\operatorname{card}(G{\cdot}a)$ divise $\operatorname{card}(G)$, qui est une puissance de p. Comme p est premier, $\operatorname{card}(G \cdot a)$ est aussi une puissance de p : $\operatorname{card}(G \cdot a) = p^r$, où $r \in \mathbb{N}$. En fait $r \geqslant 1$, car $a \notin X^G$, *i.e.* l'orbite $G \cdot a$ n'est pas ponctuelle. Ainsi p divise $p^r = \operatorname{card}(G \cdot a)$. ∎

Théorème 16 (Cauchy). Soient G un groupe fini d'ordre N et p un facteur premier de N. Alors G possède (au moins) un élément d'ordre p.

Démonstration. L'application $s : (x_1, x_2, \ldots, x_p) \mapsto (x_2, x_3 \ldots, x_p, x_1)$ est une permutation d'ordre p de $X := G^p$. Ainsi $\langle s \rangle$ est un sous-groupe d'ordre p de $\mathfrak{S}(X)$. Les points fixes de s sont évidemment les p-uplets $(x, x, \ldots, x)$, où $x \in G$. Considérons maintenant la partie Y de X suivante :

$$Y \;:=\; \big\{(x_1, \ldots, x_p) \in X \;\big|\; x_1 x_2 \cdots x_p = 1\big\}.$$

Montrons que Y est stable sous $\langle s \rangle$. Soit donc $\xi = (x_1, \ldots, x_p) \in Y$, de sorte que l'on a $s(\xi) = (x_2, x_3 \ldots, x_p, x_1)$. Alors :

$$x_2 x_3 \cdots x_p x_1 \;=\; x_1^{-1}(x_1 x_2 \cdots x_p) x_1 \;=\; x_1^{-1}(1) x_1 \;=\; 1\,,$$

d'où $s(\xi) \in Y$. Ainsi $s(Y) \subset Y$, *i.e.* $s(Y) = Y$, puisque Y est fini.

Il est clair que $(x_1, \ldots, x_{p-1}) \longmapsto \big(x_1, \ldots, x_{p-1}, (x_1 x_2 \cdots x_{p-1})^{-1}\big)$ est une bijection de G^{p-1} sur Y, et ainsi $\operatorname{card}(Y) = N^{p-1}$. En particulier, puisque p divise N et $p - 1 \geqslant 1$, p divise $\operatorname{card}(Y)$, autrement dit $\operatorname{card}(Y) \equiv 0 \mod p$. D'après la proposition précédente, p divise donc le nombre de points fixes de Y sous $\langle s \rangle$. Ces points fixes sont exactement les p-uplets $(x, x, \ldots, x)$, où $x \in G$ et $x^p = 1$. Ainsi le nombre m de solutions de l'équation $x^p = 1$ dans G est multiple de p. Or 1 est une telle solution, donc $m \geqslant 1$. Dans ces conditions $m \geqslant p$ (car p divise m), en particulier $m \geqslant 2$. Il existe donc un $x \in G$ tel que $x^p = 1$ et $x \neq 1$, ce qui signifie que x est d'ordre p. ∎

3 Quelques applications des actions de groupes

3.1 Problèmes de classification

Comme premier exemple de problème de classification, soient K un corps, $n, p \in \mathbb{N}^*$ et $X := M_{n,p}(K)$. Rappelons que deux matrices $A, B \in X$ sont dites *équivalentes* s'il existe des matrices inversibles $P \in GL_p(K)$ et $Q \in GL_n(K)$ telles que $B = QAP$. Il s'agit d'une relation d'équivalence $\mathcal{R}$ sur X. Le groupe produit $G := GL_p(K) \times GL_q(K)$ agit sur X par la formule : $(P, Q) \cdot A := QAP^{-1}$; on prend QAP^{-1}, et non QAP, pour que l'associativité mixte soit satisfaite. Alors $\mathcal{R}$ est la relation d'équivalence associée à cette action, dont les orbites sont donc les classes modulo $\mathcal{R}$.

Classifier les matrices $A \in X$ à équivalence près, c'est décrire l'ensemble des orbites de X sous G, ou expliciter un ensemble de représentants de ces orbites. Nous savons résoudre ce problème : deux matrices sont équivalentes si, et seulement si, elles ont même rang. Soit $m := \min(n, p)$. Pour tout $r \in [\![0, m]\!]$, posons $X_r := \{A \in X \mid \operatorname{rang}(A) = r\}$. Les orbites de X sous G sont : $X_0, \ldots, X_m$. Lorsque r décrit $[\![0, m]\!]$, les matrices $I_r^{(n,p)}$, définies dans le module II.7 de [L1], forment un ensemble de représentants des orbites de X sous G.

Voici un deuxième exemple. Posons $X := M_n(K)$. Deux matrices carrées $A, B \in X$ sont dites *semblables* s'il existe une matrice inversible $P \in GL_n(K)$ telle que $B = PAP^{-1}$. Il s'agit d'une relation d'équivalence $\mathcal{R}'$ sur X. Le groupe $G := GL_n(K)$ agit sur X par la formule : $P \cdot A := PAP^{-1}$. La classification des matrices carrées à similitude près, *i.e.* la description des orbites de X sous G, est un problème plus difficile que le précédent. Ce problème est résolu dans [MPA1].

Ces deux problèmes de classification peuvent ainsi être formulés dans le même langage, celui des actions de groupes. Évidemment, cette formulation ne donne pas une solution du problème de classification, solution qui est simple dans le premier cas mais pas dans le second.

3.2 Groupes de symétries

Une deuxième application des actions de groupes est l'étude de ce que l'on appelle *groupe des symétries d'une figure*. Voici ce dont il s'agit :

Définition 10.
Soient X un espace affine et A une partie ou « figure » de X. On appelle *groupe des symétries de* A le stabilisateur de A pour l'action naturelle du groupe affine $\mathcal{GA}(X)$ sur X, c'est donc le sous-groupe $\operatorname{Stab}(A) := \{g \in \mathcal{GA}(X) \mid g(A) = A\}$ de $\mathcal{GA}(X)$.

La terminologie « groupe des symétries » est classique, mais pas très heureuse, car les éléments du groupe $\operatorname{Stab}(A)$ ne sont pas forcément des symétries (affines). On parle aussi de « groupe des transformations affines de X laissant la figure A globalement invariante ».

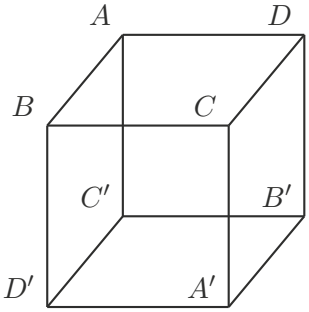
Exemple 1.
Dans le plan $\mathbb{R}^2$, soit C le cercle unité, de centre O et de rayon 1. Déterminons le groupe des symétries G de C. Intuitivement, C est une figure « très symétrique », *i.e.* G doit être « assez gros ».
Observons d'abord que O est le seul centre de symétrie de C, *i.e.* le seul point A tel que la symétrie centrale s_A de centre A appartienne à G (le vérifier). Il en résulte que O est un point fixe de G. En effet, si $f \in G$, la formule (9) montre que $s_{f(O)} = f \circ s_O \circ f^{-1}$ appartient aussi à G, de sorte que $f(O)$ est centre de symétrie de C et par suite $f(O) = O$. Cela montre que G est en fait un sous-groupe de $\mathcal{GL}(\mathbb{R}^2)$, car les transformations *linéaires* de $\mathbb{R}^2$ sont exactement les transformations affines laissant fixe O.
Un élément f de $\mathcal{GL}(\mathbb{R}^2)$ appartient à G si et seulement si, pour tout vecteur $u \in \mathbb{R}^2$, $\big(\|u\| = 1\big) \Longrightarrow \big(\|f(u)\| = 1\big)$, ce qui revient à dire que $\|f(u)\| = \|u\|$ pour tout $u \in \mathbb{R}^2$. Anticipant sur le module I.4, on en déduit que $G = \mathcal{O}(\mathbb{R}^2)$ est formé des transformations orthogonales de $\mathbb{R}^2$. En identifiant chaque endomorphisme de $\mathbb{R}^2$ avec sa matrice dans la base canonique de $\mathbb{R}^2$, G est donc formé des matrices de l'un des deux types suivants, $\theta \in \mathbb{R}$ étant arbitraire : $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ ou $\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$. La première matrice représente la rotation d'angle θ, la deuxième représente la symétrie orthogonale par rapport à la droite d'équation $x\sin(\theta/2) - y\cos(\theta/2) = 0$, le lecteur pourra le vérifier. Ainsi, G est formé des rotations de centre O et des symétries orthogonales par rapport aux droites passant par O.

Exemple 2 : groupe des symétries d'un cube.
Dans l'espace $\mathbb{R}^3$, soit $K := \{-1, 1\}^3$. Les points de K sont les huit sommets d'un cube : A, B, C, D (et A', B', C', D' leurs symétriques par rapport à O), comme indiqué sur la figure ci-après. Nous allons déterminer le groupe G des symétries de K. D'abord O est l'isobarycentre de K, c'est donc un point fixe de G, d'après l'exercice 9 de la page 73. Autrement dit $G \subset \mathcal{GL}(\mathbb{R}^3)$. Nous identifierons $\mathcal{GL}(\mathbb{R}^3)$ à $GL_3(\mathbb{R})$, en identifiant tout $u \in \mathcal{GL}(\mathbb{R}^3)$ avec sa matrice dans la base canonique $\mathcal{B} := (e_1, e_2, e_3)$ de $\mathbb{R}^3$.

1) Posons $D_i := \mathbb{R}e_i$ pour $i = 1, 2, 3$ et $\mathcal{D} := \{D_1, D_2, D_3\}$. Montrons que $\mathcal{D}$ est stable sous G. L'idée est la suivante. Notons T l'ensemble des points de $\mathbb{R}^3$ ayant une seule coordonnée non nulle, égale à ± 1. Ces points sont les centres des six « faces » du cube. Pour qu'un point M de $\mathbb{R}^3$ appartienne à T, il faut, et il suffit, qu'il existe quatre points distincts $N_1, N_2, N_3, N_4 \in K$ tels que M soit milieu de $[N_1N_2]$ et de $[N_3N_4]$. Par exemple, $(0, -1, 0)$ est milieu de $[AD']$ et de $[BC']$. Il en résulte d'abord que T est stable sous G, puis que $\mathcal{D}$ est stable sous G, car $\mathcal{D}$ est formée des droites $\mathbb{R}\overrightarrow{OM}$, M parcourant T. D'où une action naturelle de G sur $\mathcal{D}$, correspondant à un morphisme $\psi : G \to \mathfrak{S}(\mathcal{D}) \equiv \mathfrak{S}_3$.

2) Le groupe G possède deux sous-groupes évidents. Le premier, notons-le Γ, est formé des matrices de permutations P_s, $s \in \mathfrak{S}_3$. Si $s \in \mathfrak{S}_3$ et $i \in \{1, 2, 3\}$, $P_s(e_i) := e_{s(i)}$; nous savons que $\det(P_s) = \varepsilon(s)$ est la signature de s. Il est clair que la restriction de ψ

à Γ est un isomorphisme de Γ sur $\mathfrak{S}(\mathcal{D})$, ce qui montre que ψ est surjectif.
Le deuxième sous-groupe Δ de G est formé des matrices diagonales $\mathrm{Diag}(\varepsilon_1, \varepsilon_2, \varepsilon_3)$, où $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ décrit $\{-1, 1\}^3$. Il est clair que Δ est isomorphe à $\big(\{-1, 1\}^3, \times\big)$, ou encore à $\big(\mathbb{Z}/2\mathbb{Z}\big)^3, +\big)$. De plus $\Delta \subset \mathrm{Ker}\,\psi$, et il y a en fait égalité. Soit en effet $g \in \mathrm{Ker}\,\psi$. Puisque g laisse stable chacune des droites D_i, g est une matrice diagonale $\mathrm{Diag}(a_1, a_2, a_3)$. Mais $g(C) \in K$, donc $a_1, a_2, a_3 \in \{-1, 1\}$ et par suite $g \in \Delta$. Ainsi $\Delta = \mathrm{Ker}\,\psi$. D'après le théorème d'isomorphisme, $\mathfrak{S}(\mathcal{D})$ est isomorphe au quotient G/Δ. Comme $\mathfrak{S}(\mathcal{D})$ est d'ordre $3! = 6$ et Δ d'ordre $2^3 = 8$, nous avons déjà déterminé l'ordre de G : $\mathrm{card}\,(G) = 8 \times 6 = 48$.
Il est clair que l'intersection $\Gamma \cap \Delta$ est réduite à l'élément neutre : I_3 est la seule matrice de permutation diagonale. On en déduit que tout $g \in G$ s'écrit de manière unique $g = d \circ \gamma$, où $(d, \gamma) \in \Delta \times \Gamma$. Il en résulte que $\det(g) = \pm 1$ pour tout $g \in G$.



$A = (-1, -1, 1)$

$B = (1, -1, 1)$

$C = (1, 1, 1)$

$D = (-1, 1, 1)$

3) Il y a une action naturelle de G sur les « diagonales » du cube. Posons :

$$\alpha := \{A, A'\},\ \beta := \{B, B'\},\ \gamma := \{C, C'\},\ \delta := \{D, D'\},\ \mathcal{A} := \{\alpha, \beta, \gamma, \delta\}.$$

Puisque O est fixe sous G, les images par un élément $g \in G$ de deux points de K symétriques par rapport à O sont aussi symétriques par rapport à O. Ainsi $\mathcal{A}$ est stable sous G. D'où une action de G sur $\mathcal{A}$, correspondant à un morphisme $\varphi : G \to \mathfrak{S}(\mathcal{A})$.
Si $\sigma := -I_3 \in G$ est la symétrie par rapport à O, $\varphi(\sigma) = 1$. Soit r la rotation d'angle $\pi/2$ autour de l'axe Oz : matriciellement, $r = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Puisque r permute circulairement A, B, C, D, r appartient à G et $\varphi(r)$ est le 4-cycle $(\alpha\ \beta\ \gamma\ \delta)$. Soit ensuite t la symétrie orthogonale par rapport au plan d'équation $y = z$. Matriciellement, le lecteur vérifiera que $t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$. Chacun des points C, C', D, D' est fixé par t, et t échange A et B'. Ainsi $t \in G$, et $\varphi(t)$ est la transposition $(\alpha\ \beta)$. Comme $(\alpha\ \beta\ \gamma\ \delta)$ et $(\alpha\ \beta)$ engendrent $\mathfrak{S}(\mathcal{A})$, le morphisme φ est surjectif. Le groupe G étant d'ordre 48, $\mathrm{Ker}\,\varphi$ est alors d'ordre 2, *i.e.* $\mathrm{Ker}\,\varphi = \langle\sigma\rangle$.

4) Le morphisme $\det : G \to \mathbb{R}^*$ a pour image $\{-1, 1\}$, son noyau G^+ est donc d'ordre 24. Comme $\sigma \notin G^+$, la restriction de φ à G^+ est injective, c'est donc un isomorphisme

de G^+ sur $\mathfrak{S}(\mathcal{A})$. Ainsi G^+ est isomorphe à $\mathfrak{S}_4$. Cela va nous permettre de préciser la structure de G. Puisque σ commute avec tout élément de G, la restriction de la multiplication à $G^+ \times \langle\sigma\rangle$ induit un morphisme de $G^+ \times \langle\sigma\rangle$ dans G. Ce morphisme est injectif parce que $\sigma \notin G^+$, c'est donc un isomorphisme, car $G^+ \times \langle\sigma\rangle$ et G sont tous les deux d'ordre 48. Conclusion :

$$G \quad \text{est isomorphe à} \quad \mathfrak{S}_4 \times (\mathbb{Z}/2\mathbb{Z}).$$

5) Le groupe G est en fait un sous-groupe du groupe orthogonal $\mathcal{O}(\mathbb{R}^3)$ (*cf.* le module I.4). En effet, G^+ est engendré par r et $-t$, car $\varphi(r) = (\alpha\ \beta\ \gamma\ \delta)$ et $\varphi(-t) = (\alpha\ \beta)$ engendrent $\mathfrak{S}(\mathcal{A})$. Il en résulte que G est engendré par $r, -t$ et σ. Or $r, -t, \sigma \in \mathcal{O}(\mathbb{R}^3)$, et par suite $G \subset \mathcal{O}(\mathbb{R}^3)$.

À titre d'illustration, soit g le seul élément de G^+ tel que $\varphi(g) = (\beta\ \delta\ \gamma)$. Montrons que g est la rotation d'angle $2\pi/3$ autour de l'axe $[AA']$ dirigé par $\overrightarrow{A'A}$. Le triangle $BC'D$ est équilatéral, car $BC' = C'D = DB = 2\sqrt{2}$, son isobarycentre I est défini par $\overrightarrow{OI} := \dfrac{1}{3}\overrightarrow{OA}$. Un calcul de déterminant montre que la base $\left(\overrightarrow{IC'}, \overrightarrow{IB}, \overrightarrow{IA}\right)$ de $\mathbb{R}^3$ est directe, *i.e.* $\det_{\mathcal{B}}(\overrightarrow{IC'}, \overrightarrow{IB}, \overrightarrow{IA}) > 0$. La rotation en question applique donc A, C', B, D sur A, B, D, C' respectivement, donc est égale à g.

6) Reprenons le morphisme $\psi : G \to \mathfrak{S}_3$ défini en 1).

On a $\psi(r) = (1\ 2)$ et $\psi(-t) = (2\ 3)$. Comme $(1\ 2)$ et $(2\ 3)$ engendrent $\mathfrak{S}_3$, la restriction de ψ à G^+ est un morphisme surjectif ψ' de G^+ sur $\mathfrak{S}_3$. Ainsi $V := \operatorname{Ker}\psi'$ est un sous-groupe distingué d'ordre 4 de G^+. Outre l'identité, les éléments de V sont les trois rotations d'angle π et d'axes D_1, D_2, D_3. Si l'on identifie $\mathfrak{S}(\mathcal{A})$ à $\mathfrak{S}_4$ en numérotant arbitrairement $\alpha, \beta, \gamma, \delta$, $\varphi(V)$ est un sous-groupe distingué d'ordre 4 de $\mathfrak{S}_4$. Plus précisément, $\varphi(V)$ est le sous-groupe V_4 défini dans [L1], à savoir $V_4 := \big\{1, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\big\}$. Ce sous-groupe distingué V_4 de $\mathfrak{S}_4$ est lié à la résolution des équations de degré 4 (*cf.* [MPA1]).

3.3 Applications au groupe symétrique

Comme troisième application des actions de groupes, nous allons étudier le groupe $\mathfrak{S}_n$, et en particulier décrire ses classes de conjugaison.

Commençons par introduire une définition. Soient X un ensemble et s une permutation de X. Le sous-groupe $\langle s\rangle$ de $\mathfrak{S}(X)$ engendré par s agit sur X, par restriction de l'action naturelle de $\mathfrak{S}(X)$ sur X.

Définition 11. On appelle *support* de s l'ensemble :

$$\operatorname{Supp}(s) := \big\{x \in X \;\big|\; s(x) \neq x\big\},$$

c'est-à-dire le complémentaire dans X de l'ensemble des points fixes de s.

On remarquera que le support de s et l'ensemble des points fixes de s sont stables sous l'action de $\langle s\rangle$ (le vérifier).

Proposition 17. Si $s, t \in \mathfrak{S}(X)$ ont des supports disjoints, $s \circ t = t \circ s$.

Démonstration. Si $x \in X$, il s'agit de montrer que $s\big(t(x)\big) = t\big(s(x)\big)$.

C'est évident si $s(x) = x = t(x)$. Supposons $s(x) \neq x$ (la situation est symétrique en s et t). Alors $x \in \mathrm{Supp}(s)$ donc, vu l'hypothèse, $x \notin \mathrm{Supp}(t)$, *i.e.* $t(x) = x$. Ainsi $s\big(t(x)\big) = s(x)$.

De la même manière, $s(x) \in \mathrm{Supp}(s)$, donc $t\big(s(x)\big) = s(x) = s\big(t(x)\big)$. ∎

La réciproque est fausse : par exemple une permutation distincte de l'identité commute avec elle-même.

Exercice 10.
Si $s, t \in \mathfrak{S}(X)$, montrer que $\mathrm{Supp}(s \circ t \circ s^{-1}) = s\big(\mathrm{Supp}(t)\big)$.
Solution. Cela vient des équivalences, valables pour tout $x \in X$:

$$\begin{aligned} x \notin \mathrm{Supp}(s \circ t \circ s^{-1}) &\iff (s \circ t \circ s^{-1})(x) = x \\ &\iff t\big(s^{-1}(x)\big) = s^{-1}(x) \\ &\iff s^{-1}(x) \notin \mathrm{Supp}(t) \\ &\iff x \notin s\big(\mathrm{Supp}(t)\big). \end{aligned}$$

Intéressons-nous maintenant aux orbites de l'action de $\langle s \rangle$ sur X. Ces orbites permettent de caractériser les *cycles* (*cf.* le module II.2 de [L1]) :

Lemme 18. Soient X un ensemble, $s \in \mathfrak{S}(X)$ et n un entier, $n \geqslant 2$. Pour que s soit un n-cycle, il faut, et il suffit, que l'une des orbites de X sous l'action de $\langle s \rangle$ soit de cardinal n et que toutes les autres orbites soient ponctuelles.

Démonstration.

Supposons que s soit un n-cycle : $s := (a_1\ a_2\ \cdots\ a_n)$, où $a_1, \ldots, a_n \in X$ sont distincts. Posons $S := \{a_1, \ldots, a_n\}$. La définition d'un cycle montre que S est stable sous $\langle s \rangle$. Si $x \in X \setminus S$, $s(x) = x$, d'où $s^k(x) = x$ pour tout $k \in \mathbb{Z}$. Ainsi x est un point fixe sous $\langle s \rangle$, *i.e.* l'orbite de x sous $\langle s \rangle$ est $\{x\}$. Ensuite, l'orbite de a_1 sous $\langle s \rangle$ est S. En effet, cette orbite est contenue dans S $\big($parce que S est stable sous $\langle s \rangle\big)$ et, pour tout $k \in [\![1, n]\!]$, $a_k = s^{k-1}(a_1)$ appartient à cette orbite. En conclusion, S est la seule orbite non ponctuelle de X sous $\langle s \rangle$.

Supposons inversement que, parmi les orbites de X sous $\langle s \rangle$, toutes soient ponctuelles sauf une, notons-la S, qui vérifie $\mathrm{card}\,(S) = n$. Soit $a \in S$, de sorte que $S = \langle s \rangle \cdot a$.

Considérons l'action de $(\mathbb{Z},+)$ sur S donnée par $(k,x)\mapsto s^k(x)$ (*cf.* l'exercice 1 de la page 61). Cette action est transitive (la seule orbite est S). D'après le théorème 14 de la page 72, le stabilisateur de a pour cette action est un sous-groupe d'indice n de $\mathbb{Z}$, donc est égal à $n\mathbb{Z}$. Ainsi :

$$\left\{k\in\mathbb{Z}\;\middle|\; s^k(a)=a\right\} \;=\; n\mathbb{Z}.$$

Pour $k=1,\dots,n$, posons alors $a_k := s^{k-1}(a)$. Si $1\leqslant j<k\leqslant n$, on a $s^{k-j}(a)\neq a$, *i.e.* $s^k(a)\neq s^j(a)$. Ainsi $a_1,\dots,a_n$ sont distincts. Par construction, $s(a_k)=a_{k+1}$ pour $k=1,\dots,n-1$ et, puisque $s^n(a)=a$, $s(a_n)=a=a_1$.
Il en résulte que $s=(a_1\;a_2\;\cdots\;a_n)$ est un n-cycle. ∎

Théorème 19 (Décomposition canonique d'une permutation).
Soient X un ensemble fini et s une permutation de X. On peut décomposer s sous la forme *canonique* suivante :

$$s \;=\; c_1\circ c_2\circ\cdots\circ c_r\,, \tag{10}$$

où $r\in\mathbb{N}$ et $c_1,\dots,c_r$ sont des cycles à supports deux à deux disjoints. Une telle décomposition de s est unique, à l'ordre des facteurs près.

Démonstration. La conclusion est évidente si $s=\mathrm{Id}_X$, en prenant $r:=0$. Supposons $s\neq\mathrm{Id}_X$. Notons $S_1,\dots,S_r$ les orbites non ponctuelles de X sous $\langle s\rangle$ (ainsi $r\geqslant 1$). Pour tout $i\in[\![1,r]\!]$, notons c_i l'application de X dans X définie ainsi :

$$c_i(x) := \begin{cases} x & \text{si } x\in X\setminus S_i \\ s(x) & \text{si } x\in S_i\,. \end{cases}$$

Puisque $s(S_i)=S_i$, c_i est une permutation de X. Pour tout $k\in\mathbb{Z}$, c_i^k coïncide avec s^k sur S_i et avec l'identité sur $X\setminus S_i$. On en déduit que S_i est la seule orbite non ponctuelle de X sous $\langle c_i\rangle$. D'après le lemme de la page précédente, c_i est un cycle, de longueur $n_i := \mathrm{card}\,(S_i)$.
Soit $s' := c_1\circ c_2\circ\cdots\circ c_r$, montrons que $s'=s$. En vertu de la proposition 17 de la page précédente, les c_i commutent deux à deux, puisque leurs supports S_i sont deux à deux disjoints. Soit $x\in X$. Si x n'appartient à aucune des S_i, il est laissé fixe par s et aussi par chaque c_i, donc $s(x)=x=s'(x)$. Dans le cas contraire, il existe un unique i tel que $x\in S_i$. Pour tout $j\in[\![1,r]\!]$ distinct de i, $c_j(x)=x$ et $c_j\big(c_i(x)\big)=c_i(x)$. Il en résulte que $s'(x)=c_i(x)=s(x)$. D'où l'existence d'une décomposition de s du type requis.
Supposons inversement que $s=c'_1\circ c'_2\circ\cdots\circ c'_q$, les c'_j étant des cycles à supports deux à deux disjoints. Pour tout j, soit $S'_j := \mathrm{Supp}(c'_j)$. Alors s coïncide avec c'_j sur S'_j ; on en déduit que S'_j est une orbite de X sous $\langle s\rangle$. Par ailleurs, $s(x)=x$ dès que $x\in X$ n'appartient à aucune des S'_j. Il en résulte que les S'_j sont les orbites non ponctuelles de X sous $\langle s\rangle$. Ainsi $q=r$ et, à l'ordre près, $S'_i=S_i$ pour tout i. Il est alors clair que $c'_i=c_i$ pour tout i (ces deux cycles coïncident sur S_i, qui est leur support commun). ∎

Exemple. Soit $s \in \mathfrak{S}_9$ la permutation donnée par la représentation matricielle :

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 2 & 8 & 9 & 1 & 6 & 3 & 7 & 4 \end{pmatrix}.$$

On a $s(1) = 5$ et $s(5) = 1$, donc $\{1, 5\}$ est une orbite de $[\![1, 9]\!]$ sous $\langle s \rangle$, et s coïncide sur cette orbite avec la transposition $(1\ 5)$. Ensuite $s(3) = 8$, $s(8) = 7$, $s(7) = 3$, donc $\{3, 7, 8\}$ est une seconde orbite, sur laquelle s coïncide avec le 3-cycle $(3\ 8\ 7)$. Enfin $s(4) = 9$ et $s(9) = 4$, donc $\{4, 9\}$ est une orbite de $[\![1, 9]\!]$ sous $\langle s \rangle$, et s coïncide sur cette orbite avec la transposition $(4\ 9)$. D'après la preuve du théorème 19 de la page ci-contre, la décomposition canonique de s est : $s = (1\ 5) \circ (4\ 9) \circ (3\ 8\ 7)$. Les points fixes de s sont 2 et 6. On écrit symboliquement $s = (1\ 5)(4\ 9)(3\ 8\ 7)(2)(6)$, ce qui précise les points fixes.

Exercice 11.

Soient X un ensemble de cardinal fini $n \geqslant 1$, $s \in \mathfrak{S}(X)$ et m le nombre d'orbites de X sous l'action de $\langle s \rangle$. Prouver la formule $\varepsilon(s) = (-1)^{n-m}$, $\varepsilon(s)$ étant la *signature* de s.

Solution. Décomposons s comme en (10). Si $i \in [\![1, r]\!]$, notons S_i le support de c_i et $n_i := \operatorname{card}(S_i)$. D'après la preuve du théorème 19 de la page précédente, les orbites sont d'une part les S_i et d'autre part les orbites ponctuelles $\{x\}$, où x décrit l'ensemble des points fixes de s, c'est-à-dire $X \setminus \bigcup_i S_i$. Ainsi :

$$m \;=\; r + \left[n - \sum_{i=1}^{r} n_i \right] \;=\; n - \sum_{i=1}^{r} (n_i - 1).$$

Pour tout i, nous savons que $\varepsilon(c_i) = (-1)^{n_i - 1}$. La conclusion résulte alors du fait que ε est un morphisme de groupes, de $\mathfrak{S}(X)$ dans $\big(\{-1, 1\}, \times\big)$.

Théorème 20. Soient X un ensemble fini et $s, s' \in \mathfrak{S}(X)$. Alors s et s' sont conjuguées dans $\mathfrak{S}(X)$ si, et seulement si, leurs décompositions canoniques respectives comportent le même nombre de cycles de chaque longueur.

Démonstration. Nous pouvons supposer s, s' distinctes de l'identité. Notons $\ell(c)$ la longueur d'un cycle c. Précisons d'abord ce que signifie l'énoncé. Il est possible d'écrire les décompositions canoniques de s, s' ainsi :

$$s \;=\; c_1 \circ c_2 \circ \cdots \circ c_r \quad \text{et} \quad s' \;=\; c'_1 \circ c'_2 \circ \cdots \circ c'_q\,,$$

en faisant les hypothèses suivantes. Les c_i sont des cycles à supports deux à deux disjoints, $\ell(c_1) \leqslant \ell(c_2) \leqslant \cdots \leqslant \ell(c_r)$, et de même pour les c'_j. Alors s et s' sont conjuguées si, et seulement si, $r = q$ et $\ell(c_i) = \ell(c'_i)$ pour tout i.

Supposons s et s' conjuguées : il existe $g \in \mathfrak{S}(X)$ tel que $s' = g \circ s \circ g^{-1}$. Pour tout i, posons $\gamma_i := g \circ c_i \circ g^{-1}$. Alors $s' = \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_r$. D'après [L1], chaque γ_i est un cycle

de longueur $\ell(c_i)$, et ces cycles sont à supports deux à deux disjoints (*cf.* l'exercice 10). De la partie unicité du théorème 19 de la page 80 et du fait que les suites $\big(\ell(c_1),\dots,\ell(c_r)\big)$ et $\big(\ell(c'_1),\dots,\ell(c'_q)\big)$ soient croissantes, on déduit que $r=q$ et $\ell(c_i)=\ell(c'_i)$ pour tout i.

Supposons inversement que $r=q$ et $\ell(c_i)=\ell(c'_i)$ pour tout i. Pour tout i, notons S_i le support de c_i, S'_i celui de c'_i et posons $n_i := \operatorname{card}(S_i) = \ell(c_i)$, de sorte que $n_i = \operatorname{card}(S'_i) = \ell(c'_i)$. Choisissons enfin un point $a_i \in S_i$ et un point $a'_i \in S'_i$. La preuve du théorème 19 de la page 80 montre que les n_i éléments de S_i sont les $s^h(a_i)$, $h=0,\dots,n_i-1$, de même pour S'_i. Par ailleurs, soit F (resp. F') l'ensemble des points fixes de s (resp. s'). Alors F est le complémentaire de la réunion des S_i (cette réunion est le support de s), de même pour F'. Il en résulte que $\operatorname{card}(F)=\operatorname{card}(F')$. Soit donc θ une bijection de F sur F'. Il existe une unique permutation g de X prolongeant θ et telle que, pour tout $i\in[\![1,r]\!]$ et tout $h\in[\![0,n_i-1]\!]$, $g\big(s^h(a_i)\big)=s'^h(a'_i)$. Alors $c'_i = g\circ c_i\circ g^{-1}$ pour tout i, vu la formule de conjugaison d'un cycle déjà invoquée, et donc $s'=g\circ s\circ g^{-1}$. ∎

Par exemple, si $k\in[\![2,\operatorname{card}(X)]\!]$, les k-cycles forment une classe de conjugaison de $\mathfrak{S}(X)$.

Exercice 12.

Décrire les classes de conjugaison de chacun des groupes $\mathfrak{S}_3$, $\mathfrak{S}_4$ et $\mathfrak{S}_5$.

Solution. Soient en général $n\in\mathbb{N}^*$ et $s\in\mathfrak{S}_n$, $s\neq 1$. Vu le théorème précédent, la classe de conjugaison de s est caractérisée par la suite croissante $(n_1,\dots,n_r)$ des longueurs des cycles composant s, comme en (10). Les contraintes sur les n_i sont d'une part $n_1\geqslant 2$, et d'autre part $n_1+n_2+\cdots+n_r\leqslant n$ (la somme est le cardinal du support de s).

Inversement, soit $(n_1,\dots,n_r)$ une suite croissante d'entiers obéissant à ces contraintes. On peut trouver des parties $S_1,\dots,S_r$ de $[\![1,n]\!]$ deux à deux disjointes, chaque S_i étant de cardinal n_i. Pour chaque i, on choisit un cycle c_i de support n_i. Alors $(n_1,\dots,n_r)$ est la suite associée à $s:=c_1\circ c_2\circ\cdots\circ c_r$. D'où une bijection $(n_1,\dots,n_r)\longmapsto\Gamma(n_1,\dots,n_r)$ de l'ensemble des suites croissantes obéissant aux deux contraintes mentionnées sur l'ensemble des classes de conjugaison de $\mathfrak{S}_n$ autres que $\{1\}$.

Si $n:=3$, il n'y a que deux suites possibles : (2) et (3) ; $\Gamma(2)$ est formée des transpositions et $\Gamma(3)$ est formée des 3-cycles. D'où trois classes de conjugaison, en comptant $\{1\}$.

Si $n:=4$, on obtient de même, pour $k=2,3,4$, la classe $\Gamma(k)$ formée des k-cycles. La seule autre suite est $(2,2)$, et $\Gamma(2,2)$ est formée des produits de deux transpositions à supports disjoints : $\Gamma(2,2)=\big\{(1\ 2)(3\ 4),(1\ 3)(2\ 4),(1\ 4)(2\ 3)\big\}$. Il y a ici cinq classes de conjugaison.

Si $n:=5$, on obtient de même sept classes de conjugaison, à savoir $\{1\}$, $\Gamma(2)$, $\Gamma(3)$, $\Gamma(4)$, $\Gamma(5)$, $\Gamma(2,2)$ et $\Gamma(2,3)$. Le lecteur pourra calculer les cardinaux de ces classes. On trouve dans l'ordre : $1,10,20,30,24,15,20$.

3.4 Applications aux groupes finis

Proposition 21. Soit G un groupe multiplicatif fini. Pour tout $a \in G$, posons $C_G(a) := \{g \in G \mid ga = ag\}$, et notons $Z(G)$ le centre de G. Si S est un ensemble de représentants des classes de conjugaison non ponctuelles de G, on a *l'équation aux classes* suivante :

$$\operatorname{card}(G) = \operatorname{card}\big(Z(G)\big) + \sum_{a \in S} \big(G : C_G(a)\big). \tag{11}$$

Démonstration. La classe de conjugaison d'un élément a de G est ponctuelle si, et seulement si, $a \in Z(G)$. Il en résulte que $T := S \cup Z(G)$ est un ensemble de représentants des classes de conjugaison de G (orbites de l'action par conjugaison de G sur lui-même). La proposition 8 de la page 65 donne donc :

$$\operatorname{card}(G) = \operatorname{card}\big(Z(G)\big) + \sum_{a \in S} \operatorname{card}(G \cdot a).$$

Soit $a \in S$. Le stabilisateur de a dans l'action par conjugaison de G sur lui-même est $C_G(a)$, donc le théorème 14 de la page 72 montre que $\operatorname{card}(G \cdot a) = \big(G : C_G(a)\big)$. La proposition en résulte. ∎

Voici une application simple de cette équation aux classes :

Proposition 22. Soient p un nombre premier et G un p-groupe non trivial. Le centre de G n'est pas trivial.

Démonstration. Par hypothèse, il existe $n \in \mathbb{N}^*$ tel que $\operatorname{card}(G) = p^n$. Gardons les notations de la proposition 21. Si $a \in S$, $C_G(a)$ est un sous-groupe de G distinct de G. La proposition 13 de la page 72 montre donc que l'indice $\big(G : C_G(a)\big)$ est un diviseur de $\operatorname{card}(G)$ distinct de 1, *i.e.* est de la forme p^r, où $r \in [\![1, n]\!]$. En particulier cet indice est un multiple de p.

Appliquons l'équation aux classes. Dans le second membre de l'équation (11), chaque terme de la somme est multiple de p, nous venons de le montrer. Comme $\operatorname{card}(G) = p^n$ est aussi multiple de p, il en résulte que $\operatorname{card}\big(Z(G)\big)$ est multiple de p. Mais $Z(G)$ n'est pas vide (c'est un sous-groupe de G), donc $\operatorname{card}\big(Z(G)\big) \geqslant 1$, et par suite $\operatorname{card}\big(Z(G)\big) \geqslant p$. Ainsi $Z(G)$ n'est pas trivial, *i.e.* n'est pas réduit à l'élément neutre. ∎

Exercice 13.

Soit p un nombre premier. Montrer que tout groupe d'ordre p^2 est commutatif.

Solution. Soit G un groupe (multiplicatif) d'ordre p^2. D'après la proposition précédente, le centre $Z(G)$ de G contient un élément $a \neq 1$. Vu le théorème de Lagrange, a est d'ordre p ou p^2. Si cet ordre est p^2, $G = \langle a \rangle \subset Z(G)$, donc $Z(G) = G$, *i.e.* G est abélien.

Supposons a d'ordre p, et soit $b \in G \setminus \langle a \rangle$. Le sous-groupe $\langle a, b \rangle$ de G engendré par a et b contient strictement $\langle a \rangle$, donc son ordre est, vu le théorème de Lagrange, un multiple

de p distinct de p et divisant p^2. D'où $\langle a, b\rangle = G$. Considérons ensuite le groupe $C_G(b)$ (comme dans la proposition 21 de la page précédente). Ce sous-groupe de G contient b, mais aussi a, parce que $a \in Z(G)$. Ainsi $C_G(b) \supset \langle a, b\rangle = G$, *i.e.* $C_G(b) = G$, ou encore $b \in Z(G)$. Dans ces conditions $Z(G) = G$, puisque $a, b \in Z(G)$. D'où la conclusion.

Proposition 23. Soit H un sous-groupe du groupe multiplicatif G. Pour toute classe à gauche C de G modulo H et tout $g \in G$, gC est encore une classe à gauche modulo H. On obtient ainsi une action par translations à gauche de G sur G/H, et cette action est *transitive*.

Démonstration. Soit $a \in G$ tel que $C = aH$. Alors $gC = g(aH) = (ga)H$ est la classe à gauche de ga modulo H, d'où la première assertion. Pour tous $g \in G$ et $C \in G/H$, posons ensuite $g \cdot C = gC$. Vu ce qui précède, $gC \in G/H$, d'où une loi externe $G \times (G/H) \to G/H$. C'est évidemment une action de groupe : si par exemple $C \in G/H$ et $g, g' \in G$, il vient :

$$g \cdot (g' \cdot C) = g \cdot (g'C) = g(g'C) = (gg')C = (gg') \cdot C.$$

Puisque $H = 1H$ est la classe à gauche de 1 modulo H, H est un élément de G/H. Soient $C \in G/H$ et $a \in G$ tel que $C = aH$. Alors $C = a \cdot H$ appartient à l'orbite de H sous G (pour l'action de G sur G/H considérée), donc cette orbite est égale à G/H. Ainsi, l'action de G sur G/H est transitive. ∎

Soit H un sous-groupe d'indice 2 d'un groupe (multiplicatif) G. Alors H est *distingué* dans G. En effet il y a deux classes à gauche (resp. à droite) de G modulo H ; l'une d'elles étant H, l'autre est nécessairement le complémentaire $G \setminus H$. Pour tout $g \in G$, on en déduit que $gH = Hg$ (distinguer les cas $g \in H$ et $g \notin H$), autrement dit $gHg^{-1} = H$. Dans le cas où G est fini, ce résultat est aussi une conséquence de l'exercice suivant.

Exercice 14.

Soient G un groupe fini d'ordre N et H un sous-groupe de G dont l'indice $p := (G : H)$ est le plus petit facteur premier de N. Montrer que H est un sous-groupe *distingué* de G.

Solution. Soit $\varphi : G \to \mathfrak{S}(G/H)$ le morphisme associé à l'action par translations à gauche de G sur G/H. Étudions l'image L et le noyau K de φ. D'abord K est un sous-groupe distingué de G. Ensuite L est isomorphe à G/K (*cf.* le module I.1), donc $N = \operatorname{card}(L)\operatorname{card}(K)$ est multiple de $\operatorname{card}(L)$. Or L est un sous-groupe de $\mathfrak{S}(G/H)$, donc $\operatorname{card}(L)$ divise l'ordre de $\mathfrak{S}(G/H)$, à savoir $p!$. Ainsi, $\operatorname{card}(L)$ divise N et $p!$.

Or, p étant le plus petit facteur premier de N, le pgcd de N et $p!$ est p. Écrivons en effet $N = pm$. Il suffit de montrer que m et $(p-1)!$ sont premiers entre eux. Les entiers $1, 2, \ldots, p-1$ sont premiers avec m (parce que p est le plus petit facteur premier de N), leur produit $(p-1)!$ est donc premier avec m. Ainsi, $\operatorname{card}(L)$ divise $p = N \wedge p!$. L'action de G sur G/H est transitive, elle ne peut être triviale (sinon l'ensemble G/H serait réduit à un élément), autrement dit $L \neq \{1\}$. Ainsi $\operatorname{card}(L) = p$, et par conséquent $\operatorname{card}(K) = N/p = \operatorname{card}(H)$. Soit $g \in K$: on a donc $gC = C$ pour toute

classe $C \in G/H$. Ainsi $gH = H$ (prendre $C := H$), ce qui signifie que $g \in H$. Nous avons ainsi prouvé l'inclusion $K \subset H$. D'où $K = H$, puisque $\mathrm{card}\,(K) = \mathrm{card}\,(H)$. Finalement H est un sous-groupe distingué de G parce que K en est un (c'est le noyau de φ).

Comme autre application, calculons les ordres des groupes linéaires sur un corps fini :

Proposition 24. Soit K un corps fini de cardinal q. Pour tout entier $n \geqslant 1$, on a :

$$\mathrm{card}\left(GL_n(K)\right) \;=\; q^{n(n-1)/2} \prod_{k=1}^{n} (q^k - 1). \tag{12}$$

Démonstration. Raisonnons par récurrence sur n.

Le cas $n = 1$ est évident : $GL_1(K) = K^*$ est bien d'ordre $q - 1$. Supposons $n \geqslant 2$, la formule (12) étant vraie pour $n - 1$. Identifions $GL_n(K)$ à $\mathcal{GL}(K^n)$, au moyen de la base canonique $(e_1, \ldots, e_n)$ de K^n.

Considérons l'action naturelle de $GL_n(K)$ sur $K^n \setminus \{0\}$. L'exercice 3 de la page 66 montre que cette action est transitive. L'orbite de e_n sous $GL_n(K)$ est donc $K^n \setminus \{0\}$, elle est de cardinal $q^n - 1$. Le théorème 14 de la page 72 donne alors :

$$\mathrm{card}\left(GL_n(K)\right) \;=\; (q^n - 1)\,\mathrm{card}\left(GL_n(K)_{e_n}\right), \tag{$*$}$$

où $GL_n(K)_{e_n}$ est le stabilisateur de e_n. Calculons l'ordre de $GL_n(K)_{e_n}$.

Soit $M := (a_{i,j}) \in M_n(K)$. Notons $M' \in M_{n-1}(K)$ la matrice déduite de M par suppression de la dernière ligne et de la dernière colonne. D'abord $M(e_n) = e_n$ si, et seulement si, M est de la forme :

$$M \;=\; \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n-1} & 0 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n-1} & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n-1,1} & a_{n-1,2} & \cdots & a_{n-1,n-1} & 0 \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n-1} & 1 \end{pmatrix}.$$

Dans ce cas, $\det(M) = \det(M')$ (développement par rapport à la dernière colonne). Ainsi, lorsque $M(e_n) = e_n$, M appartient à $GL_n(K)$ si, et seulement si, M' appartient à $GL_{n-1}(K)$.On en déduit aisément que l'application $M \longmapsto \left(M', (a_{n,1}, \ldots, a_{n,n-1})\right)$ est une bijection de $GL_n(K)_{e_n}$ sur l'ensemble produit $GL_{n-1}(K) \times K^{n-1}$ (mais ce n'est pas un isomorphisme de groupes). Vu l'hypothèse de récurrence, il vient :

$$\mathrm{card}\left(GL_n(K)_{e_n}\right) = q^{n-1}\left[q^{(n-1)(n-2)/2} \prod_{k=1}^{n-1} (q^k - 1)\right] = q^{n(n-1)/2} \prod_{k=1}^{n-1} (q^k - 1).$$

D'où la conclusion, compte tenu de la formule $(*)$. ∎

Exercice 15.
En déduire l'ordre du groupe $SL_n(K)$.
Solution. Le morphisme $\det : GL_n(K) \to K^*$ est surjectif, et son noyau est $SL_n(K)$. D'où un isomorphisme du quotient $GL_n(K)/SL_n(K)$ sur K^*. D'après la proposition 13 de la page 72, l'ordre de $SL_n(K)$ est le quotient de l'ordre de $GL_n(K)$ par $\mathrm{card}\,(K^*) = q-1$. D'où :

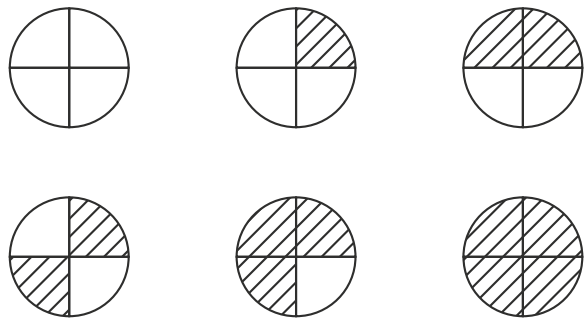
$$\mathrm{card}\,\big(SL_n(K)\big) \;=\; q^{n(n-1)/2}\prod_{k=2}^{n}(q^k-1), \quad \text{par exemple} \quad \mathrm{card}\,\big(SL_2(K)\big) \;=\; q(q^2-1).$$

3.5 Une application à la combinatoire

Voici un exemple de problème combinatoire dans lequel intervient une action de groupe. Considérons une roulette formée de N secteurs circulaires. De combien de façons peut-on la colorier, si l'on dispose de q couleurs, et si l'on ne distingue pas deux coloriages se déduisant l'un de l'autre par rotation ? Notons $T(N,q)$ le nombre cherché. Nous allons établir la formule :

$$T(N,q) \;=\; \frac{1}{N}\sum_{d|N}\varphi(d)q^{N/d}, \tag{13}$$

dans laquelle d décrit l'ensemble des diviseurs positifs de N, et φ est la fonction indicatrice d'Euler (*cf.* [L1]). Il est clair que $T(N,1) = 1$; nous retrouvons ainsi la formule $\sum_{d|N}\varphi(d) = N$. La figure ci-dessous illustre l'égalité $T(4,2) = 6$.



Formulons notre problème en termes d'actions. Numérotons les secteurs de 1 à N, dans le sens direct, et les couleurs de 1 à q. Un coloriage de la roulette peut être interprété comme une application f de $\mathbb{Z}/N\mathbb{Z}$ dans $[\![1,q]\!]$: à tout secteur d'indice $k \in [\![1,N]\!]$ on attribue la couleur d'indice $f(\bar{k})$, $\bar{k}$ étant la classe de k modulo N. Deux coloriages, correspondant aux applications $f, f' : \mathbb{Z}/N\mathbb{Z} \to [\![1,q]\!]$, se déduisent l'un de l'autre par rotation si, et seulement s'il existe $h \in \mathbb{Z}$ tel que $f'(\bar{k}) = f(\bar{k}-\bar{h})$ pour tout $k \in \mathbb{Z}$.
Considérons alors le groupe *additif* $\mathbb{Z}/N\mathbb{Z}$. Faisons-le agir par translations à gauche sur lui-même et de façon triviale sur $[\![1,q]\!]$. D'après l'exemple 10 de la page 61, on en déduit une action de $\mathbb{Z}/N\mathbb{Z}$ sur $\mathcal{F}\big(\mathbb{Z}/N\mathbb{Z}, [\![1,q]\!]\big)$, ensemble des applications de $\mathbb{Z}/N\mathbb{Z}$ dans $[\![1,q]\!]$.

Pour tous $g, x \in \mathbb{Z}/N\mathbb{Z}$ et $f \in \mathcal{F}\big(\mathbb{Z}/N\mathbb{Z}, [\![1,q]\!]\big)$, on a par définition $(g\cdot f)(x) := f(x-g)$. Compte tenu de l'alinéa précédent, $T(N,q)$ est simplement le nombre d'orbites de cette action de $\mathbb{Z}/N\mathbb{Z}$ sur $\mathcal{F}\big(\mathbb{Z}/N\mathbb{Z}, [\![1,q]\!]\big)$. Or, il se trouve qu'un théorème général donne le nombre d'orbites d'une action de groupes :

Théorème 25 (Frobenius). Soit G un groupe fini d'ordre N agissant sur un ensemble fini X. Pour tout $g \in G$, notons $\Phi(g)$ le nombre de points fixes de g (dans X). Alors le nombre d'orbites de X sous G est la moyenne arithmétique des nombres $\Phi(g)$, g décrivant G. Autrement dit, on a la formule suivante, appelée aussi *formule de Burnside* :

$$\operatorname{card}(X/G) \;=\; \frac{1}{N}\sum_{g\in G}\Phi(g).$$

Démonstration. Posons $m := \operatorname{card}(X/G)$, et soit $E := \big\{(g,x) \in G\times X \;\big|\; g\cdot x = x\big\}$. La formule de Fubini (*cf.* le module II.1 de [L1]) permet de calculer $\operatorname{card}(E)$ de deux façons. Si $g \in G$, le nombre de points $x \in X$ tels que $(g,x) \in E$ est $\Phi(g)$. Si $x \in X$, le nombre de $g \in G$ tels que $(g,x) \in E$ est l'ordre de G_x. Ainsi :

$$\sum_{g\in G}\Phi(g) \;=\; \operatorname{card}(E) \;=\; \sum_{x\in X}\operatorname{card}(G_x).$$

Les orbites $\omega_1, \ldots, \omega_m$ de X sous G forment une partition de X, d'où :

$$\sum_{x\in X}\operatorname{card}(G_x) \;=\; \sum_{i=1}^{m}\left(\sum_{x\in\omega_i}\operatorname{card}(G_x)\right).$$

Si $i \in [\![1,m]\!]$ et $x \in \omega_i$, on a $\operatorname{card}(G_x) = N/\operatorname{card}(\omega_i)$ (théorème 14 de la page 72). Ainsi :

$$\sum_{g\in G}\Phi(g) \;=\; \sum_{i=1}^{m}\left(\sum_{x\in\omega_i} N/\operatorname{card}(\omega_i)\right) \;=\; N\sum_{i=1}^{m}1 \;=\; Nm. \qquad \blacksquare$$

Revenons à notre roulette. Pour calculer $T(N,q)$ il suffit, en vertu du théorème de Frobenius, de calculer, pour tout $g \in \mathbb{Z}/N\mathbb{Z}$, le nombre $\Phi(g)$ de points fixes de g dans $\mathcal{F}\big(\mathbb{Z}/N\mathbb{Z}, [\![1,q]\!]\big)$. Soit donc $g \in \mathbb{Z}/N\mathbb{Z}$, notons d l'ordre de g dans le groupe $(\mathbb{Z}/N\mathbb{Z}, +)$. D'abord d divise N (théorème de Lagrange), écrivons $N := dt$. On peut écrire $g := \bar{k}$ pour un unique $k \in [\![1,N]\!]$, puis $k := tk'$, où $k' \in [\![1,d]\!]$ est premier avec d, et alors $\langle g\rangle = \langle \bar{t}\rangle$ (*cf.* un exercice du module II.2 de [L1]), d'où aussitôt $\Phi(g) = \Phi(\bar{t})$.

Soit maintenant $f \in \mathcal{F}\big(\mathbb{Z}/N\mathbb{Z}, [\![1,q]\!]\big)$. Pour que f soit fixe sous $\bar{t}$, il faut, et il suffit, que $f(x-\bar{t}) = f(x)$ pour tout $x \in \mathbb{Z}/N\mathbb{Z}$. Cela revient à dire que l'application $a \mapsto f(\bar{a})$ de $\mathbb{Z}$ dans $[\![1,q]\!]$, qui est *a priori* N-périodique, est en fait t-périodique. On en déduit une bijection entre l'ensemble des points fixes de $\bar{t}$ et l'ensemble V_t des applications t-périodiques de $\mathbb{Z}$ dans $[\![1,q]\!]$. Il est clair par ailleurs que $u \longmapsto \big(u(0), u(1), \ldots, u(t-1)\big)$ est une bijection de V_t sur $[\![1,q]\!]^t$, et ainsi $\Phi(g) = \Phi(\bar{t}) = q^t = q^{N/d}$.

En vertu de l'exercice précité, si d et t sont comme ci-dessus ($N = dt$), $k' \mapsto \overline{tk'}$ est une bijection de l'ensemble des entiers $k' \in [\![1, d]\!]$ premiers avec d sur l'ensemble des éléments d'ordre d de $\mathbb{Z}/N\mathbb{Z}$. Par définition de la fonction φ, $\mathbb{Z}/N\mathbb{Z}$ possède donc $\varphi(d)$ éléments d'ordre d. La formule (13) s'en déduit aussitôt.

Exercice type corrigé

I.2.0 Soit E un espace vectoriel de dimension finie $n \geqslant 1$ sur un corps K. On appelle *drapeau* de E toute suite $(V_0, \dots, V_n)$ de sous-espaces vectoriels de E, croissante pour l'inclusion, telle que $\dim(V_i) = i$ pour $i = 0, \dots, n$. Soit $\mathcal{D}$ l'ensemble des drapeaux de E.

1. Définir une action naturelle de $\mathcal{GL}(E)$ sur $\mathcal{D}$, et montrer que cette action est transitive.

2. Montrer que le stabilisateur d'un drapeau est isomorphe au sous-groupe de $GL_n(K)$ formé des matrices triangulaires supérieures inversibles.

Solution.

1. Soient $d = (V_0, \dots, V_n)$ un drapeau de E et $g \in \mathcal{GL}(E)$. Pour tout $i \in [\![0, n]\!]$, $g(V_i)$ est un sous-espace vectoriel de dimension i de E, car g est un automorphisme de E. De plus, si $i < n$, $V_i \subset V_{i+1}$, donc $g(V_i) \subset g(V_{i+1})$. Il en résulte que $\big(g(V_0), \dots, g(V_n)\big)$ est un drapeau, que l'on notera $g \cdot d$. On obtient ainsi une action de $\mathcal{GL}(E)$ sur $\mathcal{D}$. En fait $\mathcal{GL}(E)$ agit naturellement sur l'ensemble $\mathcal{V}$ des sous-espaces vectoriels de E, il agit donc diagonalement sur $\mathcal{V}^{n+1}$, et $\mathcal{D}$ est une partie de $\mathcal{V}^{n+1}$ stable sous $\mathcal{GL}(E)$. D'où une action induite de $\mathcal{GL}(E)$ sur $\mathcal{D}$, c'est l'action définie ci-dessus.

Montrons que l'action de $\mathcal{GL}(E)$ sur $\mathcal{D}$ est transitive. Pour cela, considérons deux drapeaux $d = (V_0, \dots, V_n)$ et $d' = (V'_0, \dots, V'_n)$. Il existe une base $\mathcal{B} := (e_1, \dots, e_n)$ de E *adaptée* à d en ce sens que, pour tout $k \in [\![1, n]\!]$, $V_k = \operatorname{Vect}(e_1, \dots, e_k)$. On construit une telle base par récurrence, à l'aide du théorème de la base incomplète : si $(e_1, \dots, e_k)$ est une base de V_k et si $k < n$, il existe $e_{k+1} \in V_{k+1}$ tel que $(e_1, \dots, e_{k+1})$ soit une base de V_{k+1}. Il existe de même une base $\mathcal{B}' := (e'_1, \dots, e'_n)$ de E adaptée à d'. Soit alors g le seul automorphisme de E tel que $g(e_k) := e'_k$ pour $k = 1, \dots, n$. Pour tout $k \in [\![1, n]\!]$, $g(V_k) = V'_k$, car $(e_1, \dots, e_k)$ est une base de V_k, donc $\big(g(e_1), \dots, g(e_k)\big) = (e'_1, \dots, e'_k)$ est une base de $g(V_k)$, et aussi une base de V'_k. Ainsi $g \cdot d = d'$, donc d, d' appartiennent à une même orbite sous $\mathcal{GL}(E)$.

L'action de $\mathcal{GL}(E)$ sur $\mathcal{D}$ est donc transitive.

2. Soit $d = (V_0, \dots, V_n)$ un drapeau de E. Considérons, comme en 1, une base $\mathcal{B} := (e_1, \dots, e_n)$ de E adaptée à d. En associant à tout $g \in \mathcal{GL}(E)$ sa matrice dans la base $\mathcal{B}$, nous savons que nous obtenons un isomorphisme ψ de $\mathcal{GL}(E)$ sur $GL_n(K)$. Soient $g \in \mathcal{GL}(E)$ et $A := \operatorname{Mat}_{\mathcal{B}}(g) := (a_{i,j})$. L'égalité $g \cdot d = d$ est vraie si, et seulement si, $g(V_k) = V_k$, ce qui équivaut (parce que g est un automorphisme) à $g(V_k) \subset V_k$ pour $k = 1, \dots, n$. Cela revient à dire que $g(e_k) \in V_k$ pour tout $k \in [\![1, n]\!]$. Soit $k \in [\![1, n]\!]$. Comme $g(e_k) = \sum_{i=1}^{n} a_{i,k} e_i$, $g(e_k) \in V_k$ si, et seulement si, $a_{i,k} = 0$ pour tout $i \in [\![k+1, n]\!]$. Il en résulte que $g \cdot d = d$ si, et seulement si, A est triangulaire supérieure.

Ainsi, la restriction de ψ au stabilisateur $\mathcal{GL}(E)_d$ de d est un isomorphisme de $\mathcal{GL}(E)_d$ sur le sous-groupe $T_n(K) \cap GL_n(K)$ de $GL_n(K)$ formé des matrices triangulaires supérieures inversibles, *i.e.* à termes diagonaux non nuls.

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 les sections 1 et 2

I.2.1 Soient A un anneau commutatif, G le groupe multiplicatif des éléments inversibles de A et n un entier fixé. Montrer que $(g, a) \longmapsto g^n a$ est une action de G sur A.

I.2.2 Un groupe G agit sur un ensemble X. L'action diagonale de G sur X^2 est-elle transitive ?

I.2.3 Soient E un espace vectoriel de dimension finie $n \geqslant 2$ sur un corps et $k \in [\![1, n-1]\!]$. Montrer que l'action naturelle de $\mathcal{GL}(E)$ sur l'ensemble des sous-espaces vectoriels de dimension k de E est transitive. Quel est le noyau de cette action ?

I.2.4 Soient X un espace affine de dimension finie n sur un corps et $k \in [\![0, n-1]\!]$. Montrer que l'action naturelle de $\mathcal{GA}(X)$ sur l'ensemble des sous-espaces affines de dimension k de X est transitive et fidèle.

I.2.5

1. Soit $\mathcal{H} := \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$ le demi-plan de Poincaré. Si $g := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{R})$ et $z \in \mathcal{H}$, on pose $g \cdot z := \dfrac{az+b}{cz+d}$. Vérifier la formule suivante :

$$\operatorname{Im}(g \cdot z) = \frac{\operatorname{Im} z}{|cz+d|^2}.$$

En déduire que $(g, z) \mapsto g \cdot z$ est une action de $SL_2(\mathbb{R})$ sur $\mathcal{H}$. Montrer que cette action est transitive. Montrer que le stabilisateur de i est $SO_2(\mathbb{R})$ (*cf.* le module ? ?). Que dire du stabilisateur d'un autre point de $\mathcal{H}$?

2. Soit $\mathcal{D} := \{Z \in \mathbb{C} \mid |Z| < 1\}$ le disque unité. L'application $f : z \mapsto (z-i)/(z+i)$ est une bijection de $\mathcal{H}$ sur $\mathcal{D}$ (*cf.* [L1], module II.4). En déduire une action $(g, Z) \mapsto g \cdot Z$ de $SL_2(\mathbb{R})$ sur $\mathcal{D}$, *i.e.* un morphisme $\psi : SL_2(\mathbb{R}) \to \mathfrak{S}(\mathcal{D})$. Soit s une permutation de $\mathcal{D}$. Montrer que s appartient à l'image de ψ si, et seulement s'il existe $u, a \in \mathbb{C}$ vérifiant $|u| = 1$, $|a| < 1$ et tels que :

$$s(Z) = u\,\frac{Z+a}{1+\bar{a}Z} \quad \text{pour tout} \quad Z \in \mathcal{D}.$$

I.2.6 Soient X un espace affine réel et G un sous-groupe fini du groupe affine $\mathcal{GA}(X)$. Soit par ailleurs Y une partie *convexe* non vide de X stable sous G (pour la définition de « convexe », voir le module II.1). Montrer qu'il existe un point de Y fixe sous G.

I.2.7 Soit H un sous-groupe du groupe multiplicatif G. On appelle *normalisateur* de H dans G le stabilisateur $N_G(H)$ de H dans l'action par conjugaison de G sur l'ensemble de ses sous-groupes.

1. Montrer que $N_G(H)$ est, pour l'inclusion, le plus grand sous-groupe de G dont H soit un sous-groupe distingué.

2. Soit K un sous-groupe de $N_G(H)$. Montrer que HK est un sous-groupe de G. On rappelle que $HK := \{hk \mid (h,k) \in H \times K\}$.

I.2.8 Pour tous $P \in \mathbb{C}[X]$ et $a \in \mathbb{C}^*$, on pose $a \cdot P := P(aX)$. Montrer qu'on définit ainsi une action du groupe $(\mathbb{C}^*, \times)$ sur $\mathbb{C}[X]$. Soit $a \in \mathbb{C}^*$, posons $A := \{P \in \mathbb{C}[X] \mid a \cdot P = P\}$. Montrer que A est une sous-algèbre de $\mathbb{C}[X]$, et décrire cette sous-algèbre.

I.2.9 Soit $k \in \mathbb{N}^*$. Une action d'un groupe G sur un ensemble X est dite k *fois transitive* si l'action de G sur l'ensemble des k-uplets *injectifs* $(x_1, \dots, x_k) \in X^k$ (les x_i sont deux à deux distincts) est transitive.

1. On suppose $k \geqslant 2$. Montrer qu'une action de G sur X est k fois transitive si, et seulement si, elle est transitive et s'il existe $a \in X$ tel que l'action du stabilisateur G_a sur $X \setminus \{a\}$ soit $k-1$ fois transitive.

2. Soit X un espace affine de dimension finie $n \geqslant 1$ sur un corps. Montrer que l'action naturelle de $\mathcal{GA}(X)$ sur X est deux fois transitive.

3. Soit E un espace vectoriel de dimension finie $n \geqslant 2$ sur un corps. Montrer que l'action naturelle de $\mathcal{GL}(E)$ sur l'ensemble des droites vectorielles de E est deux fois transitive, et qu'elle est trois fois transitive si, et seulement si, $n = 2$.

4. Soit $n \in \mathbb{N}^*$. Montrer que l'action naturelle de $\mathfrak{S}_n$ sur $[\![1, n]\!]$ est n fois transitive et que, si $n \geqslant 3$, l'action naturelle du groupe alterné $\mathfrak{A}_n$ sur $[\![1, n]\!]$ est $n-2$ fois transitive.

I.2.10 Considérons une action transitive d'un groupe G sur un ensemble X ayant au moins deux éléments, et soit $a \in X$.

1. Montrer que les conditions suivantes sont équivalentes :
a) Tout sous-groupe de G contenant G_a est égal à G_a ou à G.
b) Les singletons et X sont les seules parties non vides Y de X vérifiant :

$$\big(\forall g \in G\big),\ \big(g \cdot Y \cap Y \neq \varnothing\big) \Rightarrow \big(g \cdot Y = Y\big).$$

Si ces conditions sont satisfaites, l'action donnée est dite *primitive*.

2. On suppose l'action donnée primitive. Soit K un sous-groupe distingué de G. Montrer que l'action de K sur X est soit triviale, soit transitive.

3. Montrer qu'une action deux fois transitive (exercice I.2.9) est primitive.

I.2.11 Soit G le sous-groupe de $\mathcal{GA}(\mathbb{R}^2)$ formé des transformations affines f telles que $\vec{f} \in \mathcal{O}(\mathbb{R}^2)$. Les éléments de G sont appelés *isométries affines* de $\mathbb{R}^2$. Montrer que l'action naturelle de G sur l'ensemble des droites affines de $\mathbb{R}^2$ est transitive. Décrire les orbites de l'action naturelle de G sur les paires $\{D, D'\}$ de droites *affines* de $\mathbb{R}^2$.

I.2.12 Soit X un espace affine de direction E et de dimension finie n sur un corps K.

1. Pour une famille $(A_0, \dots, A_n)$ de $n+1$ points de X, montrer que les conditions suivantes sont équivalentes :

a) Le seul sous-espace affine de X contenant $A_0, \dots, A_n$ est X.

b) Pour tout $M \in X$, il existe $(\lambda_0, \dots, \lambda_n) \in K^{n+1}$ unique tel que $\sum_{i=0}^{n} \lambda_i = 1$ et que M soit le barycentre de $\big((A_0, \lambda_0), \dots, (A_n, \lambda_n)\big)$.

c) La famille $\big(\overrightarrow{A_0A_1}, \dots, \overrightarrow{A_0A_n}\big)$ est une base de E.

Si ces conditions sont vérifiées, $(A_0, \dots, A_n)$ est appelée *base affine* de X.

2. Montrer que le groupe affine $\mathcal{GA}(X)$ agit naturellement sur l'ensemble des bases affines de X, et que cette action est simplement transitive. En déduire que, si $(A_0, \dots, A_n)$ est une base affine de X, le stabilisateur de $\{A_0, \dots, A_n\}$ dans $\mathcal{GA}(X)$ est isomorphe à $\mathfrak{S}_{n+1}$.

I.2.13 Soit X un espace affine de dimension finie $n \geqslant 1$ sur un corps K. Montrer que le centre du groupe affine $\mathcal{GA}(X)$ est trivial, sauf si $n = 1$ et si K est un corps à deux éléments.

I.2.14 Soit $K = \mathbb{Z}/2\mathbb{Z}$. Montrer que le groupe affine $\mathcal{GA}(K^2)$ est isomorphe à $\mathfrak{S}_4$. Retrouver ainsi le fait que $\mathfrak{S}_4$ possède un sous-groupe distingué isomorphe à $\big((\mathbb{Z}/2\mathbb{Z})^2, +\big)$.

I.2.15 Montrer que tout groupe fini est isomorphe à un sous-groupe du groupe alterné $\mathfrak{A}_n$, pour un certain entier $n \geqslant 2$.

Exercices sur la section 3

I.2.16 Soit H un sous-groupe du groupe G. Montrer que le noyau K de l'action par translations à gauche de G sur G/H est l'intersection des conjugués de H dans G. Montrer que K est, au sens de l'inclusion, le plus grand sous-groupe distingué de G inclus dans H.

I.2.17 Soit G un groupe possédant une partie génératrice finie. À l'aide de l'exercice précédent montrer que, pour tout entier n, l'ensemble des sous-groupes d'indice n de G est fini.

I.2.18 Soient X un ensemble fini et $s \in \mathfrak{S}(X)$. Démontrer que l'ordre de s est le ppcm des longueurs des cycles figurant dans la décomposition canonique de s.

I.2.19 Écrire un algorithme donnant la décomposition canonique d'un élément de $\mathfrak{S}_n$ (décomposition comme produit de cycles à supports disjoints).

I.2.20

Construire un sous-groupe d'ordre 81 de $\mathfrak{S}_9$. On pourra utiliser l'exercice I.2.7.

I.2.21 Soit G un groupe fini d'ordre $2n$, où n est un entier *impair*. Notons $\varphi : G \to \mathfrak{S}(G)$ le morphisme associé à l'action de G sur lui-même par translations à gauche.

1. Soit c un élément d'ordre 2 de G (justifier l'existence de c). Montrer que $\varphi(c)$ est produit de n transpositions à supports deux à deux disjoints.

2. Soit $\varepsilon : \mathfrak{S}(G) \to \{-1, 1\}$ le morphisme de signature. Montrer que le noyau de $\varepsilon \circ \varphi$ est un sous-groupe distingué d'indice 2 de G.

I.2.22 Soient k un entier et X un ensemble fini de cardinal N. On suppose $1 < k < N$. Le groupe $\mathfrak{S}(X)$ agit naturellement sur l'ensemble E des parties à k éléments de X, d'où un morphisme $\varphi : \mathfrak{S}(X) \to \mathfrak{S}(E)$.

1. Montrer que l'action de $\mathfrak{S}(X)$ sur E est fidèle et transitive. Soit $A \in E$. Montrer que le stabilisateur $\mathfrak{S}(X)_A$ est isomorphe au groupe $\mathfrak{S}(A) \times \mathfrak{S}(X \setminus A)$.

2. Décrire les orbites de l'action naturelle de $\mathfrak{S}(X)$ sur l'ensemble des paires d'éléments de E.

3. À quelle condition, portant sur k et N, l'image de φ est-elle contenue dans le groupe alterné $\mathfrak{A}(E)$?

I.2.23 Soit $n \in \mathbb{N}^*$. Quelles sont les classes de conjugaison de $\mathfrak{S}_n$ formées d'éléments d'ordre 2 ? Calculer le cardinal de chacune de ces classes. En déduire que, si $n \neq 6$ et si α est un automorphisme de $\mathfrak{S}_n$, l'image par α de toute transposition est une transposition (on peut montrer que c'est faux si $n = 6$).

I.2.24 Soient n un entier strictement positif et $\mathcal{F}$ l'ensemble des applications $f : \mathbb{N} \to \mathbb{N}$ vérifiant la condition suivante :

$$f^2(x) := (f \circ f)(x) = x + n \quad \text{pour tout} \quad x \in \mathbb{N}. \qquad (*)$$

1. Soit $f \in \mathcal{F}$. Montrer que $f(x + n) = f(x) + n$ pour tout $x \in \mathbb{N}$. En déduire qu'il existe une unique permutation $s \in \mathfrak{S}_n$ telle que $f(x) \equiv s(x) \mod n$ pour tout $x \in [\![1, n]\!]$. Montrer que $s^2 = \mathrm{Id}$.

2. En étudiant l'application $f \mapsto s$, avec les notations de 1, montrer que $\mathcal{F}$ est vide si n est impair et que, si $n = 2m$ est pair, $\mathrm{card}\,(\mathcal{F}) = (2m)!/m!$.

I.2.25 Soit X un ensemble fini de cardinal $n \geqslant 3$. Montrer que n est premier si, et seulement si, pour tout cycle c de longueur n et toute transposition s de X, $\{c, s\}$ est une partie génératrice de $\mathfrak{S}(X)$.

I.2.26 Soient $n \in \mathbb{N}^*$ et S une partie génératrice de $\mathfrak{S}_n$ formée de transpositions, posons $r := \mathrm{card}\,(S)$. On veut démontrer l'inégalité $r \geqslant n - 1$. On pourra raisonner par l'absurde, en supposant $r \leqslant n - 2$, et en construisant une suite $(s_1, \ldots, s_r)$ d'éléments de S deux à deux distincts telle que, pour $k = 1, \ldots, r$, le support de $s_1 \circ s_2 \circ \cdots \circ s_k$ soit de cardinal $k + 1$.

I.2.27 Les hypothèses et notations étant celles du théorème de Frobenius, on suppose que l'action de G sur X est *transitive*. Soient $a \in X$ et $H := G_a$ son stabilisateur. Montrer que le nombre μ d'orbites de X sous l'action de H est donné par la formule suivante :

$$\mu = \frac{1}{N} \sum_{g \in G} \Phi(g)^2.$$

En posant $F := \{(g, s) \in G^2 \mid gsg^{-1} \in H\}$, on calculera de deux façons différentes la somme $\sum_{(g,s) \in F} \Phi(s)$.

I.2.28 Soit $K := \{-1, 1\}^n \subset \mathbb{R}^n$, où n est un entier, $n \geqslant 2$. Les 2^n points de K sont les sommets d'un « hypercube » de $\mathbb{R}^n$. Soit $G \subset \mathcal{GA}(\mathbb{R}^n)$ le groupe des symétries de K. Montrer que G est d'ordre $2^n n!$, et préciser la structure du groupe G (s'inspirer du cas $n = 3$).

I.2.29 Soit K le corps $\mathbb{Z}/2\mathbb{Z}$. Construire un isomorphisme entre $GL_2(K)$ et $\mathfrak{S}_3$.

I.2.30 Soient $p \geqslant 3$ un nombre premier et G un groupe multiplicatif d'ordre $p+1$, possédant un automorphisme σ d'ordre p. En considérant l'action de σ sur $G \setminus \{1\}$, montrer qu'il existe un nombre premier n tel que G soit isomorphe à $(\mathbb{Z}/2\mathbb{Z})^n$ (ainsi $p = 2^n - 1$). Réciproque ?

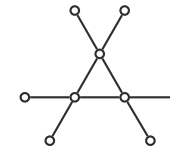
I.2.31 Soit p un nombre premier. Notons G l'ensemble des matrices $\begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}$, où a, b, c décrivent $\mathbb{Z}/p\mathbb{Z}$. Montrer que G est un sous-groupe d'ordre p^3 de $GL_3(\mathbb{Z}/p\mathbb{Z})$, et que G n'est pas commutatif (comparer avec l'exercice 13 de la page 83).

I.2.32
De combien de façons peut-on colorier les six faces d'un cube, si l'on dispose de q couleurs, et si l'on ne distingue pas deux coloriages se déduisant l'un de l'autre par rotation du cube ?

I.2.33 Soient S un ensemble et A une partie de $\mathcal{P}(S)$ formée de parties à deux éléments. Ainsi $\Gamma = (S, A)$ est un *graphe* (*cf.* le module I.8), ayant pour sommets les éléments de S et pour arêtes les éléments de A.

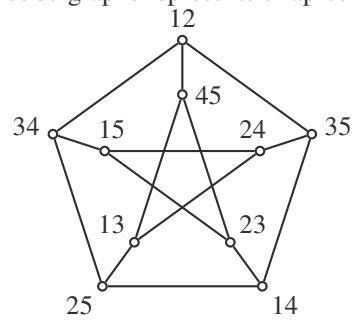
1. Soit G l'ensemble des permutations g de S telles que, pour toute arête $\{x, y\} \in A$, $\{g(x), g(y)\}$ appartienne à A. Montrer que G est un sous-groupe de $\mathfrak{S}(S)$, appelé *groupe des automorphismes* du graphe Γ.

2. Décrire le groupe G des automorphismes du graphe Γ représenté ci-dessous (il a neuf sommets), et en particulier trouver l'ordre de G.



3. Montrer que le groupe G précédent est isomorphe au groupe des symétries du cube.

I.2.34 En appliquant l'exercice I.2.22 de la page précédente avec $N := 5$ et $k := 2$, montrer que le groupe des automorphismes du graphe représenté ci-après est isomorphe à $\mathfrak{S}_5$:



I.2.35 Soient K un corps fini à q éléments et E un espace vectoriel de dimension finie $n \geqslant 1$ sur K. Calculer le cardinal de l'ensemble des drapeaux de E (*cf.* l'exercice I.2.0 de la page 88). Dénombrer ensuite les bases de E, et en déduire une autre preuve de la proposition 24 de la page 85.

I.2.36 Soient K un corps fini à q éléments et X un espace affine de dimension finie $n \geqslant 1$ sur K. Calculer l'ordre du groupe affine $\mathcal{GA}(X)$, de deux façons différentes : soit en appliquant la proposition 24 de la page 85, soit en dénombrant les bases affines de X.

I.2.37 Soit G un groupe multiplicatif fini. Montrer que l'action naturelle de $\mathrm{Aut}(G)$ (groupe des automorphismes de G) sur $G \setminus \{1\}$ est transitive si, et seulement s'il existe un nombre premier p et un entier n tels que G soit isomorphe à $\big((\mathbb{Z}/p\mathbb{Z})^n, +\big)$.

Algèbre bilinéaire I.3

Ce module est centré sur la notion de bilinéarité. Les formes bilinéaires, et les formes quadratiques qui leur sont associées, jouent un rôle important dans presque tous les domaines des mathématiques (géométries : euclidienne, symplectique…, géométrie différentielle, analyse, arithmétique, probabilités et statistiques, optimisation) ; on les rencontre aussi en mécanique (tenseurs d'inertie, oscillations, formalisme hamiltonien) et en physique (en électromagnétisme ou en mécanique quantique sous leur variante sesquilinéaire par exemple). Dans ce module, sont privilégiés les aspects algébriques élémentaires ; on trouvera des prolongements dans d'autres directions dans les modules I.6 (groupes orthogonaux et groupes unitaires), II.1, II.4 et II.5 ainsi que dans le module sur les surfaces de [L2-L3] (quadriques) et dans [MPA1] (groupes symplectiques, analyse numérique matricielle…). Nous détaillons les algorithmes classiques : LU, Gauß, Gram-Schmidt, Cholesky, QR et leurs relations. (Ces algorithmes sont présentés de façon itérative « informelle » et nous ne discutons pas le problème de leur stabilité et de leur efficacité numérique ; tout cela est repris dans [MPA1].)

Dans tout le module, on désignera par K un *corps commutatif de caractéristique nulle*. Si E est un K-espace vectoriel de dimension finie $n > 0$ et si $\mathcal{B} := (e_1, \dots, e_n)$ est une base fixée de E, on conviendra en général de noter par $X, Y, \dots$ les matrices unicolonnes formées des coordonnées des vecteurs $x, y, \dots \in E$ dans la base $\mathcal{B}$.

1 Dualité

La dualité en algèbre linéaire est sans doute la plus simple des nombreuses dualités qui apparaissent dans divers domaines des mathématiques. Elle fournit un exemple central de bilinéarité : le crochet de dualité.

1.1 Formes linéaires et hyperplans

Rappels : soit E un K-espace vectoriel, l'espace $\mathcal{L}(E, K)$ des formes linéaires sur E est appelé dual de E et noté E^*. Si E est un K-espace vectoriel de dimension finie n, l'espace dual E^* est aussi de dimension n.

Exemple. Soit $\mathcal{B} := (e_i)_{i\in I}$ une base de l'espace E, l'application $x \mapsto x_i$ de E dans K qui associe au vecteur x de E sa coordonnée d'indice i est une *forme linéaire* sur E, nous la noterons e_i^* (application $i^{ème}$ coordonnée). Cette forme est caractérisée par les relations $e_i^*(e_j) = \delta_{ij}$.

Proposition 1. La famille $(e_i^*)_{i\in I}$ est libre.

Démonstration. Si $\sum_{i\in I} \lambda_i e_i^* = 0$, en prenant la valeur en e_j, $0 = \sum_{i\in I} \lambda_i e_i^*(e_j) = \lambda_j$. ∎

Exercice 1.

Soient $(\alpha_i)_{i\in I}$ une famille d'éléments non nuls de K, $\mathcal{B} := (e_i)_{i\in I}$ une base de E et $\mathcal{B}' := (\alpha_i e_i)_{i\in I} = (e'_i)_{i\in I}$; $\mathcal{B}'$ est une base de E, quelle relation lie $e_i'^*$ et e_i^* ?

Solution. Si $x \in E$, on a : $x = \sum_{i\in I} x_i e_i = \sum_{i\in I} \frac{1}{\alpha_i} x_i e'_i = \sum_{i\in I} x'_i e'_i$ d'où $e_i'^* = \dfrac{e_i^*}{\alpha_i}$.

Exemples.

(i) Le dual de K^n est l'ensemble des applications $(x_1, x_2, \ldots, x_n) \in K^n \mapsto \sum_{i=1}^{n} a_i x_i$, $a := (a_1, \ldots, a_n)$ étant un élément de K^n.

(ii) Soient $k \in \mathbb{N}$ et $a \in K$. L'application $P \mapsto P^{(k)}(a)$ est une forme linéaire sur $K[X]$.

(iii) Soit $A \in M_n(K)$ une matrice donnée, l'application $f_A : M \mapsto \mathrm{Tr}(AM)$ de $M_n(K)$ dans K est une forme linéaire sur $M_n(K)$.

Exercice 2.

Montrer que l'application $A \mapsto f_A$ est une bijection de l'espace vectoriel $M_n(K)$ sur son dual $\big(M_n(K)\big)^*$.

Solution. L'application $A \mapsto f_A$ est linéaire. Elle est injective : si $\mathrm{Tr}\, AM = 0$ quel que soit M, on a en particulier $\mathrm{Tr}\, AE_{ij} = a_{ji} = 0$, quelque soit $(i,j) \in [\![1,n]\!]^2$, donc $A = 0$. Comme $\dim\big(M_n(K)\big) = \dim\Big(\big(M_n(K)\big)^*\Big) = n^2$, l'application f_A est bijective.

On rappelle que le noyau d'une forme linéaire f non nulle sur un espace vectoriel E est appelé *hyperplan* de E.

Proposition 2. Le noyau d'une forme linéaire f non nulle sur l'espace vectoriel E est un sous-espace vectoriel de codimension 1, réciproquement tout sous-espace vectoriel H de E de codimension 1 est le noyau d'une forme linéaire non nulle f sur E. On dit que $f(x) = 0$ est une équation de H ; toute équation de H est de la forme $\alpha f(x) = 0$, α étant un scalaire non nul. Ainsi, les formes linéaires f telles que $H = \mathrm{Ker}\, f$ forment une droite vectorielle, privée de l'origine, du dual E^*.

Démonstration. Soit f une forme linéaire non nulle sur E. Son image est K, elle est isomorphe à un supplémentaire de son noyau, donc $\mathrm{Ker}\, f = H$ est un sous-espace de codimension 1.

Réciproquement, $H \subset E$ étant un sous-espace de codimension 1, il existe $a \in E$ non nul tel que $E = H \oplus Ka$. Soit p le projecteur sur la droite $D := Ka$ parallèlement à H. Pour

tout $x \in E$, il existe un unique scalaire $f(x) \in K$ tel que $p(x) = f(x)a$; $x \mapsto f(x)$ est une forme linéaire sur E, non nulle car $f(a) = 1$, et $f(x) = 0$ est équivalent à $p(x) = 0$, c'est-à-dire à $x \in H$.
Supposons que $H = \operatorname{Ker} g$, $g \in E^*$. On a $g(a) \neq 0$. Les restrictions respectives des applications g et $g(a)f$ aux sous-espaces H et D sont égales ; ces sous espaces étant supplémentaires, on a $g = g(a)f = \alpha f$. ∎

On rappelle que le *rang* d'une famille $(e_i)_{i\in I}$ d'éléments de l'espace vectoriel E est la dimension du sous-espace vectoriel engendré par cette famille. Ce rang est un entier positif ou bien il est infini.

Définition 1. Soient $(H_i)_{i\in I}$ une famille d'hyperplans de l'espace vectoriel E et $(f_i)_{i\in I}$ une famille d'équations des hyperplans H_i. Le rang de la famille $(f_i)_{i\in I}$ est indépendant du choix des formes f_i ; on l'appelle rang de la famille d'hyperplans $(H_i)_{i\in I}$. La famille $(H_i)_{i\in I}$ est dite libre si la famille $(f_i)_{i\in I}$ l'est.

1.2 Orthogonalité

1.2.1 Le crochet de dualité

Soient E un K-espace vectoriel et E^* son dual, l'application $(f, x) \mapsto f(x)$ de $E^* \times E$ dans K est dite *crochet de dualité*. On notera $<f, x> := f(x)$.
Soit $f \in E^*$, $x \mapsto <f, x>$ est une application linéaire de E dans K ;
Soit $x \in E$, $f \mapsto <f, x>$ est une application linéaire de E^* dans K.
On dit que le crochet est une application *bilinéaire*.

1.2.2 Bidual d'un espace vectoriel

Définition 2. Soient E un K-espace vectoriel et E^* son dual. Le dual de l'espace vectoriel E^* est appelé le *bidual* de E et noté E^{**}.

Nous allons voir que tout espace vectoriel E *s'injecte* « naturellement » dans son bidual. Pour tout $x \in E$ *fixé*, l'application de E^* dans K définie par $f \mapsto <f, x>$ est une *forme linéaire* sur E^* ; notons la $\iota(x)$. Quand x varie dans E, on obtient une application $\iota : x \mapsto \iota(x)$ de E dans E^{**} telle que :

$$<\iota(x), f> = <f, x>$$

(on notera l'abus de notation pour les crochets : $< \iota(x), f >$ est le crochet de dualité entre E^* et son dual E^{**}, alors que $<f, x>$ est le crochet de dualité entre E et E^*).

Proposition 3. L'application $\iota : E \to E^{**}$ définie ci-dessus est une application *linéaire injective* ; on l'appelle *injection canonique* de E dans E^{**}.

Démonstration. Vérifions d'abord la linéarité de ι :
pour tous $(x_1, x_2) \in E^2$, $\alpha \in K$ et $f \in E^*$, on a :

$$\begin{aligned} <\iota(\alpha x_1 + x_2), f> &= <f, \alpha x_1 + x_2> = \alpha <f, x_1> + <f, x_2> \\ &= <\alpha\iota(x_1), f> + <\iota(x_2), f> = <\alpha\iota(x_1) + \iota(x_2), f> \end{aligned}$$

donc $\iota(\alpha x_1 + x_2) = \alpha\iota(x_1) + \iota(x_2)$.

Montrons maintenant l'injectivité de l'application linéaire ι. On considère le noyau de $\iota : x \in \operatorname{Ker}\iota$ si, et seulement si, pour tout $f \in E^*$, $<\iota(x), f> = 0$, c'est-à-dire $f(x) = 0$. Supposons alors $x \neq 0$, et montrons qu'il existe une forme $f \in E^*$ n'annulant pas x ($f(x) \neq 0$), pour cela, on complète le système libre de E réduit à $\{x\}$ en une base $\mathcal{B}$ de E et l'on note $e_1 = x$, $\mathcal{B} = \{e_1\} \cup \mathcal{B}'$. L'application e_1^* n'annule pas le vecteur x ($e_1^*(x) = 1$). Ainsi $\operatorname{Ker}\iota = \{0\}$, ι est donc injective. ∎

Définition 3. Soient E un K-espace vectoriel et E^* son dual ; $x \in E$ et $f \in E^*$ sont dits *orthogonaux* si $f(x) = <f, x> = 0$. Les parties $A \subset E$ et $A^* \subset E^*$ sont dites *orthogonales* si $<f, x> = 0$ pour tous $x \in A$ et $f \in A^*$. Pour toute partie $A \subset E$ l'ensemble :

$$A^\perp := \{f \in E^* \mid <f, x> = 0,\ \forall x \in A\} \subset E^*$$

est appelé orthogonal de A. Symétriquement, pour toute partie $A^* \subset E^*$, l'ensemble

$$(A^*)^\circ := \{x \in E \mid <f, x> = 0,\ \forall f \in A^*\} \subset E$$

est appelé orthogonal de A^*.

Proposition 4. Soient E un K-espace vectoriel et E^* son dual. Pour tout $A \subset E$, $A^\perp$ est un sous-espace vectoriel de E^*. Pour tout $A^* \subset E^*$, $(A^*)^\circ$ est un sous-espace vectoriel de E. De plus $E^\perp = \{0\}$ et $(E^*)^\circ = \{0\}$.

Démonstration. En effet, $A^\perp = \bigcap_{x \in A} \operatorname{Ker}\iota(x)$ est une intersection de sous-espaces de E^*, donc un sous-espace de E^*.

De même, $(A^*)^\circ = \bigcap_{f \in A^*} \operatorname{Ker} f$ est une intersection de sous-espaces de E.

On a $f \in E^\perp$ si, et seulement si, $f(x) = 0$ pour tout $x \in E$, c'est-à-dire si, et seulement si, $f = 0$. De même, $x \in (E^*)^\circ$ si, et seulement si, $x \in \operatorname{Ker}\iota$, donc si, et seulement si, $x = 0$. ∎

Exercice 3.

Soient E un K-espace vectoriel et E^* son dual. Montrer que :

- pour tout $A \subset E$, $A^\perp = (\operatorname{Vect} A)^\perp$
- pour tout $A^* \subset E^*$, $(A^*)^\circ = (\operatorname{Vect} A^*)^\circ$.

Solution. Si $A \subset B \subset E$: $B^\perp \subset A^\perp$. On a $A \subset \operatorname{Vect} A$, donc $(\operatorname{Vect} A)^\perp \subset A^\perp$. Inversement, si $f \in A^\perp$, f est orthogonal à $\operatorname{Vect} A$: si $x \in \operatorname{Vect} A$, $x = \alpha_1 a_1 + \cdots + \alpha_n a_n$, avec $\alpha_1, \ldots, \alpha_n \in K$ et $a_1, \ldots, a_n \in A$, alors $f(x) = \alpha_1 f(a_1) + \cdots + \alpha_n f(a_n) = 0$. On montre l'autre égalité de façon analogue.

1.3 Transposition

Définition 4. Soient E et F deux K-espaces vectoriels et $u \in \mathcal{L}(E,F)$. L'application ${}^t u : F^* \to E^*$ définie par $f \mapsto {}^t u(f) := f \circ u$ est appelée application *transposée* de l'application u.

Ainsi, pour tout $x \in E$, ${}^t u(f)(x) = f(u(x))$.

Proposition 5. Soient E et F deux K-espaces vectoriels et $u \in \mathcal{L}(E,F)$. L'application transposée ${}^t u$ est *linéaire* ; c'est la seule application v de F^* dans E^* telle que

$$\forall x \in E,\ \forall f \in F^*,\ <v(f), x> = <f, u(x)>.$$

Démonstration. Pour tous $(f_1, f_2) \in \mathcal{L}(E,F)$ et $\alpha \in K$:

$${}^t u(\alpha f_1 + f_2) = (\alpha f_1 + f_2) \circ u = \alpha f_1 \circ u + f_2 \circ u = \alpha {}^t u(f_1) + {}^t u(f_2).$$

Pour tout $(f,x) \in F^* \times E$: $<{}^t u(f), x> = <f \circ u, x> = f(u(x)) = <f, u(x)>$. Inversement, soit $v \in \mathcal{L}(F^*, E^*)$ telle que pour tout $(f,x) \in F^* \times E$: $<v(f), x> = <f, u(x)>$. Soit $f \in F^*$; pour tout $x \in E$, on a $<v(f) - {}^t u(f), x> = 0$, donc $(v - {}^t u)(f) \in E^{\perp}$. Ainsi $(v - {}^t u)(f) = 0_{E^*}$ quels que soient $f \in F^*$ et $v = {}^t u$. ∎

Proposition 6. Soient E et F deux K-espaces vectoriels. L'application $u \mapsto {}^t u$ de $\mathcal{L}(E,F)$ dans $\mathcal{L}(F^*, E^*)$ est linéaire injective. On l'appelle *transposition*.

Démonstration. Pour tous $u, v \in \mathcal{L}(E,F)$, $\alpha \in K$, $x \in E$, $f \in F^*$:

$$<{}^t(\alpha u + v)(f), x> = f\big(\alpha u(x) + v(x)\big) = \alpha f\big(u(x)\big) + f\big(v(x)\big) = <\alpha {}^t u(f) + {}^t v(f), x>.$$

Donc ${}^t(\alpha u + v) = \alpha {}^t u(f) + {}^t v$: la transposition est linéaire.

Supposons ${}^t u = 0$. Pour tout $(f,x) \in F^* \times E$: $f\big(u(x)\big) = <f, u(x)> = 0$, donc $u(x) \in (F^*)^\circ$, donc $u(x) = 0$, d'où $u = 0$. La transposition est injective. ∎

Proposition 7. Soient E, F, G trois K-espaces vectoriels.
Pour tous $u \in \mathcal{L}(E,F)$ et $v \in \mathcal{L}(F,G)$, on a ${}^t(v \circ u) = {}^t u \circ {}^t v$.

Démonstration. Pour tout $(g,x) \in G^* \times E$:

$$<{}^t(v \circ u)(g), x> = <g, v \circ u(x)> = <{}^t v(g), u(x)> = <{}^t u \circ {}^t v(g), x>,$$

d'où ${}^t(v \circ u) = {}^t u \circ {}^t v$. ∎

Corollaire 8. Soient E et F deux K-espaces vectoriels et $u \in \mathcal{L}(E,F)$ une bijection. Alors ${}^t u$ est aussi bijective et $({}^t u)^{-1} = {}^t(u^{-1})$.

Démonstration. On vérifie que ${}^t(\mathrm{Id}_E) = \mathrm{Id}_{E^*}$.

On a $u \circ u^{-1} = \mathrm{Id}_F$ et $u^{-1} \circ u = \mathrm{Id}_E$, d'où ${}^t(u^{-1}) \circ {}^t u = \mathrm{Id}_{F^*}$ et ${}^t u \circ {}^t(u^{-1}) = \mathrm{Id}_{E^*}$. Ainsi ${}^t u$ est bijective et son inverse est ${}^t(u^{-1})$. ∎

1.4 Dualité en dimension finie

Lorsque les espaces vectoriels étudiés sont *de dimension finie*, on peut préciser et compléter les résultats des trois derniers paragraphes.

Théorème et définition 9 (Bases duales). Soient E un K-espace vectoriel de dimension finie $n > 1$ et E^* son dual. Soit $(e_i)_{i\in[\![1,n]\!]}$ une base de E. Les formes coordonnées $(e_i^*)_{i\in[\![1,n]\!]}$ forment une base de E^*. Symétriquement, si $(f_i)_{i\in[\![1,n]\!]}$ est une base de E^*, il existe une unique base $(e_i)_{i\in[\![1,n]\!]}$ de E telle que $f_i = e_i^*$ quel que soit $i \in [\![1,n]\!]$. Les bases $(e_i)_{i\in[\![1,n]\!]}$ et $(e_i^*)_{i\in[\![1,n]\!]}$ sont dites *bases duales*.

Démonstration. La famille $(e_i^*)_{i\in[\![1,n]\!]}$ est libre (proposition 1 de la page 96) et de cardinal n, c'est donc une base de l'espace vectoriel E^* dont la dimension est n.

Soit $(f_i)_{i\in[\![1,n]\!]}$ une base de E^*.

Considérons l'application $u : E \to K^n$ définie par $x \mapsto \big(f_1(x), \dots, f_n(x)\big)$; elle est linéaire et injective ; en effet $u(x) = 0$ si, et seulement si, $f_i(x) = 0$ pour tout $i \in [\![1,n]\!]$, c'est-à-dire $x \in \big(\mathrm{Vect}(f_1,\dots,f_n)\big)^\circ = (E^*)^\circ = \{0\}$ i.e. $x = 0$. Comme $\dim E = \dim K^n = n$, l'application u est bijective. Pour tout $i \in [\![1,n]\!]$ il existe un unique $e_i \in E$ tel que $f_j(e_i) = \delta_{ij}$ pour tout $j \in [\![1,n]\!]$, il est défini par $e_i = u^{-1}(0,\dots,1,\dots 0)$ ($u(e_i)$ est le $i^{\text{ème}}$ vecteur de la base canonique de K^n). ∎

Si l'espace vectoriel E est de dimension infinie, pour toute base $(e_i)_{i\in I}$ de E, le système $(e_i^*)_{i\in I}$ est *libre* mais n'est plus une base de E^* (*cf.* [L1], exercice II.3.0).

Exemple. Les résultats élémentaires sur les polynômes d'interpolation de Lagrange (*cf.* [L1], module II.6) se traduisent de la façon suivante.
Soit $E = K_n[X]$. Soit $(a_0,\dots,a_n)$ un $(n+1)$-uple de scalaires deux à deux *distincts*. Pour tout $k \in [\![0,n]\!]$, l'application $f_k : P \mapsto P(a_k)$ est une forme linéaire sur E. La famille $(f_k)_{k\in[\![0,n]\!]}$ est une base de E^*. C'est la base duale de la famille des polynômes de Lagrange associés aux scalaires $(a_0,\dots,a_n)$.

Exercice 4.

Soit $E = \mathbb{R}_n[X]$. Soit $(a_0, \dots, a_n)$ un $(n+1)$-uple de scalaires deux à deux *distincts*. Montrer qu'il existe un $(n+1)$-uple $(c_0, \dots, c_n)$ de scalaires tel que, pour tout $P \in E$:

$$\int_0^1 P(t)\ \mathrm{d}t = \sum_{k=0}^{n} c_k P(a_k).$$

Solution.

Nous venons de voir, dans l'exemple ci-dessus, que les applications $f_k : P \mapsto P(a_k)$ forment une base de E^*. L'application $P \mapsto \displaystyle\int_0^1 P(t)\ \mathrm{d}t$ est une *forme linéaire* sur E, donc un élément de E^*. Les scalaires $c_0, \dots, c_n$ sont les *coordonnées* de celle-ci dans la base $(f_k)_{k\in[\![0,n]\!]}$.

Exercice 5.

Soient $E = K_n[X]$ et $a \in K$. Pour tout $k \in [\![0, n]\!]$, l'application $f_k : P \mapsto P^{(k)}(a)$ est une forme linéaire sur E.

Montrer que $(f_k)_{k\in[\![0,n]\!]}$ est une base de E. Quelle est sa base duale ?

Solution. D'après la formule de Taylor pour les polynômes de $K[X]$ (*cf.* [L1], module II.6), pour tous $P \in E$ et $x \in K$:

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

on voit ainsi que les applications f_k donnent les coordonnées de P dans la base de E formée des polynômes *de degrés échelonnés* $Q_k = \dfrac{(x-a)^k}{k!}$, $k \in [\![0, n]\!]$.

Exercice 6.

Soient E un K-espace vectoriel de dimension $n > 0$.

(i) Soient $\mathcal{B} := (e_1, \dots, e_n)$ et $\mathcal{B}' := (e'_1, \dots, e'_n)$ deux bases de E. On note respectivement $\mathcal{B}^* := (e_1^*, \dots, e_n^*)$ et $\mathcal{B}'^* := (e'^*_1, \dots, e'^*_n)$ les bases duales.

Quelle relation lie les matrices de passage $P := \mathrm{Pass}(\mathcal{B}, \mathcal{B}')$ et $Q := \mathrm{Pass}(\mathcal{B}^*, \mathcal{B}'^*)$?

(ii) Soient $(e) := (e_1, \dots, e_n)$ une base de E et $(f) = (f_1, \dots, f_n)$ une base de E^* telle que :

$$\forall i \in [\![1,]\!]n, \forall x \in E, x = \sum_{i=1}^{n} x_i e_i,\ f_i(x) = \sum_{h=1}^{n} a_{hi} x_i.$$

Soit $(\varepsilon) := (\varepsilon_1, \dots, \varepsilon_n)$ la base de E duale de (f), déterminer $\mathrm{Pass}((e), (\varepsilon))$.

(iii) Soient $E = \mathbb{R}^3$ muni de sa base canonique (e) et $(f) = (f_1, f_2, f_3)$ les formes linéaires sur E telles que : $\begin{cases} f_1(x) = x_1 - x_2 + x_3 \\ f_2(x) = 2x_1 - 3x_2 - x_3 \\ f_3(x) = x_1 + 2x_2 + 2x_3 \end{cases}$ déterminer $\mathrm{Pass}((e), (\varepsilon))$.

Solution.

(i) Notons $P := (p_{ij})$, avec $p_{ij} := e_i^*(e'_j)$ c'est-à-dire $e'_j = \sum_{i=1}^{n} p_{ij} e_i$.

Soit $x \in E$: $x = \sum_{i=1}^{n} x_i e_i = \sum_{i=1}^{n} x'_i e'_i$. On sait que :

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = P \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix} = PX'$$

d'où $X' = P^{-1}X$. Notons $P^{-1} := (q_{ij})$. Pour tout $j \in [\![1, n]\!]$:

$$e'^*_j(x) = x'_j = \sum_{k=1}^{n} q_{jk} x_k = \sum_{k=1}^{n} q_{jk} e_k^*(x),$$

donc $e'^*_j = \sum_{k=1}^{n} q_{jk} e_k^*$ et $Q = {}^tP^{-1}$.

(ii) Soit $A := (a_{hi}$, d'après le résultat de (i), on a $\mathrm{Pass}((e), (\varepsilon)) = {}^tA^{-1}$

(iii) Le lecteur vérifiera que $\mathrm{Pass}((e), (\varepsilon)) = M = \frac{1}{8} \begin{pmatrix} -4 & 4 & 4 \\ -5 & 1 & 3 \\ 7 & -3 & -1 \end{pmatrix}$.

Proposition 10. Soient E un K-espace vectoriel *de dimension finie* n, E^* son dual et E^{**} son bidual. L'application *linéaire canonique* ι de E dans son *bidual* est un *isomorphisme* d'espaces vectoriels.

Démonstration. On a vu que ι est linéaire injective. Comme :

$$\dim E = \dim E^* = \dim E^{**} = n,$$

cette application est bijective et l'application inverse est linéaire. ∎

Théorème 11. Soient E un K-espace vectoriel *de dimension finie* $n > 0$ et E^* son dual. Pour tout sous-espace vectoriel F de E, $\dim F^\perp = n - \dim F$. Pour tout sous-espace vectoriel F' de E^*, $\dim F'^\circ = n - \dim F'$; $(F^\perp)^\circ = F$ et $(F'^\circ)^\perp = F'$.

Démonstration. Soit $F \subset E$ un sous-espace vectoriel de dimension $r > 0$; soit $(e_1, \ldots, e_r)$ une base de F, on la complète en une base $(e_1, \ldots, e_r, e_{r+1}, \ldots, e_n)$ de E et on associe à cette dernière sa base duale $(e_1^*, \ldots, e_r^*, e_{r+1}^*, \ldots, e_n^*)$. Soit $f \in E^*$. On a $f = \sum_{i=1}^{n} f(e_i)\, e_i^*$. On a $f \in F^\perp$ si, et seulement si, $f(e_j) = a_j = 0$ quel que soit $j \in [\![1, r]\!]$. On en déduit que $F^\perp = \mathrm{Vect}(e_{r+1}^*, \ldots, e_n^*)$ est de dimension $n - r$.

De même, si $F' \subset E^*$ est un sous-espace vectoriel de dimension p, on choisit une base $(f_1, \ldots, f_p)$ de F', que l'on complète en une base $(f_1, \ldots, f_p, f_{p+1}, \ldots, f_n)$ de E^*

et l'on associe à cette dernière sa base duale $(e_1, \dots, e_p, e_{p+1}, \dots, e_n)$. Soient $x \in E$ et $x_1, \dots, x_n$ ses coordonnées dans la base duale ($e_i^* = f_i,\ i \in [\![1, n]\!]$) : $x = \sum_{i=1}^{n} x_i e_i$. On a $x \in (F')^\circ$ si, et seulement si, $e_j^*(x) = x_j = 0$ quel que soit $j \in [\![1, p]\!]$. Ainsi, $(F')^\circ = \mathrm{Vect}(e_{r+1}, \dots, e_n)$ est de dimension $n - p$.

Pour tout sous-espace vectoriel F de E, $F \subset (F^\perp)^\circ$ et

$$\dim(F^\perp)^\circ = n - \dim F^\perp = n - (n - \dim F) = \dim F,$$

donc $F = (F^\perp)^\circ$. On prouve de façon analogue que $F' = (F'^\circ)^\perp$. ∎

On prouve facilement le corollaire suivant.

Corollaire 12. Soit E un K-espace vectoriel de dimension finie $n > 0$.
Tout sous-espace vectoriel $F \subset E$ de dimension p (avec $1 \leqslant p < n$) est l'intersection d'une famille libre $(H_1, \dots, H_{n-p})$ de $n - p$ hyperplans de E.
Si $f_1(x) = 0, \dots, f_{n-p}(x) = 0$ sont $n - p$ équations respectives de ces hyperplans, un vecteur $x \in E$ appartient à F si, et seulement si, $f_i(x) = 0$ quel que soit $i \in [\![1, n-p]\!]$.
On dit que $f_1(x) = 0, \dots, f_{n-p}(x) = 0$ est *un système d'équations* de F.
Inversement, si $(H_1, \dots, H_q)$ est une famille libre de q ($1 \leqslant q \leqslant n$) hyperplans de E, son intersection $F = \bigcap_{i \in [\![1,q]\!]} H_i$ est un sous-espace vectoriel de dimension $p = n - q$ de E.

Théorème 13 (Représentation matricielle de la transposée).
Soient E et F deux K-espaces vectoriels de dimensions respectives $n > 0$ et $p > 0$. Soient $\mathcal{B} := (e_1, \dots, e_n)$ une base de E et $\mathcal{B}^* := (e_1^*, \dots, e_n^*)$ la base de E^*, duale de $\mathcal{B}$. De même, soient $\mathcal{C} := (f_1, \dots, f_p)$ une base de F et $\mathcal{C}^* := (f_1^*, \dots, f_p^*)$ la base de F^*, duale de $\mathcal{C}$. Soient $u \in \mathcal{L}(E, F)$ et $M := \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}}(u)$ sa matrice dans les bases $\mathcal{B}$ et $\mathcal{C}$, alors $\mathrm{Mat}_{\mathcal{C}^*}^{\mathcal{B}^*}({}^t u) = {}^t M$.

Démonstration. Soit $M = (m_{ij})_{1 \leqslant i \leqslant p, 1 \leqslant j \leqslant n}$; pour tout $(i, j) \in [\![1, p]\!] \times [\![1, n]\!]$, m_{ij} est la $i^{ème}$ coordonnée de $u(e_j)$ dans la base $\mathcal{C}$, d'où :

$$m_{ij} = f_i^*(u(e_j)) = <f_i^*, u(e_j)> = <{}^t u(f_i^*), e_j>,$$

m_{ij} est donc la $j^{ème}$ coordonnée de ${}^t u(f_i^*)$ dans la base $\mathcal{B}^*$. ∎

Conséquence : les *matrices* M et ${}^t M$ ayant le même rang, on en déduit que les *applications* u et ${}^t u$ ont le même rang.

2 Applications multilinéaires

2.1 Définitions et exemples

Définition 5. Soit p un entier strictement positif. Soient $E_1, \ldots, E_p, F$ des K-espaces vectoriels. Une application u de $E_1 \times E_2 \times \ldots \times E_p$ dans F est dite *p-linéaire* si elle est « linéaire en chacun des arguments », c'est-à-dire : pour tout $i \in [\![1, p]\!]$ et pour tout choix des vecteurs $a_1 \in E_1, \ldots, a_{i-1} \in E_{i-1}, a_{i+1} \in E_{i+1}, \ldots, a_p \in E_p$, l'application $x_i \mapsto u(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_p)$, de E_i dans F, est *linéaire*.
Si $p = 2$, l'application u est dite *bilinéaire*, si $p = 3$, u est dite *trilinéaire*, si $F = K$, u est dite *forme p-linéaire*. Si l'on ne précise pas p, on parle d'application *multilinéaire*.

2.1.1 Propriétés des applications multilinéaires

On vérifie sans difficulté les propriétés suivantes.
Soit $u : E_1 \times E_2 \times \ldots \times E_p \to F$ une application p-linéaire, alors :
(i) l'image par u d'un p-uple dont un vecteur est nul est nulle ;
(ii) si $p \geqslant 2$, et si u n'est pas identiquement nulle, elle n'est *pas* linéaire ;
(iii) l'ensemble des applications p-linéaires de $E_1 \times E_2 \times \ldots \times E_p$ dans F est un sous-espace vectoriel de l'espace vectoriel des applications de $E_1 \times E_2 \times \ldots \times E_p$ dans F.

Exemples.
(i) Soit E un K-espace vectoriel. L'application qui au couple $(\lambda, x) \in K \times E$ associe $\lambda x \in E$ est *bilinéaire*.
(ii) Le crochet de dualité (*cf.* page 97). Soient E un K-espace vectoriel et E^* son dual, l'application $(f, x) \to f(x)$ de $E^* \times E$ dans K est une *forme bilinéaire*.
(iii) L'application de K^p dans K qui associe à $(x_1, \ldots, x_p) \in K^p$ le produit $x_1 \ldots x_p$ est p-linéaire.
(iv) Soient E, F deux espaces vectoriels et E^*, F^* leurs duaux. Soit $(f, g) \in E^* \times F^*$. L'application $(x, y) \mapsto f(x)g(y)$ (produit de deux formes linéaires) est une *forme bilinéaire* sur $E \times F$. Plus généralement, si $(f_1, \ldots, f_k) \in E_1^* \times \ldots \times E_k^*$, l'application $(x_1, \ldots, x_k) \mapsto f_1(x_1) \ldots f_k(x_k)$ est une forme k-linéaire sur $E_1 \times \ldots \times E_k$.
(v) Le produit vectoriel : l'espace vectoriel $\mathbb{R}^3$ étant muni de son produit scalaire canonique, l'application $(\vec{v}, \vec{w}) \mapsto \vec{v} \wedge \vec{w}$ de $\mathbb{R}^3 \times \mathbb{R}^3$ dans $\mathbb{R}^3$ est bilinéaire.
(vi) Soient $E_1 := M_{n,p}(K)$, $E_2 := M_{p,q}(K)$ et $E_3 := M_{q,r}(K)$. L'application qui, au couple de matrices $(A, B) \in E_1 \times E_2$, associe leur produit AB est une application bilinéaire de $E_1 \times E_2$ dans $F := M_{n,q}(K)$. L'application qui, au triplet de matrices $(A, B, C) \in E_1 \times E_2 \times E_3$, associe leur produit ABC est une application trilinéaire de $E_1 \times E_2 \times E_3$ dans $G := M_{n,r}(K)$. Nous laissons le lecteur généraliser au produit de m matrices.
(vii) Soient E un K-espace vectoriel et $\mathcal{B}$ une base de E. L'application qui, au n-uple $(x_1, \ldots x_n) \in E^n$, associe $\det_{\mathcal{B}}(x_1, \ldots, x_n) \in K$ est une *forme n linéaire* sur E^n. On la note $\det_{\mathcal{B}}$.

Proposition 14. Soit p un entier strictement positif. Soient $E_1, \dots, E_p, F$ des K-espaces vectoriels.
(i) Soient $n \in \mathbb{N}^*$, $\xi_{ij} \in K$ des scalaires et $x_{ij} \in E_i$ des vecteurs, pour $i \in [\![1,p]\!]$ et $j \in [\![1,n]\!]$. Si u est une application p-linéaire de $E_1 \times E_2 \times \dots \times E_p$ dans F, on a :

$$u\left(\sum_{j=1}^{n} \xi_{1j}x_{1j}, \sum_{j=1}^{n} \xi_{2j}x_{2j}, \dots, \sum_{j=1}^{n} \xi_{pj}x_{pj}\right) = \sum_{(j_i)} \xi_{1j_1}\xi_{2j_2}\dots\xi_{pj_p}u\big(x_{1j_1}, x_{2j_2}, \dots, x_{pj_p}\big),$$

la somme du second membre étant étendue à tous les p-uples $(j_1, \dots, j_p)$ tels que $1 \leqslant j_i \leqslant n$.
(ii) Si chacun des espaces E_i $(1 \leqslant i \leqslant p)$ est *de dimension finie* et si $\mathcal{B}_i = (e_{ij})_{1\leqslant j\leqslant n_i}$ est une base de E_i, toute application p-linéaire de $E_1 \times \dots \times E_p$ dans l'espace vectoriel F est entièrement déterminée par les $n_1 \dots n_p$ vecteurs de F : $u(e_{1j_1}, \dots, e_{pj_p})$ $(1 \leqslant i \leqslant p$ et, pour chaque $i : 1 \leqslant j_i \leqslant n_i)$. Inversement étant donné un système arbitraire de $n_1 \dots n_p$ vecteurs $c_{j_1\dots j_p}$ de F, il *existe* une *unique* application p-linéaire $u : E_1 \times \dots \times E_p \to F$ telle que $u(e_{1j_1}, \dots, e_{pj_p}) = c_{j_1\dots j_p}$.

Démonstration.

(i) On prouve la formule par récurrence sur p. Nous laissons les détails au lecteur.

(ii) On choisit $n := \sup\limits_{1\leqslant i\leqslant p} n_i$, $x_{ij} := e_{ij}$ si $j \leqslant n_i$ et $x_{ij} := 0$ si $n_i < j \leqslant n$, et l'on pose $x_i = \sum\limits_{j_i=1}^{n_i} \xi_{ij_i}e_{ij_i}$ $(1 \leqslant i \leqslant p)$. En appliquant (i), on en déduit :

$$u(x_1, x_2, \dots, x_p) = \sum_{(j_i)} \xi_{1j_1}\xi_{2j_2}\dots\xi_{pj_p}u(e_{1j_1}, e_{2j_2}, \dots, e_{pj_p}).$$

Réciproquement, étant donné un système $c_{j_1\dots j_p}$ de $n_1 \dots n_p$ vecteurs de F, l'application $u : (x_1, \dots, x_p) \mapsto \sum\limits_{(j_i)} \xi_{1j_1}\xi_{2j_2}\dots\xi_{pj_p}c_{j_1\dots j_p}$ est évidemment p-linéaire et répond à la question. ∎

2.2 Retour sur le déterminant

Nous allons revenir sur la théorie du déterminant, introduite dans [L1], en utilisant la notion de *forme multilinéaire alternée*.

Proposition et définition 15 (Formes multilinéaires alternées).
Soient E et F deux K-espaces vectoriels et $p \in \mathbb{N}^*$.
Une application p-linéaire $u : E^p \to F$ est dite :
(i) *alternée* si $u(x_1, x_2, \dots, x_p) = 0$ chaque fois qu'il existe deux indices $i, j \in [\![1,p]\!]$ *distincts* tels que $x_i = x_j$;
(ii) *antisymétrique* si $u(x_1, x_2, \dots, x_p)$ change de signe chaque fois que l'on échange deux des arguments x_i et x_j $(i \neq j)$.
Les définitions (i) et (ii) sont équivalentes.

Démonstration.
Supposons u alternée. Pour tout $(x_1, \dots, x_p) \in E^p$, on a, pour $1 \leqslant i < j \leqslant p$:

$$\begin{aligned} 0 = \quad & u(x_1, \dots, x_{i-1}, x_i + x_j, x_{i+1}, \dots, x_{j-1}, x_i + x_j, x_{j+1}, \dots x_p) \\ = \quad & u(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{j-1}, x_j, x_{j+1}, \dots x_p) \\ & + u(x_1, \dots, x_{i-1}, x_j, x_{i+1}, \dots, x_{j-1}, x_i, x_{j+1}, \dots x_p), \end{aligned}$$

donc u est antisymétrique. Réciproquement, si u est antisymétrique :

$$\begin{aligned} & u(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{j-1}, x_i, x_{j+1}, \dots x_p) \\ & \qquad = -u(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{j-1}, x_i, x_{j+1}, \dots x_p), \end{aligned}$$

donc $2u(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{j-1}, x_i, x_{j+1}, \dots x_p) = 0$ et

$$u(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{j-1}, x_i, x_{j+1}, \dots x_p) = 0.$$

En effet, K est un corps de caractéristique nulle, dans lequel 2 est inversible. Notons que l'équivalence des deux définitions s'étend à des espaces vectoriels sur un corps de caractéristique $q \neq 2$. ∎

Soit $s \in \mathfrak{S}_p$ une permutation de p éléments. Elle s'écrit (de façon non unique) comme produit de transpositions et, si m est le nombre de ces transpositions, $(-1)^m \in \{1, -1\}$ est indépendant de la décomposition. C'est la *signature* $\varepsilon(s)$ de la permutation. La signature d'une transposition est -1. On en déduit le résultat suivant.

Corollaire 16. Soit u une forme p-linéaire alternée.
Pour toute permutation $s \in \mathfrak{S}_p$, on a $u(x_{s(1)}, \dots, x_{s(p)}) = \varepsilon(s) u(x_1, \dots, x_p)$.

Théorème 17.
Soient E un K-espace vectoriel de dimension $n > 0$ et $\mathcal{B} = (e_1, \dots, e_n)$ une base de E. Pour toute forme u n-linéaire alternée sur E : $u = u(e_1, \dots, e_n) \det_{\mathcal{B}}$. Ainsi, $\det_{\mathcal{B}}$ est l'unique forme n-linéaire alternée sur E prenant la valeur 1 sur le n-uple $(e_1, \dots, e_n)$ des vecteurs de base. L'espace vectoriel des formes n-linéaires alternées sur E est de dimension 1, c'est une droite vectorielle.

Démonstration. L'application $\det_{\mathcal{B}} : E^n \to K$ est une forme n-linéaire alternée (*cf.* [L1], module II.7). Soit u une forme n-linéaire alternée sur E^n.
En appliquant la proposition 14 (ii), on obtient :

$$u(x_1, x_2, \dots, x_n) = \sum_{(j_i)} \xi_{1j_1} \xi_{2j_2} \dots \xi_{nj_n} u(e_{j_1}, e_{j_2}, \dots, e_{j_n}).$$

L'ensemble des indices $(j_1, \dots, j_n)$ de la sommation du second membre s'identifie à l'ensemble des applications de $[\![1, n]\!]$ dans $[\![1, n]\!]^n$.
Considérons l'application $(1, 2, \dots, n) \mapsto (j_1, j_2, \dots, j_n)$. Si elle n'est pas injective, on a $u(e_{j_1}, e_{j_2}, \dots, e_{j_n}) = 0$. On peut donc sommer seulement sur les applications injectives au

lieu de sommer sur toutes les applications. Si l'application $(1, 2, \dots, n) \mapsto (j_1, j_2, \dots, j_n)$ est injective, elle est bijective, c'est une permutation s de $[\![1, n]\!]$ et l'on a donc :

$$u(e_{s(1)}, e_{s(2)}, \dots, e_{s(n)}) = \varepsilon(s) u(e_1, e_2, \dots, e_n).$$

Ainsi :

$$\begin{aligned} u(x_1, x_2, \dots, x_n) &= u(e_1, e_2, \dots, e_n) \sum_{s \in \mathfrak{S}_n} \varepsilon(s) \xi_{1j_1} \xi_{2j_2} \dots \xi_{nj_n} \\ &= u(e_1, e_2, \dots, e_n) \det_{\mathcal{B}}(x_1, x_2, \dots, x_n). \end{aligned}$$

Remarque. La notion de forme n-linéaire alternée apparaît naturellement en géométrie pour $n = 2$ ou 3. Dans le plan (orienté) de la géométrie ordinaire, identifié à $\mathbb{R}^2$ par le choix d'un repère orthonormé direct, l'aire orientée $\mathcal{A}(\vec{v}_1, \vec{v}_2)$ du parallélogramme construit sur les deux vecteurs $\vec{v}_1$ et $\vec{v}_2$ est une forme bilinéaire alternée. Dans l'espace (orienté) de la géométrie ordinaire, identifié à $\mathbb{R}^3$ par le choix d'un repère orthonormé direct, le volume orienté $\mathcal{V}(\vec{v}_1, \vec{v}_2, \vec{v}_3)$ du parallélépipède construit sur les trois vecteurs $\vec{v}_1$, $\vec{v}_2$ et $\vec{v}_3$ est une forme trilinéaire alternée. Il est donc naturel de se poser la question de la structure des formes n linéaires alternées en dimension quelconque. Même si l'on ignore tout de la théorie du déterminant, le calcul fait ci-dessus conduit *naturellement* à s'intéresser à l'expression :

$$\sum_{s \in \mathfrak{S}_n} \varepsilon(s) \xi_{1j_1} \xi_{2j_2} \dots \xi_{nj_n}$$

que nous avons prise pour *définition* du déterminant de la matrice (ξ_{ij}) dans [L1].

Application.

Grâce aux résultats ci-dessus, on peut donner une nouvelle démonstration de la formule $\det AB = \det A \det B$, pour $A, B \in M_n(K)$ (*cf.* [L1], module II.7).

(i) Soit E un K-espace vectoriel de dimension finie $n > 0$. Soit $\mathcal{B} := (e_1, \dots, e_n)$ une base de E. Pour tout $u \in \mathcal{L}(E)$ et tout $(x_1, \dots, x_n) \in E^n$:

$$\det_{\mathcal{B}}\big(u(x_1), \dots, u(x_n)\big) = \det_{\mathcal{B}}(x_1, \dots, x_n) \det_{\mathcal{B}}\big(u(e_1), \dots, u(e_n)\big).$$

Il suffit de remarquer que l'application $(x_1, \dots, x_n) \mapsto \det_{\mathcal{B}}\big(u(x_1), \dots, u(x_n)\big)$ est n linéaire alternée.

(ii) Soient $u, v \in \mathcal{L}(E)$. On a :

$$\det_{\mathcal{B}}\big(uv(e_1), \dots, uv(e_n)\big) = \det_{\mathcal{B}}\big(u(e_1), \dots, u(e_n)\big) \det_{\mathcal{B}}\big(v(e_1), \dots, v(e_n)\big).$$

On applique (i) deux fois et on utilise $\det_{\mathcal{B}}(e_1, \dots, e_n) = 1$.

(iii) Soient $A, B \in M_n(K)$. On a $\det AB = \det A \det B$.

On considère les endomorphismes u et v de E canoniquement associés aux matrices A et B et on leur applique (ii).

Exercice 7.
Soient E un K-espace vectoriel de dimension finie $n > 0$, $a \in \mathcal{L}(E)$ et $\mathcal{B} := (e_1, \dots, e_n)$ une base de E. Montrer que, pour tout $(x_1, \dots, x_n) \in E^n$:

$$\sum_{i=1}^{n} \det_{\mathcal{B}}(x_1, \dots, x_{i-1}, a(x_i), x_{i+1}, \dots, x_n) = (\operatorname{Tr} a) \det_{\mathcal{B}}(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n).$$

Solution. Notons $A = (a_{ij})$ la matrice de a dans la base $\mathcal{B}$.

(i) L'application $u : (x_1, \dots, x_n) \mapsto \sum_{i=1}^{n} \det_{\mathcal{B}}(x_1, \dots, x_{i-1}, a(x_i), x_{i+1}, \dots, x_n)$ est n-*linéaire* car a est linéaire.

(ii) Elle est *alternée* ; en effet, soit $(h, k) \in \mathbb{N}^2$ tel que $1 \leqslant h < k \leqslant n$ et que $x_h = x_k$, montrons qu'alors $u(x_1, \dots, x_n) = 0$; posons $x_h = x_k = y$, on a :

$$\begin{aligned} u(x_1, \dots, x_n) = &\sum_{i \in [\![1,n]\!] \setminus \{h,k\}}^{n} \det_{\mathcal{B}}(x_1, \dots, x_{i-1}, a(x_i), x_{i+1}, \dots, x_n) \\ &+ \det_{\mathcal{B}}(x_1, \dots, x_{h-1}, a(y), x_{h+1}, \dots, , x_{k-1}, y, x_{k+1}, \dots, x_n) \\ &+ \det_{\mathcal{B}}(x_1, \dots, x_{h-1}, y, x_{h+1}, \dots, , x_{k-1}, a(y), x_{k+1}, \dots, x_n). \end{aligned}$$

La somme des deux derniers termes du second membre est nulle (échange de y et $a(y)$ dans une forme alternée). Par ailleurs, tous les termes $\det_{\mathcal{B}}(x_1, \dots, x_{i-1}, a(x_i), x_{i+1}, \dots, x_n)$ sont nuls pour les indices $i \in [\![1, n]\!] \setminus \{h, k\}$, puisque, dans ce cas, y figure deux fois dans la liste des arguments. Finalement, on a $u(x_1, \dots, x_n) = 0$.

(iii) On en déduit que $u(x_1, \dots, x_n) = \det_{\mathcal{B}}(x_1, \dots, x_n) u(e_1, \dots, e_n)$. On a :

$$u(e_1, \dots, e_n) = \sum_{i=1}^{n} \det_{\mathcal{B}}(e_1, \dots, e_{i-1}, a(e_i), e_{i+1}, \dots, e_n) = \sum_{i=1}^{n} a_{ii} = \operatorname{Tr} A = \operatorname{Tr} a,$$

car $a(e_i) = \sum_{i=1}^{n} a_{ji} e_j$. Finalement, $u(x_1, \dots, x_n) = (\operatorname{Tr} a) \det_{\mathcal{B}}(x_1, \dots, x_n)$.

2.3 Formes bilinéaires

2.3.1 Généralités

Soient E et F deux K-espaces vectoriels. On note $\mathcal{BL}(E, F)$ le K-espace vectoriel des formes bilinéaires sur $E \times F$.

Soit $\varphi \in \mathcal{BL}(E, F)$; pour tout $y \in F$, l'application f_y qui à $x \in E$ associe $\varphi(x, y) \in K$, est une *forme linéaire* sur E, c'est-à-dire un élément de E^*, de même, pour tout $x \in E$, l'application h_x qui à $y \in F$ associe $\varphi(x, y) \in K$, est une *forme linéaire* sur F, c'est-à-dire un élément de F^*. On définit ainsi une application $d_\varphi : F \to E^*$ par $d_\varphi : y \mapsto f_y$ et une application $g_\varphi : E \to F^*$ par $g_\varphi : x \mapsto h_x$. Ces deux applications sont évidemment linéaires et l'on a, pour tout $(x, y) \in E \times F$: $<d_\varphi(y), x> = <g_\varphi(x), y> = \varphi(x, y)$.

Proposition 18. Avec les notations ci-dessus, ι désignant l'injection canonique de E dans son bidual E^{**} et ι' celle de F dans F^{**}, on a $g_\varphi = {}^t d_\varphi \circ \iota$ et $d_\varphi = {}^t g_\varphi \circ \iota'$.

Démonstration. Notons $< . , . >_G$ le crochet de dualité pour un espace vectoriel G. En revenant à la définition de ι (*cf.* proposition 3 de la page 97), on a , pour tout $(x, y) \in E \times F$:

$$<d_\varphi(y), x>_E = <\iota(x), d_\varphi(y)>_{E^*} = <{}^t d_\varphi(\iota(x)), y>_F.$$

Par ailleurs, $<d_\varphi(y), x>_E = \varphi(x, y) = <g_\varphi(x), y>_F$, d'où $g_\varphi = {}^t d_\varphi \circ \iota$. On prouve l'autre égalité de façon similaire. ∎

On introduit la notion essentielle *d'orthogonalité* pour une forme bilinéaire φ. Elle généralise celle que nous avons définie quand φ est le crochet de dualité $< . , . >$.

Définition 6. Soient E et F deux K-espaces vectoriels, soit φ une forme bilinéaire sur $E \times F$. On dit que $x \in E$ et $y \in F$ sont *orthogonaux* (ou plus précisément φ-orthogonaux) si $\varphi(x, y) = 0$.

Dans le cas où $E = F$, on prendra garde au fait que si x est orthogonal à y, ceci n'implique pas nécessairement que y soit orthogonal à x. Cette difficulté disparaît dans les cas où φ est symétrique ou antisymétrique (*cf.* ci-dessous le paragraphe 2.4 de la page 111).

Le lecteur vérifiera :

Proposition et définition 19. Soient E et F deux K-espaces vectoriels, φ une forme bilinéaire sur $E \times F$, et deux parties $A \subset E$ et $B \subset F$:

(i) $A^\perp := \{y \in F | \ \forall x \in A \quad \varphi(x, y) = 0\}$ est un sous-espace vectoriel de F , appelé *sous-espace orthogonal* de A et $B^\perp := \{x \in E \mid \ \forall y \in B \quad \varphi(x, y) = 0\}$ est un sous-espace vectoriel de E , appelé *sous-espace orthogonal* de B ;

(ii) les deux assertions suivantes sont équivalentes :
   a) $A \subset B^\perp$
   b) $B \subset A^\perp$
   Lorsqu'elles sont vérifiées, on dit que les parties A et B sont orthogonales

Définition 7. Soient E et F deux K-espaces vectoriels, soit φ une forme bilinéaire sur $E \times F$, elle est dite *non dégénérée* si l'on a :

$$F^\perp = \{x \in E \mid \ \forall y \in F \ \varphi(x, y) = 0\} = \{0_E\}$$

$$E^\perp = \{y \in F \mid \ \forall x \in E \ \varphi(x, y) = 0\} = \{0_F\}$$

ou, de façon équivalente, si, et seulement si, les *deux* applications d_φ et g_φ sont *injectives*. Sinon, φ est dite *dégénérée*.

2.3.2 Formes bilinéaires en dimension finie

On se propose d'étudier $\mathcal{BL}(E,F)$ lorsque les K-espaces vectoriels E et F sont *de dimension finie*.

Théorème et définition 20 (Représentation matricielle).
Soient E et F deux K-espaces vectoriels de dimension finie strictement positive. Soient $\varphi \in \mathcal{BL}(E,F)$, $\mathcal{B} := (e_1,\dots,e_n)$ une base de E et $\mathcal{C} := (\varepsilon_1,\dots,\varepsilon_p)$ une base de F. On appelle *matrice associée à φ dans les bases $\mathcal{B}$ de E et $\mathcal{C}$ de F* la matrice $M := (m_{ij})_{(i,j)\in[\![1,n]\!]\times[\![1,p]\!]}$ où, pour tout $(i,j) \in [\![1,n]\!]\times[\![1,p]\!]$, $m_{ij} = \varphi(e_i,\varepsilon_j)$.
On note $M := \mathrm{Mat}(\varphi;\mathcal{B},\mathcal{C})$. Pour tous $x := \sum_{i=1}^{n} x_i e_i \in E$ et $y := \sum_{j=1}^{p} y_j \varepsilon_j \in F$, on a $\varphi(x,y) = {}^tXMY$.
L'application $\varphi \mapsto \mathrm{Mat}(\varphi;\mathcal{B},\mathcal{C})$ est un isomorphisme d'espaces vectoriels de $\mathcal{BL}(E,F)$ sur $M_{n,p}(K)$ et $\dim \mathcal{BL}(E,F) = np$.

Démonstration. On a :

$$\varphi(x,y) = \sum_{i=1}^{n} x_i \sum_{j=1}^{p} y_j \varphi(e_i,\varepsilon_j) = \sum_{i=1}^{n} x_i \sum_{j=1}^{p} y_j m_{ij} = {}^tXMY.$$

L'application $\varphi \mapsto \mathrm{Mat}(\varphi;\mathcal{B},\mathcal{C})$ est clairement linéaire injective. Elle est surjective : si $M \in M_{n,p}(K)$, elle est l'image de la forme bilinéaire définie par $\varphi(x,y) = {}^tXMY$. ∎

Le lecteur vérifiera facilement le résultat suivant :

Proposition 21. Soient $\mathcal{B}$ une base de E et $\mathcal{C}$ une base de F, de bases duales $\mathcal{B}^*$ et $\mathcal{C}^*$; pour toute forme bilinéaire $\varphi \in \mathcal{BL}(E,F)$, on a :

$$\mathrm{Mat}(\varphi;\mathcal{B},\mathcal{C}) = \mathrm{Mat}_{\mathcal{C}}^{\mathcal{B}^*}\, d_\varphi = \mathrm{Mat}_{\mathcal{B}}^{\mathcal{C}^*}\, g_\varphi.$$

Changement de bases

Proposition 22. Soient $\mathcal{B}$ et $\mathcal{B}'$ deux bases de E, soient $\mathcal{C}$ et $\mathcal{C}'$ deux bases de F ; on pose $P := \mathrm{Pass}(\mathcal{B},\mathcal{B}')$ et $Q := \mathrm{Pass}(\mathcal{C},\mathcal{C}')$. Pour toute forme bilinéaire $\varphi \in \mathcal{BL}(E,F)$, si $M := \mathrm{Mat}(\varphi;\mathcal{B},\mathcal{C})$ et $M' := \mathrm{Mat}(\varphi;\mathcal{B}',\mathcal{C}')$, alors : $M' = {}^tPMQ$. Les matrices M et M' sont donc *équivalentes*.

Démonstration. Pour tout $(x,y) \in E\times F$, on a

$$\varphi(x,y) = {}^tXMY = {}^t(PX')M(QY') = {}^tX'({}^tPMQ)Y' = {}^tX'M'Y',$$

d'où $M' = {}^tPMQ$. ∎

Rang d'une forme bilinéaire

Si E et F sont deux espaces vectoriels de dimension finie strictement positive, on appelle *rang* d'une forme bilinéaire $\varphi \in \mathcal{BL}(E,F)$ le rang d'une matrice associée à φ. Cette définition est légitime car deux matrices associées à φ, étant *équivalentes*, ont le même rang. Si E ou F est de dimension 0, $\varphi \in \mathcal{BL}(E,F)$ est nulle ; par définition son rang est 0. Le rang de φ est aussi le rang commun de d_φ et g_φ (*cf.* la proposition 21 de la page ci-contre).

Proposition 23. Soient E et F deux K-espaces vectoriels de dimension finie strictement positive, pour toute forme bilinéaire $\varphi \in \mathcal{BL}(E,F)$, les assertions suivantes sont équivalentes :
(i) φ est non dégénérée ;
(ii) d_φ (ou g_φ) est bijective ;
(iii) $\dim E = \dim F = \operatorname{rang}\varphi$.

Démonstration.
(i)⇒(ii). Si φ est non dégénérée, les applications linéaires d_φ et g_φ sont injectives, donc :

$$\dim F \leqslant \dim E^* = \dim E \qquad \text{et} \qquad \dim E \leqslant \dim F^* = \dim F,$$

ainsi $\dim E = \dim F = \dim E^* = \dim F^*$ et les applications d_φ et g_φ sont bijectives.
(ii)⇒(iii). L'application linéaire d_φ est bijective, donc $\dim F = \dim E^* = \dim E$.
On a $\operatorname{rang}\varphi = \operatorname{rang} d_\varphi = \dim E^* = \dim F = \dim E$.
(iii)⇒(i). Si $\dim E = \dim F = \operatorname{rang}\varphi$, ce nombre est aussi le rang commun de d_φ et g_φ, qui sont donc bijectives ; ainsi, φ est non dégénérée. ∎

2.4 Formes bilinéaires symétriques, antisymétriques

2.4.1 Généralités

On suppose dans toute cette section que $E = F$. On parlera, par abus de langage, de formes bilinéaires sur E (au lieu de $E \times E$) et l'on notera $\mathcal{BL}(E) := \mathcal{BL}(E,E)$.
Si $\dim E = n > 0$ et si $\mathcal{B} := (e_1, \ldots, e_n)$ est une base de E, on notera :

$$\operatorname{Mat}(\varphi;\mathcal{B}) := \operatorname{Mat}(\varphi;\mathcal{B},\mathcal{B}) = \big(\varphi(e_i,e_j)\big)_{i,j\in[\![1,n]\!]}.$$

De la proposition 22 de la page précédente, l'on déduit immédiatement le résultat suivant.

Proposition 24. Soient E un K-espace vectoriel de dimension $n > 0$, $\mathcal{B}$ et $\mathcal{B}'$ deux bases de E et $P := \operatorname{Pass}(\mathcal{B},\mathcal{B}')$. Pour toute forme bilinéaire $\varphi \in \mathcal{BL}(E)$, si $M := \operatorname{Mat}(\varphi;\mathcal{B})$ et $M' := \operatorname{Mat}(\varphi;\mathcal{B}')$, alors : $M' = {}^tPMP$.

Définition 8. Deux matrices $M, M' \in M_n(K)$ sont dites *congruentes* s'il existe $P \in GL_n(K)$ telle que $M' = {}^tPMP$.

La relation de congruence est une relation d'équivalence sur $M_n(K)$. Les classes d'équivalence sont les orbites d'une *action* du groupe $GL_n(K)$ sur l'ensemble $M_n(K)$ que le lecteur précisera.

Définition 9. Soient E un K-espace vectoriel et $\varphi \in \mathcal{BL}(E)$. On dit que :
(i) φ est *symétrique* si, quelque soit $(x, y) \in E^2$: $\varphi(x, y) = \varphi(y, x)$;
(ii) φ est *antisymétrique* si, quel que soit $(x, y) \in E^2$: $\varphi(x, y) = -\varphi(y, x)$.

Une forme $\varphi \in \mathcal{BL}(E)$ est antisymétrique si, et seulement si, $\varphi(x, x) = 0$ pour tout $x \in E$.

Remarque. On notera que, si F est un sous-espace vectoriel de E et φ une forme bilinéaire sur E, la *restriction* $\varphi_{|F}$ de φ à $F \times F$ est une forme bilinéaire sur F. De plus, si φ est symétrique (respectivement antisymétrique), il en est de même pour sa restriction.

Exemples.
(i) L'espace $E := K^n$ étant muni de sa base canonique, l'application $\varphi : K^n \times K^n \to K$ définie par $\varphi(x, y) = \sum_{i=1}^{n} x_i y_i$ est bilinéaire symétrique. Si $K := \mathbb{R}$, la forme φ est appelée *produit scalaire canonique* de $\mathbb{R}^n$ et notée $(\,|\,)$.
(ii) L'espace $E := K^2$ étant muni de sa base canonique, l'application $\varphi : K^2 \times K^2 \to K$ définie par $\varphi(x, y) := x_1 y_2 - x_2 y_1$ (le déterminant des vecteurs colonnes X et Y) est bilinéaire antisymétrique.

Proposition 25. Soit E un K-espace vectoriel. L'ensemble $\mathcal{S}(E)$ des formes bilinéaires symétriques sur E et l'ensemble $\mathcal{A}(E)$ des formes bilinéaires antisymétriques sur E sont deux sous-espaces vectoriels *supplémentaires* de $\mathcal{BL}(E)$.

Démonstration. Le corps K étant de caractéristique nulle, une forme à la fois symétrique et antisymétrique est identiquement nulle ; donc $\mathcal{S}(E) \cap \mathcal{A}(E) = \{0\}$. Si $\varphi \in \mathcal{BL}(E)$, la forme bilinéaire définie par $\frac{1}{2}\big(\varphi(x, y) + \varphi(y, x)\big)$ est symétrique, la forme bilinéaire définie par $\frac{1}{2}\big(\varphi(x, y) - \varphi(y, x)\big)$ est antisymétrique et l'on a, pour tous $x, y \in E$:

$$\varphi(x, y) = \frac{1}{2}\big(\varphi(x, y) + \varphi(y, x)\big) + \frac{1}{2}\big(\varphi(x, y) - \varphi(y, x)\big).$$

La preuve du résultat suivant est laissée au lecteur.

Proposition 26. Soient E un K-espace vectoriel et $\varphi \in \mathcal{BL}(E)$, elle est symétrique si, et seulement si, $d_\varphi = g_\varphi$, antisymétrique si, et seulement si, $d_\varphi = -g_\varphi$. Une forme symétrique ou antisymétrique est non-dégénérée si, et seulement si, d_φ (ou g_φ) est injective.

Exemples.
(i) Soient E un K-espace vectoriel et $f, g \in E^*$:
  a) L'application $(x, y) \mapsto f(x)f(y)$ définit une forme bilinéaire symétrique sur E.
  b) L'application $(x, y) \mapsto f(x)g(y)$ définit une forme bilinéaire sur E.
  c) Les applications $(x, y) \mapsto f(x)g(y) + f(y)g(x)$ et $(x, y) \mapsto f(x)g(y) - f(y)g(x)$ définissent des formes bilinéaires sur E, respectivement symétriques et antisymétriques.
(ii) L'application $(A, B) \mapsto \operatorname{Tr} AB$ de $M_n(K)^2$ dans K est une forme bilinéaire symétrique sur $M_n(K)$.

Exercice 8.
Soient E un K-espace vectoriel de dimension $n > 0$ et $\mathcal{B} := (e_1, \dots, e_n)$ une base de E. Soit $f \in E^*$. Quelle est la matrice associée à la forme bilinéaire $\varphi : (x, y) \mapsto f(x)f(y)$? On suppose que la forme f n'est pas identiquement nulle ; quel est le rang de φ ?
Réciproquement, si $M \in S_n(K)$ est une matrice symétrique de rang 1, caractériser la forme $\varphi \in \mathcal{BL}(E)$ telle que $M = \operatorname{Mat}(\varphi; \mathcal{B})$.

Solution. Notons, pour tout $i \in [\![1, n]\!]$, $f(e_i) = a_i$. On sait que $M = \big(\varphi(e_i, e_j)\big)_{i,j\in[\![1,n]\!]}$, donc $M = \big(f(e_i)f(e_j)\big) = (a_i a_j)$ est de rang 1 si l'un au moins des a_i, $i \in [\![1, n]\!]$, est non nul, c'est-à-dire si, et seulement si, la forme f n'est pas identiquement nulle.
Inversement, supposons la matrice $M = (m_{ij})_{i,j\in[\![1,n]\!]}$ symétrique et de rang 1. Il existe $(a_1, \dots, a_n) \in K^n \setminus \{0\}$ et $(b_1, \dots, b_n) \in K^n \setminus \{0\}$, tels que, pour tous $i, j \in [\![1, n]\!]$, on ait $m_{ij} = a_i b_j = a_j b_i$.

Soit $i_0 \in [\![1, n]\!]$ tel que $a_{i_0} \neq 0$. On a, pour tout $j \in [\![1, n]\!]$, $b_j = \frac{b_{i_0}}{a_{i_0}} a_j = \lambda a_j$ et $m_{ij} = \lambda a_i a_j$. On définit une forme linéaire $f \in E^*$ par $f(e_i) = a_i$, pour tout $i \in [\![1, n]\!]$. On a alors, pour tous $x, y \in E$: $\varphi(x, y) = \lambda f(x)f(y)$.

La preuve du résultat suivant est immédiate.

Proposition 27. Soient E un K-espace vectoriel de dimension $n > 0$ et $\mathcal{B}$ une base de E. Une forme bilinéaire φ sur E est symétrique (respectivement antisymétrique) si, et seulement si, la matrice associée $\operatorname{Mat}(\varphi; \mathcal{B})$ est symétrique (respectivement antisymétrique).
Les espaces vectoriels $\mathcal{S}(E)$ et $\mathcal{A}(E)$ sont respectivement isomorphes aux espaces vectoriels $S_n(K)$ et $A_n(K)$. On a $\dim \mathcal{S}(E) = \dfrac{n(n+1)}{2}$ et $\dim \mathcal{A}(E) = \dfrac{n(n-1)}{2}$.

2.4.2 Formes bilinéaires symétriques sur un K-espace vectoriel

Vecteurs conjugués

La notion centrale de cette section est celle de *conjugaison* par rapport à une forme bilinéaire *symétrique*. Elle jouera un rôle important dans la suite de ce module.

Définition 10 (Conjugaison). Soit φ une forme bilinéaire *symétrique* sur le K-espace vectoriel E. Deux vecteurs x et y de E sont dits *conjugués* s'ils sont orthogonaux, c'est-à-dire si $\varphi(x,y)=0$. Une famille $(e_i)_{i\in I}$ de vecteurs de E est dite φ-conjuguée si pour tous $i,j\in I$, $i\neq j$, on a $\varphi(e_i,e_j)=0$.
La famille est dite *φ-orthonormée* si pour tous $i,j\in I$, on a $\varphi(e_i,e_j)=\delta_{ij}$.

Remarque. Nous employons le mot *conjugués*, de préférence à *orthogonaux*, pour des raisons historiques (les théories des coniques et des quadriques). Notons que, la forme φ étant symétrique, la conjugaison est une relation symétrique : x et y sont conjugués si, et seulement si, y et x le sont. Conformément à la tradition, nous utiliserons à nouveau l'expression *orthogonaux* dans le cas des produits scalaires.

Exemple. Reprenons l'exemple (ii) de la page 113. Tout couple $(A,B)\in S_n(K)\times A_n(K)$ est formé de vecteurs φ-conjugués : $\operatorname{Tr} AB=\operatorname{Tr}{}^t(AB)=-\operatorname{Tr} AB$, donc $\operatorname{Tr} AB=0$.

Proposition 28. Soient E un K-espace vectoriel et $\varphi\in\mathcal{S}(E)$. Pour toute partie A de E, $A^\perp:=\{x\in E\mid\ \varphi(a,x)=0\ \forall a\in A\}$ est un sous–espace vectoriel de E égal à $(\operatorname{Vect} A)^\perp$.

Démonstration. On a évidemment $0\in A^\perp$. Pour tous $\alpha\in K$, $x,y\in A^\perp$ et $a\in A$, on a : $\varphi(a,\alpha x+y)=\alpha\varphi(a,x)+\varphi(a,y)=0$, donc $A^\perp$ est un sous–espace vectoriel de E. On a l'inclusion $A\subset\operatorname{Vect} A$, d'où $(\operatorname{Vect} A)^\perp\subset A^\perp$; de plus, si $x\in A^\perp$, il est conjugué de tout vecteur de A, d'où, par linéarité, de toute combinaison linéaire de vecteurs de A ; ainsi : $(\operatorname{Vect} A)^\perp=A^\perp$. ∎

Conséquence pratique.
Soient F un sous–espace vectoriel de E et $\mathcal{C}=(v_i)_{i\in I}$ une base de F, alors :

$$F^\perp=\big\{y\in E\mid\ \varphi(y,v_i)=0\ \forall i\in I\big\}.$$

Définition 11 (Noyau). On appelle *noyau de* φ le sous-espace vectoriel de E :

$$E^\perp=\{x\in E\mid\ \varphi(x,y)=0\ \forall y\in E\}.$$

Remarque. On a $E^\perp=\operatorname{Ker} d_\varphi$; ainsi, φ est non dégénérée si, et seulement si, $E^\perp=\{0\}$, c'est-à-dire si, et seulement si, seul le vecteur nul est conjugué de tous les vecteurs de E. Si E est de dimension finie n, $\operatorname{rang}\varphi=\operatorname{rang} d_\varphi=n-\dim E^\perp$: le rang de φ est la codimension de son noyau dans E. Plus généralement, si $E^\perp$ est *de codimension finie* r dans E, on dit aussi que φ est de rang r.

Exercice 9.

Soient $E:=\mathbb{C}^2$ et $\varphi(x_1,x_2):=x_1^2+2ax_1x_2-x_2^2$. Quel est le noyau de φ ?

Solution.

Le vecteur $x := x_1e_1 + x_2e_2$ appartient à $\operatorname{Ker}\varphi$ si, et seulement si, $\begin{cases} x_1 + ax_2 = 0 \\ ax_1 - x_2 = 0. \end{cases}$

Si $a^2 + 1 \neq 0$, ce système admet l'unique solution $x_1 = x_2 = 0$ et $\operatorname{Ker}\varphi = \{0\}$.
Si $a^2 + 1 = 0$, soit $a = \pm i$, $\operatorname{Ker}\varphi$ est la droite de $\mathbb{C}^2$ engendrée par le vecteur $(1, \pm i)$.

Définition 12 (Vecteurs isotropes).
Soient E un K-espace vectoriel et $\varphi \in \mathcal{S}(E)$.
Un vecteur $x \in E$ est dit *φ-isotrope* si $\varphi(x, x) = 0$.
L'ensemble des vecteurs φ-isotropes est appelé *cône isotrope* de φ.
La forme φ est dite *définie* si 0 est son seul vecteur isotrope.

La terminologie *cône isotrope* est justifiée par le fait que tout vecteur proportionnel à un vecteur isotrope est isotrope : le cône isotrope est une réunion de droites.

> ***Remarque.*** Tout vecteur de $\operatorname{Ker}\varphi$ est évidemment isotrope, si φ est définie, elle est donc non dégénérée (on prendra garde au fait que la réciproque est fausse).

Bases de vecteurs φ-conjugués

Le théorème suivant est le résultat central de cette section.

Théorème 29 (Existence d'une base de vecteurs φ-conjugués).
Soit E un K-espace vectoriel de dimension finie n.
Pour toute forme $\varphi \in \mathcal{S}(E)$, il existe une base de E formée de vecteurs φ-conjugués.
Si $n > 0$, la matrice de φ dans une telle base $\mathcal{B} = (e_1, \ldots, e_n)$ est *diagonale* : $D = \operatorname{Diag}(\varphi(e_1, e_1), \ldots, \varphi(e_n, e_n))$.
Si φ est de rang r, la base $\mathcal{B}$ contient $n - r$ vecteurs *isotropes* et ceux-ci forment une base de $E^\perp$.

Démonstration. Raisonnons par *récurrence* sur n.

- Si $n = 0$ ou $n = 1$, le résultat est évident.
- Soit $n > 1$, supposons le théorème vrai pour tout K-espace vectoriel de dimension $n - 1$. Soient E un K-espace vectoriel de dimension n et $\varphi \in \mathcal{S}(E)$, ou bien $\varphi = 0$ et toute base de E convient, ou bien $\varphi \neq 0$, dans ce dernier cas, il existe un vecteur non nul $e_1 \in E$ tel que $\varphi(e_1, e_1) \neq 0$ (sinon φ serait à la fois symétrique et antisymétrique, donc nulle). Soit $H := (Ke_1)^\perp$: c'est un hyperplan de E $\big(H$ est le noyau de la forme linéaire non-nulle $x \mapsto \varphi(e_1, x)\big)$. Soit $\varphi_1 = \varphi_{|H}$, d'après l'hypothèse de récurrence, il existe une base $\mathcal{B}' = (e_2, \ldots, e_n)$ de H, formée de vecteurs φ_1-conjugués, $\mathcal{B} = (e_1, \ldots, e_n)$ est alors une base de E formée de vecteurs φ-conjugués.

Le rang r de φ est le rang de la matrice diagonale $D = \operatorname{Mat}(\varphi; \mathcal{B})$, la base $\mathcal{B}$ contient donc exactement $n - r$ vecteurs isotropes. Ces vecteurs appartiennent à $E^\perp$, comme l'espace vectoriel $E^\perp$ est de dimension $n - r$ ($E^\perp = \operatorname{Ker} d_\varphi$), ils en forment une base. ∎

3 Formes quadratiques

3.1 Généralités sur les formes quadratiques

3.1.1 Définitions et généralités

Théorème et définition 30. Soit Φ une application du K-espace vectoriel E dans K. On dit que Φ est une *forme quadratique* sur E s'il existe une forme bilinéaire ψ sur E telle que, pour tout $x \in E$: $\Phi(x) = \psi(x,x)$. Parmi les formes vérifiant cette condition, il *existe* une *unique* forme bilinéaire symétrique φ dite *forme polaire* de Φ.

Démonstration. Soit $\psi \in \mathcal{BL}(E)$ telle que $\Phi(x) = \psi(x,x)$. On peut écrire $\psi = \varphi + \alpha$, avec $\varphi \in \mathcal{S}(E)$ et $\alpha \in \mathcal{A}(E)$ et l'on a $\Phi(x) = \psi(x,x) = \varphi(x,x)$.

Si $\varphi_1, \varphi_2 \in \mathcal{S}(E)$ sont deux formes telles que $\Phi(x) = \varphi_1(x,x) = \varphi_2(x,x)$, alors $\varphi_1(x,x) - \varphi_2(x,x) = (\varphi_1 - \varphi_2)(x,x) = 0$, la forme $\varphi_1 - \varphi_2$ est alors à la fois symétrique et antisymétrique, donc nulle, et $\varphi_1 = \varphi_2$. ∎

Identités dites de polarisation

Soit φ la forme polaire de la forme quadratique Φ, on a, pour tout $(x,y) \in E^2$:

$$\Phi(x+y) - \Phi(x) - \Phi(y) = 2\varphi(x,y) \quad \text{et} \quad \Phi(x+y) - \Phi(x-y) = 4\varphi(x,y).$$

La vérification est facile : $\Phi(x+y) = \varphi(x+y, x+y) = \varphi(x,x) + 2\varphi(x,y) + \varphi(y,y)$, et $\Phi(x-y) = \varphi(x,x) - 2\varphi(x,y) + \varphi(y,y)$.

Définition 13. Soit Φ une forme quadratique sur E ; elle est dite *définie* (respectivement *non dégénérée*) si sa forme polaire φ est définie (respectivement non dégénérée). On appelle cône isotrope de Φ le cône isotrope de φ. Si E est de dimension finie, le rang de Φ est par définition celui de φ.

L'application qui associe à une forme quadratique sa forme polaire est clairement linéaire et bijective, d'où :

Proposition 31. Soit E un K-espace vectoriel. L'ensemble des formes quadratiques sur E, noté $Q(E)$, est un K-espace vectoriel isomorphe à $\mathcal{S}(E)$.

Si $\dim E = n$, $\dim Q(E) = \dfrac{n(n+1)}{2}$.

Le résultat suivant (très utile) se vérifie aisément.

Proposition 32. Soit $F \subset E$ un sous-espace vectoriel. La *restriction* à F d'une forme quadratique Φ sur E, notée $\Phi_{|F}$, est une *forme quadratique* sur F et la forme polaire de cette restriction est la restriction à F de la forme polaire de Φ. Si la forme quadratique Φ est *définie*, la restriction $\Phi_{|F}$ est *définie*.

Si la forme quadratique Φ est *non dégénérée*, $\Phi_{|F}$ ne l'est pas nécessairement. Soient, par exemple :

$$E := \mathbb{R}^2, \qquad \Phi(x_1, x_2) := x_1^2 - x_2^2 \quad \text{et} \quad F := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 = x_2\}.$$

On vérifie que Φ est non dégénérée ; cependant, $\Phi_{|F} = 0$ est dégénérée.

Exemples.

(i) Le carré d'une forme linéaire sur un K-espace vectoriel E.

Soit $f \in E^*$. L'application $x \mapsto \big(f(x)\big)^2$ est une forme quadratique de forme polaire $(x, y) \mapsto \varphi(x, y) = f(x)f(y)$.

(ii) Le produit de deux formes linéaires sur un K-espace vectoriel E.

Soient $f, g \in E^*$. L'application $x \mapsto f(x)g(x)$ est une forme quadratique de forme polaire $(x, y) \mapsto \varphi(x, y) = \frac{1}{2}\big(f(x)g(y) + f(y)g(x)\big)$.

Cela permet de fabriquer des formes quadratiques par *combinaisons linéaires* de formes quadratiques des types décrits ci-dessus.

Règle dite de dédoublement

Soit $\mathcal{B} := (e_i)_{i\in I}$ une base du K-espace vectoriel E. Soit $x := \sum_{i\in I} x_i e_i$. Pour tout $i \in I$, l'application $x \mapsto x_i^2$ est une forme quadratique de forme polaire $(x, y) \mapsto x_i y_i$. Pour tous $i, j \in I$, l'application $x \mapsto x_i x_j$ est une forme quadratique de forme polaire $(x, y) \mapsto \frac{1}{2}(x_i y_j + x_j y_i)$.

Caractérisation d'une forme quadratique

Proposition 33. Soit E un K-espace vectoriel. Une application $\Phi : E \to K$ est une forme quadratique sur E si, et seulement si, elle vérifie les conditions :
(i) pour tout $(\alpha, x) \in K \times E$: $\Phi(\alpha x) = \alpha^2 \Phi(x)$;
(ii) l'application de $E \times E$ dans K définie par $(x, y) \mapsto \Phi(x + y) - \Phi(x) - \Phi(y)$ est bilinéaire.

Démonstration. Une forme quadratique Φ vérifie évidemment ces deux conditions. Inversement, si une application $\Phi : E \to K$ vérifie ces conditions, on a :

$$\frac{1}{2}\big(\Phi(2x) - 2\Phi(x)\big) = \frac{1}{2}\big(4\Phi(x) - 2\Phi(x)\big) = \Phi(x)$$

ce qui prouve que $\Phi(x)$ est la forme quadratique associée à l'application bilinéaire :

$$\varphi : (x, y) \mapsto \frac{1}{2}\big(\Phi(x + y) - \Phi(x) - \Phi(y)\big).$$

3.1.2 Formes quadratiques sur un espace vectoriel de dimension finie

Matrice associée à une forme quadratique

Définition 14.
Soient E un K-espace vectoriel de dimension $n > 0$ et $\mathcal{B} := (e_1, \dots, e_n)$ une base de E. Soit $\Phi \in Q(E)$. On appelle matrice de la forme quadratique Φ dans la base $\mathcal{B}$ la matrice de la forme polaire φ de Φ dans la base $\mathcal{B}$. On note $\mathrm{Mat}(\Phi; \mathcal{B}) := \mathrm{Mat}(\varphi; \mathcal{B})$.

Proposition 34.
Soient E un K-espace vectoriel de dimension $n > 0$ et $\mathcal{B} = (e_1, \dots, e_n)$ une base de E. L'application qui associe à une forme quadratique Φ sur E sa matrice dans la base $\mathcal{B}$ est un isomorphisme d'espaces vectoriels de $Q(E)$ sur $S_n(K)$.

Exercice 10.
Soit Φ la forme quadratique sur $\mathbb{R}^3$ qui au vecteur (x_1, x_2, x_3) associe

$$\Phi(x) := x_1^2 - 3x_3^2 + 6x_1x_2 - x_1x_3 + 3x_2x_3.$$

Quelle est la matrice associée à Φ dans la base canonique $\mathcal{B}$ de $\mathbb{R}^3$?
Solution.
D'après la règle de dédoublement, la forme quadratique $x \mapsto x_1^2$ a pour forme polaire $(x, y) \mapsto x_1y_1$, la forme quadratique $x \mapsto 2x_1x_2$ a pour forme polaire $(x, y) \mapsto x_1y_2 + x_2y_1$, etc… en changeant les indices.
On obtient $M = \begin{pmatrix} 1 & 3 & -1/2 \\ 3 & 0 & 3/2 \\ -1/2 & 3/2 & -3 \end{pmatrix}$

Formes quadratiques et polynômes homogènes de degré deux

À tout *polynôme* P à n indéterminées sur le corps K (*cf.* module I.7), on associe une *fonction polynomiale* $\tilde{P}$ sur K^n. Le corps K étant de *caractéristique nulle*, donc *infini*, cette application est *injective*. Ceci nous permet « d'identifier », dans toute la suite, un polynôme à la fonction polynomiale qu'il définit. Un polynôme est *homogène de degré* $m \in \mathbb{N}$ si, et seulement s'il est somme de monômes de degré (total) m.
Soient E un K-espace vectoriel, $\mathcal{B} := (e_1, \dots, e_n)$ une base de E. Soit $\Phi \in Q(E)$ une forme quadratique de forme polaire φ. Pour tout $(x, y) \in E^2$, on note $x := \sum_{i=1}^{n} x_i e_i$ et $y := \sum_{j=1}^{n} y_j e_j$. On a $\varphi(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \varphi(e_i, e_j) x_i y_j$, soit :

$$\Phi(x) = \sum_{i=1}^{n} \alpha_{ii} x_i^2 + 2 \sum_{1 \leqslant i < j \leqslant n} \alpha_{ij} x_i x_j,$$

en posant $\alpha_{ij} := \varphi(e_i, e_j)$. À la forme quadratique Φ sur E, on associe ainsi un polynôme homogène de degré deux à n indéterminées sur K. Inversement, un tel polynôme s'écrit, de manière unique, $P(x_1, \ldots, x_n) = \sum_{i=1}^{n} a_{ii}x_i^2 + 2 \sum_{1 \leqslant i < j \leqslant n} a_{ij}x_i x_j$ et l'application de E dans K définie par $x \mapsto P(x_1, \ldots, x_n)$ est une forme quadratique, l'application :

$$\varphi : (x, y) \in E^2 \mapsto \sum_{i=1}^{n} a_{ii}x_i y_i + \sum_{1 \leqslant i < j \leqslant n} a_{ij}(x_i y_j + x_j y_i)$$

est bilinéaire symétrique et $\varphi(x, x) = \Phi(x)$. Si $P \in K[X_1, \ldots, X_n]$ est un polynôme homogène de degré m, on a la *formule d'Euler* (voir page 325) : $\sum_{i=1}^{n} X_i \dfrac{\partial P}{\partial X_i} = mP$. Si $m = 2$: $P = \dfrac{1}{2} \sum_{j=1}^{n} X_j \dfrac{\partial P}{\partial X_j}$, et en appliquant à nouveau la formule d'Euler aux $\dfrac{\partial P}{\partial X_j}$, qui sont homogènes de degré un, on obtient :

$$P = \frac{1}{2} \sum_{i,j=1}^{n} X_i X_j \frac{\partial^2 P}{\partial X_i \partial X_j}.$$

Noter que les polynômes $\dfrac{\partial^2 P}{\partial X_i \partial X_j}$ sont des *constantes* (ce sont les $2a_{ij}$).

Proposition 35. Soit E un K-espace vectoriel de dimension finie $n > 0$, l'espace vectoriel $Q(E)$ des formes quadratiques sur E est isomorphe à l'espace vectoriel des polynômes homogènes, de degré deux, à n indéterminées sur K. (L'isomorphisme dépend du choix d'une base de E.)

Exemples.

(i) L'espace vectoriel K^n étant muni de sa base canonique $(e_1, \ldots, e_n)$, l'application $(x_1, \ldots, x_n) \mapsto \sum_{i=1}^{n} x_i^2$ est une forme quadratique.

Sa forme polaire est $\varphi(x, y) = \sum_{i=1}^{n} x_i y_i$.

(ii) Soit $E := M_2(K)$. L'application $\Phi : M \in E \mapsto \det M \in K$ est une forme quadratique. En effet, si $M := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $M = aE_{11} + bE_{12} + cE_{21} + dE_{22}$ dans la base canonique de $M_2(K)$ et $\det M = ad - bc$ est un polynôme homogène de degré deux en les *quatre* variables a, b, c, d. La forme polaire de Φ est :

$$\varphi : (M, N) \mapsto \frac{1}{2}(a\delta + \alpha d - b\gamma - \beta c) \qquad \text{si} \qquad M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{et} \quad N = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}.$$

Décomposition en carrés

Théorème 36 (Décomposition en carrés).
Soit Φ une forme quadratique de rang r sur le K-espace vectoriel de dimension $n > 0$.
(i) Il existe r formes linéaires *indépendantes* sur E : $f_1, \dots, f_r$ et r scalaires *non nuls* : $\alpha_1, \dots, \alpha_r$ tels que $\Phi = \sum_{i=1}^{r} \alpha_i f_i^2$.
(ii) Si la forme quadratique Φ est combinaison linéaire à coefficients non nuls des carrés de s formes linéaires indépendantes sur E, alors $s = r$.

Démonstration.
(i) Soit φ la forme polaire de Φ. On sait qu'il existe une base $\mathcal{B} := (e_1, \dots, e_r, e_{r+1}, \dots e_n)$ de E, formée de vecteurs φ-conjugués, les r vecteurs $e_1, \dots, e_r$ étant non isotropes et les $(n-r)$ vecteurs $e_{r+1}, \dots e_n$ isotropes. Pour tout $x = \sum_{i=1}^{n} x_i e_i$, on a alors :

$$\Phi(x) = \sum_{i=1}^{n} \varphi(e_i, e_i) x_i^2 = \sum_{i=1}^{r} \Phi(e_i)(e_i^*(x))^2.$$

Ainsi, Φ est combinaison linéaire à coefficients non nuls des carrés des r formes linéaires indépendantes $e_1^*, \dots, e_r^*$.
(ii) On complète le système libre des s formes données en une base $\mathcal{C}'$ de E^*. La matrice de Φ, dans la base $\mathcal{C}$ de E dont $\mathcal{C}'$ est la base duale, est diagonale et de rang s.
On a donc $s = \operatorname{rang} \varphi = r$. ∎

Si, pour tout $i \in [\![1, r]\!]$, $\Phi(e_i)$ admet une *racine carrée* λ_i dans le corps K, on a, dans la situation (i) ci-dessus :

$$\Phi(x) = \sum_{i=1}^{r} \Phi(e_i)(e_i^*(x))^2 = \sum_{i=1}^{r} (\lambda_i e_i^*(x))^2$$

et Φ est la somme des carrés de r formes linéaires indépendantes. Cette hypothèse est vraie, d'une part, si $K := \mathbb{C}$ ou, d'autre part, si $K := \mathbb{R}$ et si les $\Phi(e_i)$ sont tous des réels positifs.

Corollaire 37. Soient E un $\mathbb{C}$-espace vectoriel, de dimension $n > 0$. Toute forme quadratique Φ de rang r sur E est la somme des carrés de r formes linéaires indépendantes sur E. La forme quadratique Φ est non dégénérée si, et seulement s'il existe des bases Φ-orthonormées de E.

Démonstration. Il reste seulement à prouver la dernière assertion.
Avec les notations ci-dessus : $\mathcal{B}' = (\lambda_1 e_1^*, \dots, \lambda_r e_r^*, e_{r+1}^*, \dots, e_n^*)$ est une base de E^*. Soit $\mathcal{B} = (\varepsilon_1, \dots, \varepsilon_n)$ la base de E dont elle est la base duale ; pour tout $x = \sum_{i=1}^{n} \xi_i \varepsilon_i$, on a $\Phi(x) = \sum_{i=1}^{r} \xi_i^2$. Ainsi, $\operatorname{Mat}(\Phi; \mathcal{B}) = \operatorname{Diag}(\underbrace{1, \dots, 1}_{r \text{ fois}}, 0, \dots, 0)$.
Le résultat s'en déduit immédiatement : Φ est non dégénérée si, et seulement si, $r = n$. ∎

Remarque. Le cas particulier des formes quadratiques sur un $\mathbb{R}$-espace vectoriel est particulièrement important ; il sera étudié dans la section 4.

3.2 Décomposition LU, décomposition de Gauß

Dans cette section, on décrit deux *algorithmes* permettant le *calcul effectif* de bases d'un K-espace vectoriel de dimension finie conjuguées pour une forme quadratique donnée Φ. Le premier (la décomposition LU) utilise essentiellement l'algorithme du pivot de Gauß ; il nécessite de faire certaines hypothèses sur Φ. Le second (algorithme de Gauß) s'applique à toute forme quadratique.

3.2.1 Décomposition LU d'une matrice de $M_n(K)$

Définition 15. Soit $A \in M_n(K)$. Pour $k \in [\![1, n]\!]$ on appelle déterminant mineur principal d'ordre k de A le déterminant :

$$D_k(A) = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kk} \end{vmatrix}.$$

On conviendra que $D_0(A) = 1$.

Théorème 38. Soit $A \in M_n(K)$ une matrice dont tous les déterminants mineurs principaux sont *non nuls*. Il existe un couple unique (L, U) formé d'une matrice triangulaire inférieure *unipotente* L et d'une matrice triangulaire supérieure U tel que $A = LU$. Le $k^{ème}$ élément diagonal de U est $\dfrac{D_k(A)}{D_{k-1}(A)}$, $k \in [\![1, n]\!]$.

On rappelle qu'une matrice triangulaire inférieure est dite unipotente si tous ses termes diagonaux sont égaux à 1. Les matrices triangulaires inférieures unipotentes forment un *groupe* pour la multiplication. (La notation L, U est d'origine anglo-saxonne : L = « lower », U = « upper ».)

Démonstration. Soit $A \in M_n(K)$ une matrice dont tous les mineurs principaux sont *non nuls*.

Unicité. Supposons que $A = LU = L'U'$. Alors $L'^{-1}L = U'U^{-1}$. La matrice $L'^{-1}L$ est triangulaire inférieure *unipotente* et la matrice $U'U^{-1}$ est triangulaire supérieure ; donc $L'^{-1}L = U'U^{-1} = I_n$. Ainsi, $L = L'$ et $U = U'$.

Existence. On notera $E_{ij}^{(n)}$ les matrices élémentaires formant la base canonique de $M_n(K)$, en allégeant en E_{ij} s'il n'y a pas d'ambiguïté. On rappellera, *si besoin est*, en indice supérieur $^{(n)}$, la taille d'une matrice carrée (autre que l'identité I_n).

Appliquons à $A := (a_{ij})$ l'algorithme du pivot de Gauß (*cf.* [L1], module II.4). Puisque $a_{11} = D_1 \neq 0$, on peut l'utiliser comme *pivot* et remplacer chaque ligne $L_i(A)$ de A ($i \in [\![2, n]\!]$) par la ligne $L_i(A) - \dfrac{a_{i1}}{a_{11}} L_1(A)$ dont le premier terme

est *nul*. On pose $\ell_{i1} = \dfrac{a_{i1}}{a_{11}}$ et l'on introduit le produit (*commutatif*) de *matrices de transvection* :

$$T_1 := \prod_{i=2}^{n}(I - \ell_{i1}E_{i1}) = I - \sum_{i=2}^{n}\ell_{i1}E_{i1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -\ell_{21} & 1 & 0 & \dots & 0 \\ -\ell_{31} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\ell_{n1} & 0 & 0 & \dots & 1 \end{pmatrix}.$$

On pose $T_1A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$, où A_1 est une matrice carrée d'ordre $n-1$;

pour $i \in [\![2, n]\!]$, les lignes de T_1A sont les $L_i(A) - \ell_{i1}L_1(A)$.

On vérifie que les mineurs principaux de T_1A sont égaux à ceux de A (une matrice triangulaire unipotente est de déterminant un) ; donc les mineurs principaux de A_1 sont $D_k(A_1) = \dfrac{D_{k+1}(A)}{a_{11}}$, $k \in [\![1, n-1]\!]$ et ils sont tous non nuls. En particulier, le terme $(1,1)$ de A_1 est $\dfrac{D_2(A)}{a_{11}} = \dfrac{D_2(A)}{D_1(A)}$; il est non nul et, si $n > 2$, il peut être utilisé comme pivot pour la modification des lignes de A_1. On obtient ainsi une matrice triangulaire inférieure unipotente :

$$T_2^{(n-1)} = I_{n-1} - \sum_{i=3}^{n}\ell_{i2}E_{i2}^{(n-1)}$$

d'ordre $n-1$; on pose $T_2 = T_2^{(n)} = I_n - \sum\limits_{i=3}^{n}\ell_{i2}E_{i2}^{(n)} = I - \sum\limits_{i=3}^{n}\ell_{i2}E_{i2}$, donc :

$$T_2T_1A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1k} \\ 0 & \frac{D_2(A)}{D_1(A)} & \star & \dots & \star \\ 0 & 0 & & & \\ \vdots & \vdots & & A_2 & \\ 0 & 0 & & & \end{pmatrix},$$

où A_2 est une matrice carrée d'ordre $n-2$. Les mineurs principaux de T_2T_1A sont égaux à ceux de A.

En itérant la méthode (à titre d'exercice, le lecteur pourra formaliser ceci en utilisant une récurrence), on détermine des matrices triangulaires inférieures unipotentes $T_j := I - \sum\limits_{i=j+1}^{n}\ell_{ij}E_{ij}$, $j \in [\![1, n-1]\!]$ telles que la matrice $U = (u_{ij}) = T_{n-1}\dots T_2T_1A$ soit triangulaire supérieure. On a donc $A = LU$, avec $L = (T_{n-1}\dots T_1)^{-1} = T_1^{-1}\dots T_{n-1}^{-1}$, qui est évidemment triangulaire infé-

rieure unipotente. Les mineurs principaux de U sont égaux à ceux de A, et on a, pour tout $k \in [\![1,n]\!]$: $D_k(A) = D_k(U) = \prod_{i=1}^{k} u_{ii}$, donc $u_{kk} = \dfrac{D_k(A)}{D_{k-1}(A)}$. ∎

Exemple.
Dans les calculs pratiques, on ne vérifie pas *a priori* que les mineurs principaux sont non nuls : si l'algorithme peut se poursuivre c'est que l'on est dans ce cas !

$$A = \begin{pmatrix} 1 & 2 & 0 & -1 \\ 3 & 8 & 2 & 1 \\ 4 & 0 & 3 & 2 \\ -2 & 1 & 4 & 0 \end{pmatrix} \quad T_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 \\ -4 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{pmatrix} \quad T_1A = \begin{pmatrix} 1 & 2 & 0 & -1 \\ 0 & 2 & 2 & 4 \\ 0 & -8 & 3 & 6 \\ 0 & 5 & 4 & -2 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 2 & 2 & 4 \\ -8 & 3 & 6 \\ 5 & 4 & -2 \end{pmatrix} \quad T_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & -\frac{5}{2} & 0 & 1 \end{pmatrix} \quad T_2T_1A = \begin{pmatrix} 1 & 2 & 0 & -1 \\ 0 & 2 & 2 & 4 \\ 0 & 0 & 11 & 22 \\ 0 & 0 & -1 & -12 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 11 & 22 \\ -1 & -12 \end{pmatrix} \quad T_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{11} & 1 \end{pmatrix} \quad T_3T_2T_1A = \begin{pmatrix} 1 & 2 & 0 & -1 \\ 0 & 2 & 2 & 4 \\ 0 & 0 & 11 & 22 \\ 0 & 0 & 0 & -10 \end{pmatrix}$$

On a $L = T_1^{-1}T_2^{-1}T_3^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ 4 & -4 & 1 & 0 \\ -2 & \frac{5}{2} & -\frac{1}{11} & 1 \end{pmatrix}$.

Remarque importante. On constate dans l'exemple précédent que les parties subdiagonales des $3 = 4-1$ premières colonnes de L s'obtiennent tout simplement en recopiant successivement celles de T_1, T_2, T_3 en changeant tous les signes. Le cas général est similaire, L est connue *sans aucun calcul* à partir des $T_1, \ldots, T_{n-1}$: les parties subdiagonales des $n-1$ premières colonnes de L s'obtiennent en recopiant successivement celles de T_1, T_2, …, T_{n-1} en changeant tous les signes.

Cette *règle* est justifiée de la façon suivante. Pour tout $j \in [\![1, n-1]\!]$, on pose $T_j := I - \sum_{i=j+1}^{n} \ell_{ij}E_{ij} = I - V_j$. On a $T_j^{-1} = I + V_j$, donc

$$L = (I + V_1)(I + V_2)\ldots(I + V_{n-1}).$$

On observe que, si $(h,k) \in [\![1, n-1]\!]^2$ est tel que $h < k$, on a $V_hV_k = 0$ (attention au sens du produit !) : en effet quels que soient $i \in [\![h+1, n]\!]$ et $\ell \in [\![k+1, n]\!]$, on a $h < \ell$, donc $E_{ih}E_{\ell k} = 0$, on en déduit que $L = I + V_1 + \cdots + V_{n-1}$.

Si la matrice A est *symétrique*, on a une variante *très importante* de la décomposition $A = LU$.

Corollaire 39. Soit $A \in S_n(K)$ une matrice symétrique dont les mineurs principaux sont tous non nuls. Il existe un couple unique (L, D) formé d'une matrice L *triangulaire inférieure unipotente* et d'une matrice *diagonale* D tel que $A = LD{}^tL$.
On a $D = \mathrm{Diag}\left(\dfrac{D_k(A)}{D_{k-1}(A)}; k \in [\![1, n]\!]\right)$.

Démonstration. On reprend l'algorithme de la décomposition LU. À chaque étape, au lieu d'utiliser le pivot pour transformer seulement les *lignes*, on l'utilise pour transformer aussi les *colonnes* de façon similaire. On commence donc par écrire, au lieu de T_1A, la matrice *symétrique* $T_1A{}^tT_1$; on raye ensuite la première ligne et la première colonne de $T_1A{}^tT_1$, et l'on obtient à nouveau une matrice *symétrique* A_1', d'ordre $n-1$. On continue de façon évidente et l'algorithme ainsi modifié conduit clairement au résultat.

On peut aussi raisonner directement sur la décomposition et écrire de façon unique $A = LU = LDU'$ (avec U' triangulaire supérieure *unipotente* et D diagonale), d'où ${}^tA = A = LDU' = {}^tU'D{}^tL$ et, d'après l'unicité de la décomposition LU : $L = {}^tU'$, donc $U' = {}^tL$ et $A = LD{}^tL$. ∎

Conséquence

Soit φ une forme bilinéaire sur un K-espace vectoriel de dimension finie $n > 0$. On choisit une base $\mathcal{B}$ de E, *si* la matrice $A := \mathrm{Mat}(\varphi; \mathcal{B}) \in S_n(K)$ a tous ses déterminants mineurs principaux non nuls, on peut écrire comme ci-dessus $A = LD{}^tL$, donc $D = L^{-1}A({}^tL)^{-1} = L^{-1}A{}^tL^{-1}$, on peut alors définir une nouvelle base $\mathcal{B}'$ de E par la matrice de passage $\mathrm{Pass}(\mathcal{B}, \mathcal{B}') = {}^tL^{-1}$ et l'on a :

$$\mathrm{Mat}(\varphi; \mathcal{B}') = {}^t({}^tL^{-1})A{}^tL^{-1} = L^{-1}A{}^tL^{-1} = D.$$

On voit ainsi que la base $\mathcal{B}'$ est φ*-conjuguée*. Si l'on écrit les coordonnées de $x \in E$ en vecteurs colonnes : X pour les coordonnées dans la base $\mathcal{B}$ et X' pour celles dans la base $\mathcal{B}'$, les formules de passage s'écrivent : $X = {}^tL^{-1}X'$, $X' = {}^tLX$. On a donc :

$$X' = (I + {}^tV_1 + \cdots + {}^tV_{n-1})X = (I + {}^tV_{n-1}) \dots (I + {}^tV_2)(I + {}^tV_1)X.$$

Remarque. Nous donnerons un peu plus loin une autre méthode *algorithmique*, dite de *décomposition de Gauß*. Celle-ci permet de construire des bases φ-conjuguées dans tous les cas (sans condition sur les mineurs principaux de la matrice associée). Lorsque la condition de non nullité des mineurs principaux est satisfaite, on verra plus loin que la décomposition de Gauß, peut être considérée comme une « reformulation » de la décomposition LU pour les matrices symétriques (*cf.* page 128).

Exercice 11.

Calculer le couple (L, D) pour la matrice symétrique $A := \begin{pmatrix} 1 & 2 & -1 & 0 \\ 2 & 3 & 0 & 1 \\ -1 & 0 & 5 & -4 \\ 0 & 1 & -4 & 2 \end{pmatrix}$.

Calculer L^{-1}. Si φ est la forme bilinéaire définie par la matrice A dans la base canonique $\mathcal{B}$ de $\mathbb{R}^4$, trouver une base φ-orthogonale. Écrire la forme quadratique Φ définie par φ comme combinaison linéaire de carrés de formes linéaires indépendantes.

Solution. On a

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & -2 & 1 & 0 \\ 0 & -1 & -\frac{1}{4} & 1 \end{pmatrix} \quad D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & \frac{5}{2} \end{pmatrix} \quad L^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -3 & 2 & 1 & 0 \\ -\frac{11}{4} & \frac{3}{2} & \frac{1}{4} & 1 \end{pmatrix}.$$

On obtient une base $\mathcal{B}'$ φ-orthogonale en utilisant la matrice de passage

$$\mathrm{Pass}(\mathcal{B}, \mathcal{B}') = {}^tL^{-1} = \begin{pmatrix} 1 & -2 & -3 & -\frac{11}{4} \\ 0 & 1 & 2 & \frac{3}{2} \\ 0 & 0 & 1 & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Les nouvelles coordonnées $(x'_1, \ldots, x'_4)$, de $x \in \mathbb{R}^4$, s'écrivent en fonction des anciennes $(x_1, \ldots, x_4)$:

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \\ x'_4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & -1 & 0 \\ 0 & 1 & -2 & -1 \\ 0 & 0 & 1 & -\frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.$$

On a $\Phi(x) = x'^2_1 - x'^2_2 + 8x'^2_3 + \frac{5}{2}x'^2_4 = (x_1 + 2x_2 - x_3)^2 - (x_2 - 2x_3 - x_4)^2 + 8(x_3 - \frac{1}{4}x_4)^2 + \frac{5}{2}x_4^2$. Notons que la première étape de l'algorithme LU (sous sa version symétrique) donne dans ce cas :

$$\Phi(x) = {}^tXAX = {}^tX'BX' = x'^2_1 + \Phi_1(x_2, x_3, x_4) = (x_1 + 2x_2 - x_3)^2 + \Phi_1(x_2, x_3, x_4),$$

où $B = T_1 A {}^tT_1$ et $\Phi_1(x_2, x_3, x_4) = -x_2^2 + 4x_3^2 + 2x_4^2 + 4x_2x_3 + 2x_2x_4 - 8x_3x_4$ est la forme quadratique définie sur $\mathbb{R}^3$ par la matrice A'_1.

3.2.2 Décomposition de Gauß

Soient E un K-espace vectoriel de dimension $n > 0$ et Φ une forme quadratique sur E. Dans ce paragraphe, nous décrivons un algorithme (dit de Gauß) permettant de trouver n formes linéaires *indépendantes*, telles que Φ soit une combinaison linéaire de leurs *carrés* (on notera que l'on ne fait *aucune hypothèse restrictive* sur Φ.) La *base duale* de la base définie par ces formes est une base de E formée de vecteurs Φ-*conjugués* ; l'algorithme permet ainsi de construire des bases Φ-*conjuguées*.

La méthode est basée sur deux identités très simples :

$$\forall (a, b) \in K^2, \qquad a^2 + 2ab = (a + b)^2 - b^2 \quad \text{et} \quad 4ab = (a + b)^2 - (a - b)^2.$$

Exemples.

(i) Soit Φ la forme quadratique définie sur $\mathbb{R}^3$ par :

$$\Phi(x_1, x_2, x_3) := x_1^2 - x_2^2 + 6x_3^2 + 2x_1x_2 - 4x_1x_3 - 8x_2x_3.$$

On écrit :

$$\begin{aligned}\Phi(x_1, x_2, x_3) &= x_1^2 + 2x_1(x_2 - 2x_3) - x_2^2 + 6x_3^2 - 8x_2x_3\\ &= (x_1 + x_2 - 2x_3)^2 - (x_2 - 2x_3)^2 - x_2^2 + 6x_3^2 - 8x_2x_3\\ &= (x_1 + x_2 - 2x_3)^2 - 2(x_2^2 + 2x_2x_3) + 2x_3^2\\ &= (x_1 + x_2 - 2x_3)^2 - 2(x_2 + x_3)^2 + 4x_3^2\\ &= x_1'^2 - 2x_2'^2 + 4x_3'^2,\end{aligned}$$

avec $x_1' = x_1 + x_2 - 2x_3,\ x_2' = x_2 + x_3,\ x_3' = x_3$. On vérifie que les trois formes obtenues forment un système libre.

(ii) Soit Φ la forme quadratique définie sur $\mathbb{R}^3$ par $\Phi(x_1, x_2, x_3) := x_1x_2 + 4x_2x_3 + x_3x_1$. On observe que, dans cet exemple, les coefficients des termes x_1^2, x_2^2, x_3^2 sont *tous nuls*. On écrit :

$$\begin{aligned}\Phi(x_1, x_2, x_3) &= (x_1 + 4x_3)(x_2 + x_3) - 4x_3^2\\ &= \frac{1}{4}\big((x_1 + 4x_3 + x_2 + x_3)^2 - (x_1 + 4x_3 - x_2 - x_3)^2\big) - 4x_3^2\\ &= \frac{1}{4}\big((x_1 + x_2 + 5x_3)^2 - (x_1 - x_2 + 3x_3)^2\big) - 4x_3^2\\ &= \frac{1}{4}(x_1'^2 - x_2'^2) - 4x_3'^2,\end{aligned}$$

avec $x_1' = x_1 + x_2 + 5x_3,\ x_2' = x_1 - x_2 + 3x_3,\ x_3' = x_3$. On vérifie que les trois formes obtenues forment un système libre.

Revenons maintenant au cas général. Soient E un K-espace vectoriel de dimension $n > 0$, $\mathcal{B} := (e_1, \ldots, e_n)$ une base de E et Φ une forme quadratique sur E, pour tout $x \in E$, $x := \sum_{i=1}^{n} x_i e_i$, on a :

$$\Phi(x) = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{1 \leqslant i < j \leqslant n} a_{ij}x_ix_j.$$

Si $n = 1$, $\Phi(x) = ax^2$ et le résultat est acquis. On va raisonner par *récurrence* en utilisant les trois cas suivants (qui sont *disjoints*), pour lesquels on supposera que $n \geqslant 2$.

Cas 0. On suppose que la forme quadratique Φ est *identiquement nulle*.

On écrit alors : $\Phi(x) = 0x_1^2 + \cdots + 0x_n^2$.

Cas 1. On suppose qu'il existe $i \in [\![1,n]\!]$ tel que $a_{ii} \neq 0$. Moyennant éventuellement une permutation des indices, on peut supposer que $a_{11} \neq 0$. Comme dans le premier exemple ci-dessus, on écrit alors :

$$\Phi(x) = a_{11}x_1^2 + 2\sum_{2\leqslant i\leqslant n} a_{i1}x_ix_1 + \Psi(x_2,\dots,x_n)$$

$$\Phi(x) = a_{11}\big(x_1 + \sum_{i=2}^{n}\frac{a_{i1}}{a_{11}}x_i\big)^2 - \frac{1}{a_{11}}\big(\sum_{i=2}^{n} a_{i1}x_i\big)^2 + \Psi(x_2,\dots,x_n).$$

Ainsi $\Phi(x) = a_{11}x_1'^2 + \Phi_1(x_2,\dots,x_n)$, avec $x_1' = x_1 + \sum_{i=2}^{n}\frac{a_{i1}}{a_{11}}x_i$:

$$\Phi_1(x_2,\dots,x_n) = \Psi(x_2,\dots,x_n) - \frac{1}{a_{11}}\big(\sum_{i=2}^{n} a_{i1}x_i\big)^2.$$

On constate que $\Phi_1(x_2,\dots,x_n)$ est un polynôme homogène de degré 2 en les $n-1$ variables $(x_2,\dots,x_n)$.

Cas 2. On suppose que $\forall i \in [\![1,n]\!]\,,\ a_{ii} = 0$ sans que Φ soit identiquement nulle.

Si $n = 2$, on écrit $\Phi(x) = 2a_{12}x_1x_2 = \frac{a_{12}}{2}\big((x_1+x_2)^2 - (x_1-x_2)^2\big)$.

Si $n \geqslant 3$, on peut supposer, quitte à permuter les coordonnées, que $a_{12} \neq 0$.

On a alors $\Phi(x) = 2\sum_{1\leqslant i<j\leqslant n} a_{ij}x_ix_j$ et :

$$\begin{aligned}\frac{\Phi(x)}{2} &= a_{12}x_1x_2 + \sum_{j=3}^{n} a_{1j}x_1x_j + \sum_{j=3}^{n} a_{2j}x_2x_j + \Theta(x_3,\dots,x_n)\\ &= a_{12}\Big(x_1 + \frac{1}{a_{12}}\sum_{j=3}^{n} a_{2j}x_j\Big)\Big(x_2 + \frac{1}{a_{12}}\sum_{j=3}^{n} a_{1j}x_j\Big)\\ &\qquad - \frac{1}{a_{12}}\Big(\sum_{j=3}^{n} a_{2j}x_j\Big)\Big(\sum_{j=3}^{n} a_{1j}x_j\Big) + \Theta'(x_3,\dots,x_n)\\ &= \frac{a_{12}}{4}(x_1'^2 - x_2'^2) + \Theta''(x_3,\dots,x_n)\end{aligned}$$

où l'on a posé :

$$x_1' = x_1 + x_2 + \frac{1}{a_{12}}\Big(\sum_{j=3}^{n}(a_{2j}+a_{1j})x_j\Big), \quad x_2' = x_1 - x_2 + \frac{1}{a_{12}}\Big(\sum_{j=3}^{n}(a_{2j}-a_{1j})x_j\Big),$$

Θ'' étant un polynôme quadratique des $n-2$ variables $(x_3,\dots,x_n)$.

Les formes $x_1', x_2', x_3, \dots, x_n$ forment un système libre.

Le cas 1 permet de passer de n à $n-1$, le cas 2 termine l'algorithme ou permet de passer de n à $n-2$, le cas 0 termine l'algorithme. On obtient donc par récurrence les formes cherchées $x_1', \dots, x_n'$ et l'on vérifie facilement qu'elles forment un système libre (nous laissons les détails au lecteur).

Exemple. La forme quadratique :

$$\Phi(x) := x_1^2 + 3x_2^2 + 5x_3^2 + 2x_4^2 + 4x_1x_2 - 2x_1x_3 + 2x_2x_4 - 8x_3x_4$$

sur $\mathbb{R}^4$ admet pour matrice la matrice A de l'exercice 11 de la page 124. La première étape de l'algorithme de Gauß (cas 1, $a_{11} \neq 0$) donne :

$$\Phi(x) = x_1'^2 + \Phi_1(x_2, x_3, x_4) = (x_1 + 2x_2 - x_3)^2 + \Phi_1(x_2, x_3, x_4),$$

où $\Phi_1(x_2, x_3, x_4) = -x_2^2 + 4x_3^2 + 2x_4^2 + 4x_2x_3 + 2x_2x_4 - 8x_3x_4$ est la forme quadratique définie sur $\mathbb{R}^3$ par la matrice A_1' de l'algorithme LU. En revenant à la solution de l'exercice, on constate le parfait « parallélisme » entre les premières étapes des algorithmes de Gauß et LU « symétrique ». On poursuit par deux nouvelles étapes de l'algorithme de Gauß (cas 1). On obtient :

$$\Phi(x) = x_1'^2 - x_2'^2 + 8x_3'^2 + \frac{5}{2}x_4'^2 = (x_1+2x_2-x_3)^2 - (x_2-2x_3-x_4)^2 + 8(x_3-\frac{1}{4}x_4)^2 + \frac{5}{2}x_4^2,$$

et l'on constate à nouveau le parallélisme entre les algorithmes de Gauß et LU dans ces deux nouvelles étapes.

Décomposition de Gauß et décomposition LU.
Quand la matrice A de la forme quadratique Φ dans la base $\mathcal{B}$ vérifie la condition de non nullité de tous ses mineurs principaux, l'algorithme LU peut être interprété comme une simple « traduction » de l'algorithme de Gauß (on a dans ce cas $n-1$ applications du cas 1 de cet algorithme). On raisonne comme dans le cas de l'exercice 11 de la page 124, en utilisant la description donnée dans le paragraphe *Conséquences* de la page 124. Les formes x_i' sont données par les formules $x_i' = (I + {}^tV_i)X$, elles se lisent donc immédiatement sur L.

4 Formes quadratiques sur un espace vectoriel réel

Dans tout ce paragraphe $K := \mathbb{R}$ est le corps des réels. On parlera de formes quadratiques *réelles*. La *relation d'ordre* sur $\mathbb{R}$ joue un rôle fondamental ($\mathbb{R}$ est un *corps ordonné*.) Il est bon, avant d'aller plus loin, de revenir aux cas simples, mais essentiels pour la suite, de dimension 1 et 2, et de rappeler quelques questions de *signes* dans ces cas là. (La restriction d'une forme quadratique Φ à un sous-espace étant une forme quadratique, les restrictions, en particulier, à des *droites* ou des *plans* vectoriels donnent des informations fort intéressantes sur Φ.)

En *dimension* 1, une base étant choisie, toute forme quadratique s'écrit $\Phi(x) := ax^2$ ($a \in \mathbb{R}$). Si $a \neq 0$, $\Phi(x) = 0$ si, et seulement si, $x = 0$. On a $\Phi(x) \geqslant 0$ si $a > 0$ et $\Phi(x) \leqslant 0$ si $a < 0$, les inégalités étant strictes pour $x \neq 0$.

En *dimension* 2, une base étant choisie, toute forme quadratique s'écrit

$$\Phi(x_1, x_2) := ax_1^2 + 2bx_1x_2 + cx_2^2 \quad (a, b, c \in \mathbb{R}).$$

D'après ce que nous venons de rappeler en dimension 1, le polynôme homogène Φ de degré 2 reste de signe constant ou identiquement nul sur toute droite issue de l'origine,

privée de l'origine : si $x_2 \neq 0$, on pose $x_1 = tx_2$ et l'on écrit :

$$\Phi(x_1, x_2) = ax_1^2 + 2bx_1x_2 + cx_2^2 = x_2^2(at^2 + 2bt + c),$$

dont le signe est celui de $at^2 + 2bt + c$ (si $x_1 \neq 0$, on peut échanger les rôles des variables et considérer $a + 2b\tau + c\tau^2$.) On est ainsi ramené au problème classique du *signe du trinôme* : $at^2 + 2bt + c$ garde un signe constant (au sens large) si, et seulement si, son *discriminant* réduit $b^2 - ac$ est négatif ou nul, c'est-à-dire si, et seulement si, l'équation du second degré $at^2 + 2bt + c$ n'a pas de racine réelle ou a une racine réelle double.

4.1 Formes positives, définies positives Inégalité de Cauchy-Schwarz

Définition 16. Soient E un $\mathbb{R}$-espace vectoriel, Φ une forme quadratique sur E, φ sa forme polaire. On dit que Φ et φ sont *positives* si $\forall x \in E$, $\Phi(x) \geqslant 0$. On dit que Φ et φ sont *définies positives* si $\forall x \in E\backslash\{0\}$, $\Phi(x) > 0$. On dit que Φ et φ sont *négatives* (resp. *définies négatives*) si $-\Phi$ et $-\varphi$ sont positives (resp. définies positives).

On notera respectivement $Q^+(E)$ et $Q^{++}(E)$ *l'ensemble* des formes quadratiques positives et l'ensemble des formes quadratiques définies positives sur l'espace vectoriel réel E. Ce ne sont évidemment pas des espaces vectoriels ; ce sont par contre des *cônes* : $Q^+(E)$ est invariant par homothétie positive, $Q^{++}(E)$ par homothétie strictement positive.

Exemples.

(i) Soit $E := M_n(\mathbb{R})$. Soit $\varphi : M_n(\mathbb{R}) \times M_n(\mathbb{R}) \to \mathbb{R}$ définie par $\varphi(A, B) := \operatorname{Tr} {}^tAB$. L'application φ est bilinéaire. Elle est *définie positive* :

- $\forall A \in M_n(\mathbb{R})\,,\ \operatorname{Tr} {}^tAA = \sum_{(i,j)\in[\![1,n]\!]^2} a_{ij}^2 \geqslant 0,$
- $\operatorname{Tr} {}^tAA = 0$ si, et seulement si, les nombres *réels* a_{ij} sont tous nuls, quel que soit $(i, j) \in [\![1, n]\!]^2$, c'est-à-dire si, et seulement si, $A = 0$.

(ii) Soit $(a, b) \in \mathbb{R}^2$, avec $a < b$. On note $E = \mathcal{C}([a, b], \mathbb{R})$ le $\mathbb{R}$ espace vectoriel des fonctions réelles continues sur $[a, b]$. On définit $\varphi : E \times E \to \mathbb{R}$ par : $(f, g) \mapsto \int_a^b f(t)g(t)\ \mathrm{d}t$, elle est bilinéaire et *définie positive*.

Exercice 12.

Soit $E = \mathcal{C}([a, b], \mathbb{R})$, comme dans l'exemple précédent. Soit $w \in \mathcal{C}([a, b], \mathbb{R})$, $w \geqslant 0$, une fonction continue, *positive*, non identiquement nulle donnée. On définit une forme bilinéaire φ par :

$$\varphi : (f, g) \in E \times E \mapsto \int_a^b f(t)g(t)w(t)\ \mathrm{d}t.$$

Elle est positive, mais n'est pas en général définie positive. Si on la restreint au sous-espace F des fonctions polynomiales sur $[a, b]$, sa restriction $\psi = \varphi_{|_F}$ est définie positive.

Solution. On a $\Phi(f) = \varphi(f,f) = \displaystyle\int_a^b f(t)^2 w(t)\ \mathrm{d}t \geqslant 0$. Si $f \in E$ est telle que

$$\Phi(f) = \varphi(f,f) = \int_a^b f(t)^2 w(t)\ \mathrm{d}t = 0,$$

on a $w(t)f(t)^2 = 0$, quelque soit $t \in [a,b]$ puisque la fonction $t \mapsto w(t)f(t)^2$ est continue et positive sur l'intervalle $[a,b]$. Cela n'entraîne pas nécessairement que f est identiquement nulle, mais seulement qu'elle l'est sur l'ensemble $\{t \in [a,b] \mid w(t) \neq 0\}$, qui peut être distinct de $[a,b]$ (trouver un exemple !). Cet ensemble est ouvert, non vide ; soit $c \in [a,b]$ tel que $w(c) \neq 0$ (il en existe), la fonction w étant continue, il existe un intervalle ouvert I, de centre c dans $[a,b]$, sur lequel w ne s'annule pas. On en déduit que, pour tout $t \in I$, $f(t) = 0$. Si la fonction f est polynomiale (i.e. la restriction à $[a,b]$ d'une fonction polynomiale), elle a ainsi une *infinité* de zéros distincts ; elle est donc *identiquement nulle* : $\psi : F \times F \to \mathbb{R}$ est bien *définie positive*.

Proposition 40. Soit Φ une forme quadratique *définie* sur l'espace vectoriel réel E. Alors Φ est *définie positive* ou *définie négative*.

Démonstration. Supposons qu'il existe $(x,y) \in E^2$ tel que $\Phi(x) > 0$ et $\Phi(y) < 0$. On considère le *trinôme* (en $t \in \mathbb{R}$) :

$$h(t) = \Phi(x+ty) = \Phi(y)t^2 + 2\varphi(x,y)t + \Phi(x),$$

son *discriminant* (réduit) $\varphi(x,y)^2 - \Phi(x)\Phi(y)$ est strictement positif et l'équation du second degré $h(t) = 0$ a deux racines *réelles distinctes* t_1 et t_2.

Alors $\Phi(x+t_1y) = \Phi(x+t_2y) = 0$, et, Φ étant *définie*, on a $x + t_1y = x + t_2y = 0$, donc $y = 0$, ce qui est impossible.

On remarquera que cette preuve ne fait intervenir que la *restriction* de Φ au sous-espace vectoriel, de dimension inférieure ou égale à 2, engendré par x et y. ∎

Inégalité de Cauchy-Schwarz

Théorème 41. Soit Φ une forme quadratique positive sur l'espace vectoriel réel E. Pour tout $(x,y) \in E^2$ on a :

$$(\varphi(x,y))^2 \leqslant \Phi(x)\Phi(y).$$

Si de plus Φ est *définie positive*, l'égalité $(\varphi(x,y))^2 = \Phi(x)\Phi(y)$ est vraie si, et seulement si, x et y sont *colinéaires*.

Démonstration. Soit $(x,y) \in E^2$ fixé. On considère le *trinôme* (en $t \in \mathbb{R}$) :

$$h(t) := \Phi(x+ty) = \Phi(y)t^2 + 2\varphi(x,y)t + \Phi(x).$$

Puisque Φ est *positive*, on a $h(t) \geqslant 0$, pour tout réel t.

- Supposons $\Phi(y) \neq 0$; le trinôme h ne prend que des valeurs positives ou nulles, son discriminant (réduit) $(\varphi(x,y))^2 - \Phi(x)\Phi(y)$ est donc *négatif ou nul*, ce qui donne $(\varphi(x,y))^2 \leqslant \Phi(x)\Phi(y)$.
- Supposons maintenant $\Phi(y) = 0$. La fonction h est alors affine : $h(t) = 2\varphi(x,y)t + \Phi(x)$. Si l'on avait $\varphi(x,y) \neq 0$, elle prendrait des valeurs *strictement négatives*, ce qui n'est pas ; on a donc $\varphi(x,y) = 0$ et $0 = (\varphi(x,y))^2 \leqslant \Phi(x)\Phi(y)$.

L'inégalité est ainsi établie dans tous les cas.

Si x et y sont colinéaires, on a, sans aucune hypothèse sur la forme quadratique Φ : $\big(\varphi(x,y)\big)^2 = \Phi(x)\Phi(y)$.

Supposons que la forme quadratique Φ soit réelle et *définie positive* et qu'il existe $(x,y) \in E^2$ tel que $(\varphi(x,y))^2 = \Phi(x)\Phi(y)$, alors, ou bien $y = 0$, et x et y sont colinéaires, ou bien $y \neq 0$; dans ce dernier cas $\Phi(y) > 0$, h est de degré 2 et son discriminant est nul, l'équation $h(t) = 0$ a donc une (unique) racine réelle t_0 : $h(t_0) = \Phi(x + t_0 y) = 0$, donc $x + t_0 y = 0$ et x et y sont colinéaires.

On remarquera que cette démonstration ne fait intervenir que la *restriction* de Φ au sous-espace vectoriel, de dimension inférieure ou égale à 2, engendré par x et y. ∎

Corollaire 42. Soit Φ une forme quadratique *positive* sur un espace vectoriel réel E. Son *cône isotrope* et le *noyau* de sa forme polaire φ sont égaux. Ainsi, Φ est *définie positive* si, et seulement si, elle est *non dégénérée*.

Démonstration. Si $\Phi \in Q(E)$, $\operatorname{Ker}\varphi$ est toujours contenu dans le cône isotrope de Φ. La forme Φ étant positive on vérifie l'inclusion symétrique : si $\Phi \in Q^+(E)$ et si x est isotrope, $\Phi(x) = 0$ et, pour tout $y \in E$, on a $\varphi(x,y) = 0$ d'après *l'inégalité de Cauchy-Schwarz*, donc $x \in E^\perp$ et $x \in \operatorname{Ker}\varphi$. ∎

Corollaire 43 (Inégalité de Minkowski). Soit Φ une forme quadratique *positive* sur le $\mathbb{R}$-espace vectoriel E. Pour tous $x, y \in E$: $\sqrt{\Phi(x+y)} \leqslant \sqrt{\Phi(x)} + \sqrt{\Phi(y)}$.

Démonstration. Les termes considérés étant positifs, l'inégalité ci-dessus est vraie si, et seulement si, $\Phi(x+y) \leqslant \Phi(x) + \Phi(y) + 2\sqrt{\Phi(x)\Phi(y)}$, soit $2\varphi(x,y) \leqslant 2\sqrt{\Phi(x)\Phi(y)}$; d'après l'inégalité de Cauchy-Schwarz (théorème 41), cette inégalité est vraie. ∎

4.2 Signature d'une forme quadratique réelle Théorème d'inertie de Sylvester

Soit E un $\mathbb{R}$-espace vectoriel de dimension finie $n > 0$. Soit $\Phi \in Q(E)$, de forme polaire φ. Nous savons qu'il existe une base $\mathcal{B}$ de E formée de vecteurs deux à deux Φ-conjugués : $\operatorname{Mat}(\varphi;\mathcal{B}) = \operatorname{Diag}(\alpha_1, \ldots, \alpha_n)$. Soit p le nombre de termes strictement positifs de $\operatorname{Diag}(\alpha_1, \ldots, \alpha_n)$. Soit q le nombre de termes strictement négatifs de cette même matrice. Un tel couple (p,q) dépend-il de la base choisie $\mathcal{B}$? La réponse est non, c'est le théorème d'inertie de Sylvester.

Définition 17 (Signature d'une forme quadratique réelle).
Soit $\Phi \in Q(E)$ une forme quadratique sur un $\mathbb{R}$-espace vectoriel de dimension $n > 0$, notons $\Phi_{|F}$ sa restriction à un sous-espace vectoriel $F \subset E$. Soit $\mathcal{P}$ (resp. $\mathcal{Q}$) la famille des sous-espaces vectoriels F de E, tels que la *forme quadratique* $\Phi_{|F}$ soit *définie positive* (resp. *définie négative*). On pose $p := \sup_{F\in\mathcal{P}} (\dim F)$; on pose de même, $q := \sup_{F\in\mathcal{Q}} (\dim F)$. Le couple $(p, q) \in [\![0, n]\!]^2$ est, par définition, la *signature* de Φ.

Si $F = \{0\}$, alors $\Phi_{|F}$ est définie positive et définie négative ; ainsi, les ensembles $\mathcal{P}$ et $\mathcal{Q}$ ne sont pas vides.

Théorème 44 (Théorème d'inertie de Sylvester). Soit E un $\mathbb{R}$-espace vectoriel de dimension $n > 0$, soient Φ une forme quadratique sur E et (p, q) sa signature ; alors, pour toute base $\mathcal{B}$ de E formée de vecteurs deux à deux Φ-conjugués, la matrice diagonale $\mathrm{Mat}(\varphi; \mathcal{B}) := \mathrm{Diag}(\alpha_1, \ldots, \alpha_n)$ est formée de p réels strictement positifs, de q réels strictement négatifs et de $n - p - q$ zéros. Le rang de Φ est $p + q$.

Démonstration. Soit $\mathcal{B} := (e_1, \ldots, e_n)$ une base de vecteurs Φ-conjugués de E, on note $x := \sum_{i=1}^{n} x_i e_i \in E$. La matrice diagonale $\mathrm{Mat}(\varphi; \mathcal{B}) = \mathrm{Diag}(\alpha_1, \ldots, \alpha_n)$ a k termes strictement positifs, h termes strictement négatifs et $n-k-h$ termes nuls. Quitte à changer l'ordre des vecteurs de la base, on peut supposer $\alpha_1, \ldots, \alpha_k > 0$, $\alpha_{k+1}, \ldots, \alpha_{k+h} < 0$, $\alpha_{k+h+1} = \cdots = \alpha_n = 0$ (en convenant de ne pas écrire k s'il est nul, et de même pour h). On pose $E' := \mathrm{Vect}(e_1, \ldots, e_k)$ et $E'' := \mathrm{Vect}(e_{k+1}, \ldots, e_n)$. La restriction de Φ à E' est *définie positive* :

$$\Phi(x) = \alpha_1 x_1^2 + \cdots + \alpha_k x_k^2,$$

on a donc (toujours) $k \leqslant p$. La restriction de Φ à E'' est *négative* :

$$\Phi(x) = \alpha_{k+1} x_{k+1}^2 + \cdots + \alpha_{k+h} x_{k+h}^2.$$

Soit $F \in \mathcal{P}$ un sous-espace vectoriel de E, de dimension p, tel que $\Phi_{|F}$ soit définie positive (un tel F existe car l'ensemble $\{\dim F | F \in \mathcal{P}\} \subset \mathbb{N}$ est *fini* et non vide). On a $F \cap E'' = \{0\}$: sinon, il existerait $x \in F \cap E''$ non nul, et l'on aurait $\Phi(x) > 0$ et $\Phi(x) \leqslant 0$.

On en déduit que $F + E'' = F \oplus E''$ et $\dim(F + E'') = \dim F + \dim E'' = p + n - k \leqslant n$, d'où $p \leqslant k$. Comme $k \leqslant p$, on a finalement $p = k$. On démontrerait de même que $q = h$. ∎

Corollaire 45. Soit E un $\mathbb{R}$-espace vectoriel de dimension $n > 0$.
Soit $\Phi \in Q(E)$ de signature (p, q). Il existe une base $\mathcal{B}_0$ de E telle que :

$$\mathrm{Mat}(\varphi; \mathcal{B}_0) = \mathrm{Diag}(\underbrace{1, \ldots, 1}_{p \text{ fois}}, \underbrace{-1, \ldots, -1}_{q \text{ fois}}, \underbrace{0, \ldots, 0}_{n-p-q \text{ fois}}).$$

Si Φ est définie positive, $p = n$ et la base $\mathcal{B}_0$ est φ-orthonormée.

Démonstration. Soit $\mathcal{B} := (\mathrm{e}_1, \ldots, e_n)$ une base de vecteurs Φ-conjugués, comme ci-dessus. On note $x := \sum_{i=1}^{n} x_i e_i \in E$, on a $\mathrm{Mat}(\varphi; \mathcal{B}) = \mathrm{Diag}(\alpha_1, \ldots, \alpha_n)$ et :

$$\Phi(x) = \sum_{i=1}^{p} \alpha_i x_i^2 - \sum_{i=p+1}^{p+q} (-\alpha_i) x_i^2.$$

On pose $\beta_i := \sqrt{\alpha_i}$ pour $i = 1, \ldots, p$, $\beta_i := \sqrt{-\alpha_i}$ pour $i = p+1, \ldots, p+q$ et $\beta_i := 1$ pour $i = p+q+1, \ldots, n$, puis $\varepsilon_i = \frac{1}{\beta_i} e_i$, pour $i = 1, \ldots, n$, on a alors $x = \sum_{i=1}^{n} \xi_i \varepsilon_i$ et :

$$\Phi(x) = \sum_{i=1}^{p} \xi_i^2 - \sum_{i=p+1}^{p+q} \xi_i^2.$$

Exercice 13.

Soit Φ une forme quadratique *non dégénérée* sur un $\mathbb{R}$-espace vectoriel de dimension $2n$. On suppose qu'il existe un sous-espace vectoriel isotrope $F \subset E$ (i.e. tel que $\Phi_{|F} = 0$) de dimension n. Quelle est la signature de Φ ?

Solution. Soit (p, q) la signature de Φ. On peut écrire $E = H^+ \oplus H^-$, avec $\Phi_{|H^+}$ définie positive, $\Phi_{|H^-}$ définie négative, $\dim H^+ = p$ et $\dim H^- = q$. On a $F \cap H^+ = \{0\}$, donc $\dim(F + H^+) = \dim F + \dim H^+ = n + p \leqslant \dim E = 2n$, ce qui entraîne $p \leqslant n$. On a, de même, $\dim H^- = q \leqslant n$, comme $p + q = 2n$, on a nécessairement $p = q = n$. La signature de Φ est donc (n, n).

4.3 Matrices symétriques réelles définies positives Décomposition de Cholesky

Un peu d'Histoire

André Louis Cholesky, officier français, mort au front en 1918 un peu avant la fin de la première guerre mondiale, a élaboré la méthode qui porte aujourd'hui son nom dans le cadre de travaux de géodésie (en relation avec la triangulation et la cartographie). Ce genre de problèmes conduit à des systèmes linéaires, où le nombre n des équations est inférieur à celui p des inconnues, systèmes que l'on cherche à résoudre par la *méthode des moindres carrés*. On est ainsi conduit à la résolution d'une équation dite *normale* de la forme $Ax = b$ (*cf.* page 165), A étant une matrice *symétrique* et positive ; on peut résoudre cette équation normale en employant l'ingénieuse *méthode de Cholesky* décrite ci-dessous. Cette méthode fut oubliée puis exhumée par un mathématicien anglais, J. Todd vers 1946 ; elle a aujourd'hui de très nombreuses applications.

Cela, en particulier, motive l'étude des matrices de $S_n(\mathbb{R})$ appelées positives, par laquelle nous allons commencer.

Du corollaire 45 de la page précédente, on déduit le résultat suivant.

Proposition 46. Toute matrice $A \in S_n(\mathbb{R})$ est congruente à une matrice diagonale de la forme $\mathrm{Diag}(\underbrace{1, \ldots, 1}_{p \text{ fois}}, \underbrace{-1, \ldots, -1}_{q \text{ fois}}, \underbrace{0, \ldots, 0}_{n-p-q \text{ fois}})$ ($p, q \in [\![0, n]\!]$).

Définition 18. Une matrice $A \in S_n(\mathbb{R})$ est dite
- *positive* si, pour tout $x \in \mathbb{R}^n$, on a ${}^tXAX \geqslant 0$;
- *définie positive* si, pour tout $x \in \mathbb{R}^n \setminus \{0\}$, on a ${}^tXAX > 0$.

On notera $S_n^+(\mathbb{R})$ $\big($resp. $S_n^{++}(\mathbb{R})\big)$ *l'ensemble* des matrices positives (resp. définies positives) de $S_n(\mathbb{R})$.
La matrice A est évidemment positive (resp. définie positive) si, et seulement si, la forme quadratique Φ sur $\mathbb{R}^n$, dont la matrice dans la base canonique est A, l'est ; on en déduit le résultat suivant.

Proposition 47. Soit $A \in S_n(\mathbb{R})$. Elle est *positive* si, et seulement si, il existe une matrice inversible $P \in GL_n(\mathbb{R})$ et un entier $p \in [\![0, n]\!]$ tels que $A = {}^tPDP$, avec $D := \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}$. La matrice A est *définie positive* si, et seulement si, il existe une matrice $P \in GL_n(\mathbb{R})$ telle que $A = {}^tPP$.

Le résultat suivant est important.

Proposition 48. Soit $M \in M_{m,n}(\mathbb{R})$.
La matrice ${}^tMM$ appartient à $S_n^+(\mathbb{R})$ et $\operatorname{Ker} M = \operatorname{Ker} {}^tMM$, $\operatorname{rang} M = \operatorname{rang} {}^tMM$.

Démonstration. Posons $A = {}^tMM$. Soient $x \in \mathbb{R}^n$ et X le vecteur colonne associé. Notons $Y = MX = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$. On a ${}^tXAX = {}^tYY = \sum_{i=1}^{m} y_i^2 \geqslant 0$. On note que ${}^tXAX = 0$ si, et seulement si, $MX = 0$, i. e. $x \in \operatorname{Ker} M$, d'où $\operatorname{Ker} A \subset \operatorname{Ker} M$. Inversement on a $\operatorname{Ker} M \subset \operatorname{Ker} {}^tMM$, d'où $\operatorname{Ker} {}^tMM = \operatorname{Ker} M$. On en déduit $\operatorname{rang} M = \operatorname{rang} {}^tMM$ en utilisant le théorème du rang. ∎

Proposition 49. Une matrice symétrique réelle est *définie positive* si, et seulement si, tous ses *mineurs principaux* sont *strictement positifs*.

Démonstration. Supposons $A \in S_n^{++}(\mathbb{R})$. Il existe $P \in GL_n(\mathbb{R})$ telle que $A = {}^tPP$, donc $\det A = (\det P)^2 > 0$. Pour tout $k \in [\![1, n]\!]$, on pose $F_k = \operatorname{Vect}(e_1, \dots, e_k) \subset \mathbb{R}^n$ et $A_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & \vdots & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}$. La matrice A_k est la matrice de la forme quadratique $\Phi_k = \Phi_{|F_k}$; elle est donc, comme Φ, définie positive. D'où $\det A_k = D_k(A) > 0$.
Réciproquement, si tous les mineurs principaux de A sont strictement positifs, on peut appliquer à A la décomposition LU (sous sa forme symétrique, *cf.* le corollaire 39 de la page 124), ainsi $A = LD{}^tL$, où $D = \operatorname{Diag}\left(\dfrac{D_k(A)}{D_{k-1}(A)};\ k \in [\![1, n]\!]\right)$.
La matrice A est donc bien définie positive. ∎

On en déduit la décomposition dite *de Cholesky* d'une matrice de $S_n^{++}(\mathbb{R})$ (le lecteur pourra se reporter à l'exercice I.3.15 de la page 140 pour une démonstration du résultat n'utilisant pas la décomposition LU).

Proposition 50 (Décomposition de Cholesky).
Pour toute matrice $A \in S_n^{++}(\mathbb{R})$, il existe une *unique* matrice $T \in M_n(\mathbb{R})$ *triangulaire supérieure*, d'éléments diagonaux *strictement positifs* telle que $A = {}^tTT$.

Démonstration. Montrons d'abord *l'existence* d'une telle décomposition.

On part de la décomposition LU symétrique : $A = LD{}^tL$ rappelée ci-dessus. Pour tout $k \in [\![1, n]\!]$, les nombres réels $\dfrac{D_k(A)}{D_{k-1}(A)}$ sont *strictement positifs*, on définit la matrice diagonale $\Delta := \mathrm{Diag}\left(\sqrt{\dfrac{D_k(A)}{D_{k-1}(A)}}\ ;\ k \in [\![1, n]\!]\right)$.

On a alors $D = \Delta^2$ et $A = LD{}^tL = L\Delta\Delta{}^tL = {}^tTT$, si $T = \Delta{}^tL$.

Passons à la preuve de *l'unicité*. Soit $A = {}^tTT$ une décomposition de Cholesky. On écrit $T = \Delta_0 L_0$, où Δ_0 est une matrice diagonale, à termes strictement positifs, et L_0 une matrice triangulaire supérieure unipotente. On en déduit que $A = L_1U_1$, avec $L_1 = {}^tL_0$ et $U_1 = \Delta_0^2 L_0$. Il est clair que L_1 est triangulaire inférieure unipotente et U_1 triangulaire supérieure. D'après l'unicité de la décomposition LU, ces deux matrices sont déterminées de façon unique par A. Il en est donc de même de L_0 et de Δ_0^2, puis de Δ_0 et, finalement, de T. ∎

Pour *calculer effectivement* la décomposition de Cholesky $A = {}^tTT$, on peut utiliser la décomposition LU, mais il vaut mieux procéder plus directement et utiliser l'algorithme de Cholesky (qui est basé seulement sur *l'existence* du résultat).

Algorithme de Cholesky.

Soit $A := (a_{ij}) \in S_n^{++}(\mathbb{R})$. On *sait* qu'il existe $B = {}^tT$ triangulaire inférieure à coefficients positifs telle que $A = B{}^tB$ et l'on veut *calculer* B. On doit avoir :

$$a_{ij} = \sum_{k=1}^{n} b_{ik}b_{jk} = \sum_{k=1}^{\inf\{i,j\}} b_{ik}b_{jk}, \quad i, j \in [\![1, n]\!].$$

La matrice A étant symétrique, il suffit de considérer ces relations pour $i \leqslant j$, soit $a_{ij} = \sum_{k=1}^{i} b_{ik}b_{jk}$, pour $i, j \in [\![1, n]\!]$. La première colonne de B est donc :

$$b_{11} = \sqrt{a_{11}},\ b_{21} = \frac{a_{12}}{b_{11}}, \ldots,\ b_{n1} = \frac{a_{1n}}{b_{11}}.$$

On calcule les colonnes suivantes par récurrence. Supposons déjà calculées les $j - 1$

premières colonnes, la $j^{ème}$ est alors : $b_{ij} = 0$ si $i < j$, $b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2}$, et :

$$b_{ij} = \frac{a_{ji} - \sum_{k=1}^{j-1} b_{jk} b_{ik}}{b_{jj}}, \quad \text{si} \quad i > j.$$

Notons que les $a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2$ successivement calculés sont *nécessairement* strictement positifs, puisque l'on *sait* qu'il y a une solution unique.

On trouvera dans le module I.8 page 379 une version *récursive* de l'algorithme de Cholesky.

Exercice type corrigé

I.3.0 1) Soit Φ la forme quadratique sur $\mathbb{R}^n$ $(n \geqslant 2)$ qui à tout $x = (x_1, \ldots, x_n)$ associe :

$$\Phi(x) := \sum_{i=1}^{n} x_i^2 + \sum_{1 \leqslant i < j \leqslant n} x_i x_j.$$

En utilisant la méthode de Gauß, décomposer Φ en une somme de carrés de formes linéaires indépendantes.

2) Soit Ψ la forme quadratique sur $\mathbb{R}^n$ $(n \geqslant 2)$ qui à tout $x = (x_1, \ldots, x_n)$ associe $\Psi(x) := \sum_{1 \leqslant i < j \leqslant n} x_i x_j$.

Utiliser de même la méthode de Gauß pour décomposer Ψ en combinaison linéaire de carrés de formes linéaires indépendantes.

Solution. 1) On a :

$$\begin{aligned}
\Phi(x) &= x_1^2 + x_1 \sum_{i=2}^{n} x_i + \sum_{i=2}^{n} x_i^2 + \sum_{2 \leqslant i < j \leqslant n} x_i x_j \\
&= \big(x_1 + \frac{1}{2}(x_2 + \cdots + x_n)\big)^2 + \frac{3}{4} \sum_{i=2}^{n} x_i^2 + \frac{1}{2} \sum_{2 \leqslant i < j \leqslant n} x_i x_j \\
&= \big(x_1 + \frac{1}{2}(x_2 + \cdots + x_n)\big)^2 + \frac{3}{4}\big(x_2 + \frac{1}{3}(x_3 + \cdots + x_n)\big)^2 \\
&\quad + \frac{4}{2.3} \sum_{i=3}^{n} x_i^2 + \frac{1}{3} \sum_{3 \leqslant i < j \leqslant n} x_i x_j.
\end{aligned}$$

Si $n \geqslant 4$, nous pouvons donc formuler l'hypothèse : pour tout $k \in [\![4, n]\!]$,

$$\begin{aligned}
\Phi(x) = \big(x_1 + \frac{1}{2}(x_2 + \cdots + x_n)\big)^2 &+ \frac{3}{4}\big(x_2 + \frac{1}{3}(x_3 + \cdots + x_n)\big)^2 + \cdots \\
&+ \frac{k}{2(k-1)}\big(x_{k-1} + \frac{1}{k}(x_k + \cdots + x_n)\big)^2 \\
&+ \frac{k+1}{2k} \sum_{i=k}^{n} x_i^2 + \frac{1}{k} \sum_{k \leqslant i < j \leqslant n} x_i x_j.
\end{aligned}$$

On la vérifie par récurrence, en utilisant, si $n \geqslant 5$, l'égalité suivante :

$$\begin{aligned}
&(k+1)\sum_{i=k}^{n} x_i^2 + 2\sum_{k\leqslant i<j\leqslant n} x_i x_j \\
&\quad = (k+1)\big(x_k + \frac{1}{k+1}(x_{k+1}+\cdots+x_n)\big)^2 \\
&\qquad + \left(k+1-\frac{1}{k+1}\right)\sum_{i=k+1}^{n} x_i^2 + \left(2-\frac{2}{k+1}\right)\sum_{k+1\leqslant i<j\leqslant n} x_i x_j \\
&\quad = (k+1)\big(x_k + \frac{1}{k+1}(x_{k+1}+\cdots+x_n)\big)^2 \\
&\qquad + \frac{k(k+2)}{k+1}\sum_{i=k+1}^{n} x_i^2 + \frac{2k}{k+1}\sum_{k+1\leqslant i<j\leqslant n} x_i x_j.
\end{aligned}$$

Pour tout $k \in [\![1, n]\!]$ on définit la forme linéaire :

$$X_k := \sqrt{\frac{k+1}{2k}}\big(x_k + \frac{1}{k+1}(x_{k+1}+\cdots+x_n)\big)$$

et l'on obtient $\Phi(x) = \sum_{k=1}^{n} X_k^2$.

2) On a :

$$\begin{aligned}
\Psi(x) &= x_1x_2+\cdots+x_1x_n+x_2x_3+\cdots+x_2x_n + \sum_{3\leqslant i<j\leqslant n} x_i x_j \\
&= (x_1+x_3+\cdots+x_n)(x_2+x_3+\cdots+x_n) - (x_3+\cdots+x_n)^2 + \sum_{3\leqslant i<j\leqslant n} x_i x_j \\
&= \frac{1}{4}\Big(\big(x_1+x_2+2(x_3+\cdots+x_n)\big)^2 - (x_1-x_2)^2\Big) - \big(\sum_{i=3}^{n} x_i^2 + \sum_{3\leqslant i<j\leqslant n} x_i x_j\big).
\end{aligned}$$

D'après 1) l'on sait que :

$$\begin{aligned}
\sum_{i=3}^{n} x_i^2 + \sum_{3\leqslant i<j\leqslant n} x_i x_j &= \sum_{k=1}^{n-2} \frac{k+1}{2k}\big(x_{k+2} + \frac{1}{k+1}(x_{k+3}+\cdots+x_n)\big)^2 \\
&= \sum_{k=3}^{n} \frac{k-1}{2(k-2)}\big(x_k + \frac{1}{k-1}(x_{k+1}+\cdots+x_n)\big)^2.
\end{aligned}$$

Ainsi :

$$\Psi(x) = \big(\frac{x_1+x_2}{2}+x_3+\cdots+x_n\big)^2 - \big(\frac{x_1-x_2}{2}\big)^2 - \sum_{k=3}^{n} \frac{k-1}{2(k-2)}\big(x_k + \frac{1}{k-1}(x_{k+1}+\cdots+x_n)\big)^2.$$

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

Exercices sur les sections 1 et 2

I.3.1 *Matrices magiques.* Si $A = (a_{ij}) \in M_n(K)$, pour tout $i \in [\![1, n]\!]$, on désigne par $l_i(A)$ la somme des termes de la $i^{ème}$ ligne de A, par $c_j(A)$ la somme des termes de sa $j^{ème}$ colonne et par $d(A)$ la somme des termes de son « anti-diagonale » : $d(A) = \sum_{k+h=n+1} a_{kh}$. On note :

$$\mathcal{M}_0 = \{A \in M_n(K) |\ l_i(A) = c_j(A) = 0\ \forall\, i, j \in [\![1, n]\!]\},$$
$$\mathcal{M}_1 = \{A \in M_n(K) |\ l_i(A) = c_j(A)\ \forall\, i, j \in [\![1, n]\!]\},$$
$$\mathcal{M}_2 = \{A \in M_n(K) |\ l_i(A) = c_j(A) = \operatorname{Tr} A = d(A) = 0\ \forall\, i, j \in [\![1, n]\!]\},$$
$$\mathcal{M}_3 = \{A \in M_n(K) |\ l_i(A) = c_j(A) = \operatorname{Tr} A = d(A)\ \forall\, i, j \in [\![1, n]\!]\}.$$

Montrer que $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ sont des sous–espaces vectoriels de $M_n(K)$ et calculer leurs dimensions respectives. Les éléments de $\mathcal{M}_3$ sont appelés *matrices magiques* (elles généralisent les *carrés magiques*).

I.3.2 Montrer que tout hyperplan du K-espace vectoriel $M_n(K)$ contient une matrice inversible.

I.3.3 $*$ Soit $\lambda \in K$. L'application $f : M_n(K) \to K$ définie par $f(A) := \lambda \operatorname{Tr} A$ est une forme linéaire sur le K-espace vectoriel $M_n(K)$ et vérifie, pour tous $A, B \in M_n(K)$: $f(AB) = f(BA)$.
a) Montrer qu'inversement si g est une forme linéaire sur $M_n(K)$ vérifiant $g(AB) = g(BA)$, pour tous $A, B \in M_n(K)$, il existe un scalaire $\mu \in K$ tel que $g = \mu \operatorname{Tr}$.
b) Montrer qu'une matrice $A \in M_n(K)$ s'écrit comme somme de matrices de la forme $BC - CB$ si, et seulement si, $\operatorname{Tr} A = 0$.
Remarque : en fait, par une autre méthode, on peut démontrer que, si $\operatorname{Tr} A = 0$, il existe $B, C \in M_n(K)$ telles que : $A = BC - CB$ (*cf.* l'exercice I.5.2 de la page 256).

I.3.4 $*$ Soient E, F, G des K-espaces vectoriels et $u : E \to F$, $v : F \to G$ des applications linéaires.
a) On considère la transposée ${}^t v : G^* \to F^*$ de v, montrer que $\operatorname{Ker} v \subset F$ et $\operatorname{Im} {}^t v \subset F^*$ sont orthogonaux.
b) On suppose dans toute la suite que F est *de dimension finie*. Montrer que $(\operatorname{Ker} v)^\perp = \operatorname{Im} {}^t v$ et que $(\operatorname{Im} {}^t v)^\circ = \operatorname{Ker} v$. Montrer que les applications v et ${}^t v$ sont toutes deux de rang fini et que ces rangs sont égaux.
c) On suppose que $v \circ u = 0$, montrer que : ${}^t u \circ {}^t v = 0$, $\operatorname{Im} u \subset \operatorname{Ker} v$ et $\operatorname{Im} {}^t v \subset \operatorname{Ker} {}^t u$. On note $H = \operatorname{Ker} v / \operatorname{Im} u$ et $H' = \operatorname{Ker} {}^t u / \operatorname{Im} {}^t v$ les espaces vectoriels quotients. Montrer que les dimensions de H et H' sont égales. Montrer que le crochet de dualité de F induit une forme bilinéaire φ sur $H \times H'$. Montrer que cette forme est non-dégénérée et en déduire un isomorphisme entre H' et H^*.

I.3.5 Soit $E = \mathbb{R}_n[X]$. Pour tout entier $k \in [\![0, n]\!]$ on définit une forme linéaire f_k sur E par $P \mapsto f_k(P) := \displaystyle\int_0^1 t^k P(t)\, \mathrm{d}t$.
a) Montrer que $(f_k)_{k \in [\![0,n]\!]}$ est une base du dual E^* de E.
b) Soit $\mathcal{B}$ la base canonique de E. Quelle est la matrice de passage de la base duale de $\mathcal{B}$ à la base $(f_k)_{k \in [\![0,n]\!]}$?

I.3.6 On définit les *polynômes de Newton* :

$$P_0(x) = 1 \qquad \text{et} \qquad \forall n \in \mathbb{N}^*,\ P_n(x) = x(x-1)\dots(x-n+1).$$

Soit $\Delta : K[X] \to K[X]$ l'application linéaire définie par $\Delta(P)(x) = P(x+1) - P(x)$.
a) Montrer que $\deg \Delta(P) = \deg P - 1$ pour tout polynôme P non constant. Déterminer $\operatorname{Ker} \Delta$.
Montrer que, pour tout $n > 0$: $\Delta\left(\dfrac{P_n}{n!}\right) = \dfrac{P_{n-1}}{(n-1)!}$.
b) Montrer que $\left(\dfrac{P_n}{n!}\right)_{n\in\mathbb{N}}$ est une base de $K[X]$.
Soit $P \in K_p[X]$, il s'écrit de manière unique $P = \displaystyle\sum_{n=0}^{p} a_n \frac{P_n}{n!}$.
Calculer les coefficients a_n en utilisant les polynômes $\Delta^n(P)$ (on peut penser à l'analogie entre la « différence finie » Δ et la dérivation $\mathrm{d}/\,\mathrm{d}x$ et à la formule de Taylor).
c) Le système $\left(\dfrac{P_n}{n!}\right)_{n\in[\![0,p]\!]}$ est une base de $K_p[X]$. Expliciter la base duale en utilisant Δ.

I.3.7 Soient $p, n \in \mathbb{N}^*$, $p > n$ et E un K-espace vectoriel de dimension n. Que peut-on dire d'une forme p-linéaire alternée sur E ?

Exercices sur les sections 3 et 4

I.3.8 Soient $a \in \mathbb{R}$ et Φ_a la forme quadratique définie sur $\mathbb{R}^3$ par :

$$\Phi_a(x_1, x_2, x_3) := a(x_1^2 + x_2^2 + x_3^2) - 2x_1x_2 + 2x_2x_3 + 2x_3x_1.$$

a) Pour quelles valeurs de a la forme Φ_a est-elle non dégénérée ?
b) Montrer qu'il existe une base de $\mathbb{R}^3$, indépendante de a, formée de vecteurs Φ_a-conjugués pour tout $a \in \mathbb{R}$.

I.3.9 Soit $n \geqslant 2$. On note $(x_1, \dots, x_n)$ les coordonnées de $x \in \mathbb{R}^n$ dans la base canonique. Déterminer le rang et la signature des formes quadratiques suivantes sur $\mathbb{R}^n$:

$$\Phi_1(x) := \sum_{i,j} (i+j-1)\, x_i x_j, \quad \Phi_4(x) := \sum_{i,j} \cosh((i+j)\alpha)\, x_i x_j,\ \alpha \in \mathbb{R},$$

$$\Phi_2(x) := \sum_{i,j} \inf(i,j)\, x_i x_j, \qquad \Phi_5(x) := \sum_{i,j} \sin(\theta_i + \theta_j)\ x_i x_j, \quad (\theta_1, \dots, \theta_n) \in \mathbb{R}^n,$$

$$\Phi_3(x) := \sum_{i<j} (x_i - x_j)^2, \qquad \Phi_6(x) := \sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} \alpha_i x_i\Big)^2, \quad (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n.$$

I.3.10 On considère la forme quadratique Φ sur $\mathbb{R}^n$ dont la matrice associée dans la base canonique est $A := (a_{ij})$, avec $a_{ij} := 2\inf(i,j) - 1$ si $i \neq j$ et $a_{ii} := i^2 + i - 1$.
Quelle est la signature de Φ ?

I.3.11 a) Soit $P \in \mathbb{R}[X]$, montrer que la série $\sum P(n)P(-n)e^{-n}$ converge.

b) Soit Φ l'application $P \in \mathbb{R}[X] \mapsto \sum_{n=0}^{+\infty} P(n)P(-n)e^{-n}$, vérifier que Φ est une forme quadratique sur $\mathbb{R}[X]$. Quelle est la signature de sa restriction à $\mathbb{R}_p[X]$?

I.3.12 Soient E un K-espace vectoriel et $\mathcal{B} = (e_1, e_2, \dots, e_n)$ une base de E. Soient f une forme linéaire non nulle sur E et Φ la forme quadratique définie sur E par $\Phi(x) := \big(f(x)\big)^2$. Déterminer la matrice associée à Φ : $M := \mathrm{Mat}(\Phi; \mathcal{B})$. Quel est son rang ?
Réciproquement, soit $M \in S_n(K)$, de rang 1, caractériser la forme quadratique Φ sur E telle que $M = \mathrm{Mat}(\Phi; \mathcal{B})$.

I.3.13 Soient $a_1, \dots, a_n$ des réels strictement positifs, deux à deux distincts. Montrer que la forme quadratique Φ qui à $(x_1, \dots, x_n) \in \mathbb{R}^n$ associe $\sum_{i=1}^{n}\sum_{j=1}^{n} \dfrac{x_i x_j}{a_i + a_j}$ est définie positive.

On pourra remarquer que, pour tout $a > 0$, $\displaystyle\int_0^\infty e^{-at}\,\mathrm{d}t = \frac{1}{a}$ ou raisonner par récurrence.

I.3.14 Soit $A := \begin{pmatrix} b_1 & c_1 & 0 & 0 & \dots & 0 \\ a_2 & b_2 & c_2 & 0 & \dots & 0 \\ 0 & a_3 & b_3 & c_3 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & \dots & 0 & 0 & a_n & b_n \end{pmatrix}$ une matrice de $M_n(K)$. On définit la suite $(d_k)_{k\in[\![0,n]\!]}$ par : $d_0 = 1$, $d_1 = b_1$, et $d_k = b_k d_{k-1} - a_k c_{k-1} d_{k-2}$ pour $k \in [\![2, n]\!]$.

a) Montrer que, pour tout $k \in [\![0, n]\!]$, d_k est égal au $k^{ème}$ mineur principal D_k de A.

b) On suppose, dans cette question et dans la suivante, que tous les mineurs principaux de A sont non nuls. On note :

$$L := \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ a_2\frac{d_0}{d_1} & 1 & 0 & \dots & 0 \\ 0 & a_3\frac{d_1}{d_2} & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & & a_n\frac{d_{n-2}}{d_{n-1}} & 1 \end{pmatrix} \quad \text{et} \quad U := \begin{pmatrix} \frac{d_1}{d_0} & c_1 & 0 & & \dots & 0 \\ 0 & \frac{d_2}{d_1} & c_2 & & \dots & 0 \\ 0 & 0 & \frac{d_3}{d_2} & c_3 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \frac{d_{n-1}}{d_{n-2}} & c_{n-1} \\ 0 & \dots & \dots & & 0 & \frac{d_n}{d_{n-1}} \end{pmatrix}.$$

Comparer A et LU.

c) On note $\Delta = \mathrm{Diag}\left(\dfrac{d_1}{d_0}, \dots, \dfrac{d_n}{d_{n-1}}\right)$. Montrer que l'on peut résoudre le système $AX = V$ (où x, v sont deux vecteurs de $\mathbb{R}^n$, v étant connu et x inconnu), en utilisant la décomposition $A = (L\Delta)(\Delta^{-1}U)$. Calculer en fonction de n le nombre d'additions, multiplications, divisions requises pour cette résolution.

I.3.15 Montrer, sans utiliser la décomposition LU, qu'une matrice de $S_n(\mathbb{R})$ dont tous les mineurs principaux sont strictement positifs appartient à $S_n^{++}(\mathbb{R})$.

Espaces préhilbertiens

I.4

1 Espaces vectoriels préhilbertiens réels Espaces euclidiens

Les notions de *distance* et *d'orthogonalité* sont à la base de la géométrie euclidienne (Euclide, mathématicien grec, III^{e} siècle avant J.C.). Les concepts initiaux ont évolué au cours des siècles ; aujourd'hui, les notions de distance et d'orthogonalité sont dérivées de celle de produit scalaire. Au tout début du XX^{e} siècle (vers 1901), il y a un tournant important, quand David Hilbert (mathématicien allemand, 1862-1943) jette les bases d'une *géométrie de l'infini*.

Si, de nos jours, ses idées sont présentées de façon abstraite, il est bon d'avoir à l'esprit ses motivations initiales. Son principal objectif était la résolution de problèmes fondamentaux, issus des mathématiques ou de la physique. Citons, au premier rang de ceux-ci, le problème de Dirichlet en théorie du potentiel, et plus généralement des questions étudiées un peu avant lui par Fredholm dans sa théorie des équations intégrales. D. Hilbert considère les fonctions comme des « points » d'un espace. Il introduit ainsi des espaces, que l'on appelle aujourd'hui *fonctionnels* et les munit d'une *géométrie*, généralisant subtilement celle d'Euclide. Il peut alors définir une notion de *forme quadratique en dimension infinie* et établir un analogue du théorème fondamental de décomposition en carrés que nous avons prouvé plus haut. Dans ce contexte, il découvre un outil essentiel, l'espace ℓ^2 des *suites de carré sommable*, le prototype de ce que l'on appelle aujourd'hui un *espace de Hilbert* (*cf.* module II.1). La systématisation de la notion d'espace de Hilbert est due à Schmidt et Fréchet (vers 1908) et surtout, bien plus tard (vers 1932), à von Neumann, en relation avec la *mécanique quantique* (plus précisément avec l'unification par Dirac de la « mécanique des matrices » de Heisenberg et de l'équation de Schrödinger). Dans ce module, nous nous limiterons aux aspects les plus algébriques de la théorie et aux résultats spécifiques de la dimension finie (suggérant seulement par quelques exemples les difficultés du cas général) ; on trouvera dans le module II.1 l'importante notion (introduite par Hilbert, puis développée par Banach) d'espace *complet* et dans le module II.4 une application à un problème fondamental de dimension infinie : la théorie des *séries de Fourier*. Les noyaux intégraux et la théorie de Fredholm sont étudiés dans [MPA], avec d'autres applications.

1.1 Produit scalaire, norme euclidienne

1.1.1 Produit scalaire

Définition 1. On appelle *produit scalaire* sur le $\mathbb{R}$-espace vectoriel E toute forme bilinéaire symétrique *définie positive*. On appelle *espace vectoriel préhilbertien réel* le couple formé d'un $\mathbb{R}$-espace vectoriel E et d'un produit scalaire sur E. Si la dimension de E est *finie*, l'espace vectoriel préhilbertien réel E est dit *euclidien*.

Pour abréger, nous dirons souvent *espace préhilbertien* au lieu d'espace vectoriel préhilbertien réel. Pour tous $x, y \in E$, on notera en général $(\, x \mid y \,)$ leur produit scalaire et $\big(E, (\,|\,)\big)$ l'espace vectoriel préhilbertien réel associé. Lorsqu'il n'y aura pas d'ambiguïté, on allégera la notation $\big(E, (\,|\,)\big)$ en E.

Exemples.

(i) L'espace vectoriel $E = \mathbb{R}^n$ muni du produit scalaire canonique est un espace euclidien. C'est la structure euclidienne canonique de $\mathbb{R}^n$.

(ii) Soit E l'ensemble des suites $x := (x_n)_{n\geqslant 0} \in \mathbb{R}^{\mathbb{N}}$, *de carré sommable*, c'est à dire telles que $\sum\limits_{n\geqslant 0} x_n^2 < +\infty$. Soient $x := (x_n), y := (y_n) \in E$, on a : $|x_n y_n| \leqslant \frac{1}{2}(x_n^2 + y_n^2)$, la série $\sum\limits_{n\geqslant 0} x_n y_n$ est donc absolument convergente. On vérifie que E est un sous-espace vectoriel de $\mathbb{R}^{\mathbb{N}}$ et que l'application $(x, y) \mapsto (\, x \mid y \,) := \sum\limits_{n=0}^{+\infty} x_n y_n$ est un produit scalaire sur E. (Pour tous $x, y \in E$ et $\alpha \in \mathbb{R}$, $\alpha x + y \in E$ car $(\alpha x_n + y_n)^2 = \alpha^2 x_n^2 + y_n^2 + 2\alpha x_n y_n$.) L'espace E muni du produit scalaire $(\,|\,)$est un espace préhilbertien, c'est, par définition, l'espace $\ell^2(\mathbb{R})$.

(iii) Soient $a, b \in \mathbb{R}$, avec $a < b$. On note $E := \mathcal{C}([a, b], \mathbb{R})$ le $\mathbb{R}$ espace vectoriel des fonctions continues sur $[a, b]$, à valeurs réelles. On définit $\varphi : E \times E \to \mathbb{R}$ par $(f, g) \mapsto \varphi(f, g) := \displaystyle\int_a^b f(t)g(t)\, \mathrm{d}t$. Cette application est bilinéaire et *définie positive* : la fonction f^2 est positive et continue sur $[a, b]$, donc, $\displaystyle\int_a^b f(t)^2\, \mathrm{d}t = 0$ si, et seulement si, $f = 0$.

Remarque. Étant donné un espace vectoriel réel E, on peut définir sur lui plusieurs produits scalaires. Par exemple :

- Sur $E = \mathbb{R}^2$, on peut définir le produit scalaire canonique $(\, x \mid y \,) = x_1 y_1 + x_2 y_2$ et le produit scalaire $< x, y >= 2x_1 y_1 + 3x_2 y_2 + x_1 y_2 + x_2 y_1$.
- En reprenant l'exercice 12 de la page 129, on voit qu'il existe sur $E := \mathbb{R}[X]$ une infinité de produits scalaires (dépendant du choix du segment $[a, b]$ et de celui de la fonction w).

Remarque. Soient $\big(E,(\,|\,)\big)$ un espace préhilbertien réel et F un sous-espace vectoriel de E. Alors $(\,|\,)$ définit par restriction un produit scalaire $(\,|\,)_{|F}$ sur F et $(F,(\,|\,)_{|F})$ est un espace préhilbertien réel.

1.1.2 Norme euclidienne

Nous allons étendre les notions de norme et de distance vues dans [L1], module IV.6.
Soit $\big(E,(\,|\,)\big)$ un espace préhilbertien réel. On note $\|x\| := \sqrt{(\,x\mid x\,)} = (\,x\mid x\,)^{1/2}$.

Proposition 1. Soit E un espace préhilbertien réel. On a :
(i) pour $x\in E$, $\|x\|=0$ si, et seulement si, $x=0$;
(ii) pour tous $(x,y)\in E^2$, $\|x+y\|\leqslant\|x\|+\|y\|$ (inégalité triangulaire) ;
(iii) pour tout $(\alpha,x)\in\mathbb{R}\times E$, $\|\alpha x\|=|\alpha|\|x\|$.

Démonstration. L'assertion (i) est vraie car le produit scalaire est *défini*. Pour (ii), on a $\|\alpha x\|^2=(\,\alpha x\mid\alpha x\,)=\alpha^2\|x\|^2$, d'où $\|\alpha x\|=|\alpha|\|x\|$. L'inégalité triangulaire (iii) est l'inégalité de Minkowski (corollaire 43 de la page 131). ∎

Ces trois propriétés caractérisent une *norme*. Nous dirons donc que $\|x\|$ est la norme euclidienne du vecteur $x\in E$. À toute norme on peut associer une *distance*. Ici c'est l'application $d : E\times E\to\mathbb{R}$ définie par $d(x,y) := \|x-y\| = \sqrt{\|x\|^2+\|y\|^2-2(\,x\mid y\,)}$. On l'appelle *distance euclidienne* de x à y. Pour $E=\mathbb{R}^n$ avec sa structure euclidienne canonique, cette distance est la longueur du vecteur joignant le « point » x au « point » y.

Proposition 2. Soit $\big(E,(\,|\,)\big)$ un espace préhilbertien. L'application distance euclidienne $d : E^2\to\mathbb{R}$ possède les propriétés suivantes :
(i) si $(x,y)\in E^2$ est tel que $d(x,y)=0$, alors $x=y$ (propriété de séparation) ;
(ii) pour tout $(x,y,z)\in E^3$, $d(x,z)\leqslant d(x,y)+d(y,z)$ (inégalité triangulaire) ;
(iii) pour tout $(\alpha,x,y)\in\mathbb{R}\times E^2$, $d(\alpha x,\alpha y)=|\alpha|d(x,y)$;
(iv) pour tout $(a,x,y)\in E^3$, $d(x+a,y+a)=d(x,y)$ (invariance par translation).

La preuve est laissée au lecteur.
Un vecteur de norme 1 de E sera dit *unitaire*. Pour $a\in E$, l'ensemble des vecteurs $x\in E$ tels que $d(a,x)<1$ (resp. $d(a,x)\leqslant 1$) forment la boule ouverte (resp. la boule fermée) de centre a et de rayon 1. C'est le point de départ de l'étude de ce que l'on appelle la topologie de E (*cf.* le module II.1).

1.1.3 Égalité de la médiane ou du parallélogramme

Soit E un espace préhilbertien réel. Pour tout $(x,y)\in E^2$, on a :

$$\|x+y\|^2=\|x\|^2+\|y\|^2+2(\,x\mid y\,)\qquad\text{et}\qquad\|x-y\|^2=\|x\|^2+\|y\|^2-2(\,x\mid y\,).$$

On en déduit :

$$\|x+y\|^2+\|x-y\|^2=2(\|x\|^2+\|y\|^2).$$

C'est l'égalité de la médiane ou du parallélogramme. Cette égalité est bien connue en géométrie plane élémentaire : dans ce cadre elle est due à Apollonius de Perge $(-262,\ -190)$. Soit $ABCD$ un parallélogramme de centre O (*cf.* figure I.4.1), on a :

$$AC^2 + DB^2 = 2(AB^2 + BC^2) \qquad \text{(égalité dite du parallélogramme),}$$

ou bien $AC^2 + 4BO^2 = 2(AB^2 + BC^2)$, soit :

$$\frac{AC^2}{2} + 2BO^2 = AB^2 + B^2 \qquad \text{(égalité dite de la médiane).}$$

Cette égalité caractérise en fait une norme euclidienne (*cf.* exercice I.4.23 de la page 211.)

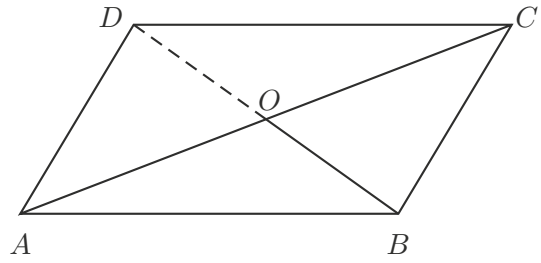


FIGURE I.4.1 – Formule de la médiane.

1.2 Orthogonalité, projecteurs orthogonaux, symétries orthogonales

1.2.1 Orthogonalité

Définition 2. Soit φ un produit scalaire sur l'espace vectoriel réel E. Deux vecteurs φ-conjugués de E seront dits *orthogonaux*. Si $(x_i)_{i\in I}$ est un système de vecteurs deux à deux orthogonaux de E, ce système est dit *orthogonal*. Si de plus chaque vecteur du système est de norme 1, le système est dit *orthonormé*.

Proposition 3. Soit E un espace préhilbertien. Un système orthogonal de E ne contenant pas 0 est libre.

Démonstration. Soit $(x_i)_{i\in I}$ un système orthogonal de E ne contenant pas 0.
Soit $(\lambda_i)_{i\in I} \in \mathbb{R}^I$ telle que $\sum\limits_{i\in I} \lambda_i x_i = 0$. Soit $j \in I$, on a :

$$0 = (\, x_j \mid \sum_{i\in I} \lambda_i x_i \,) = \sum_{i\in I} \lambda_i (\, x_j \mid x_i \,) = \lambda_j \|x_j\|^2,$$

d'où $\lambda_j = 0$. On en déduit que les réels λ_i $(i \in I)$ sont tous nuls, le système est donc libre. ∎

Du corollaire 45 de la page 132, on déduit immédiatement le résultat essentiel suivant :

Théorème 4. Tout espace vectoriel euclidien admet une base orthonormée.

Remarque. Tout espace vectoriel euclidien admet une base orthonormée. Par contre un espace préhilbertien de dimension infinie n'admet pas toujours une base orthonormée (on peut montrer, en utilisant la notion de suite de Cauchy, que $\ell^2(\mathbb{R})$ n'admet pas de base orthonormée). L'espace vectoriel réel $E := \mathbb{R}[X]$, de dimension infinie, muni d'un produit scalaire, comme dans l'exercice 12 de la page 129, admet une base orthonormée (*cf.* plus loin page 157).

Proposition 5.
Soit E un espace préhilbertien admettant une base orthonormée $(e_i)_{i\in I}$. Si $x, y \in E$, on note $x := \sum_{i\in I} x_i e_i$, $y := \sum_{i\in I} y_i e_i$. On a, pour tout $i \in I$, $x_i = (\, e_i \mid x\,)$, d'où $x = \sum_{i\in I} (\, e_i \mid x\,) e_i$. On a aussi $(\, x \mid y\,) = \sum_{i\in I} x_i y_i$, $\|x\|^2 = (\, x \mid x\,) = \sum_{i\in I} x_i^2$.

On généralise le célèbre théorème de Pythagore de la géométrie élémentaire :

Proposition 6 (Théorème de Pythagore).
Soient E un espace préhilbertien et $x, y \in E$. Les vecteurs x et y sont orthogonaux si, et seulement si, $\|x+y\|^2 = \|x\|^2 + \|y\|^2$. Si $x_1, \ldots, x_k$ sont des vecteurs deux à deux orthogonaux de E, on a $\|x_1 + \cdots + x_k\|^2 = \|x_1\|^2 + \cdots + \|x_k\|^2$.

Démonstration. On a $\|x+y\|^2 = (\, x+y \mid x+y\,) = \|x\|^2 + \|y\|^2 + 2(\, x \mid y\,) = \|x\|^2 + \|y\|^2$, le résultat s'en déduit. Le lecteur traitera le cas de k vecteurs ($k > 2$). ∎

Définition 3. Soit E un espace préhilbertien.
Pour tout *sous-ensemble* $A \subset E$, on appelle orthogonal de A, et l'on note $A^\perp$ le sous-ensemble $\{x \in E \mid (\, x \mid a\,) = 0\,,\ \forall\, a \in A\}$.

Proposition 7. Soit E un espace préhilbertien. Pour tout *sous-ensemble* $A \subset E$, $A^\perp$ est un *sous-espace vectoriel* de E égal à $(\operatorname{Vect} A)^\perp$. On a $E^\perp = \{0\}$.

Démonstration. La première partie de la proposition a été prouvée dans un cadre plus général (*cf.* la proposition 28 de la page 114). Montrons $E^\perp = \{0\}$, si $x \in E^\perp$, le vecteur x est orthogonal à lui-même, donc $\|x\|^2 = (\, x \mid x\,) = 0$ et $x = 0$. ∎

Remarque. Pour montrer qu'un vecteur x d'un espace préhilbertien E est nul, on utilise souvent le critère : $x = 0$ si, et seulement si, $(\, x \mid y\,) = 0$ pour tout vecteur $y \in E$.

Définition 4. Soit E un espace préhilbertien. Deux sous-espaces vectoriels $F, G \subset E$ sont dits *orthogonaux*, si pour tout $(x, y) \in F \times G : (x \mid y) = 0$. Deux sous-ensembles $A, B \subset E$ sont dits *perpendiculaires* si les sous-espaces $A^\perp$ et $B^\perp$ sont orthogonaux.

En utilisant la proposition 6, on montre le résultat suivant.

Proposition et définition 8. Soient E un espace préhilbertien et $E_1, \ldots, E_p$ des sous-espaces vectoriels de E deux à deux orthogonaux.

Alors leur somme $F = E_1 + \cdots + E_p$ est *directe*, on note $F = E_1 \overset{\perp}{\oplus} \ldots \overset{\perp}{\oplus} E_p$, et l'on dit que F est la *somme directe orthogonale* des sous-espaces $E_1, \ldots, E_p$.

1.2.2 Supplémentaire orthogonal

Soit E un espace préhilbertien. Si F est un sous-espace de E, on a évidemment $F \cap F^\perp = \{0\}$. Peut-on dire, par analogie avec des situations géométriques élémentaires, que $F^\perp$ est un sous-espace supplémentaire de F dans E ? Nous allons voir que la réponse est positive si E est euclidien, par contre, si E est de dimension infinie, il est possible que $F \overset{\perp}{\oplus} F^\perp \neq E$.

Exemple. Soit F le sous-espace vectoriel de $E = \ell^2(\mathbb{R})$ formé des suites *à support fini*, on note $e_n \in \ell^2(\mathbb{R})$ la suite $(0, \ldots, 0, 1, 0, \ldots) \in \ell^2(\mathbb{R})$ dont tous les termes sont nuls sauf le $(n+1)^{ème}$ que l'on suppose égal à 1, le système $(e_n)_{n \in N}$ est une base orthonormée de F. Soit $x = (x_n) \in \ell^2(\mathbb{R})$ un vecteur orthogonal à F, $(x \mid \mathrm{e}_n) = x_n = 0$, pour tout $n \in \mathbb{N}$, d'où $x = 0$. On a ainsi $F^\perp = \{0\}$, donc $F + F^\perp = F \neq E$.

Exercice 1.

Soit $E := \mathcal{C}([-1, 1]; \mathbb{R})$ le $\mathbb{R}$-espace vectoriel des fonctions continues à valeurs réelles, sur le segment $[-1, 1]$, muni du produit scalaire $(f \mid g) := \displaystyle\int_{-1}^{1} f(t)g(t)\,\mathrm{d}t$. C'est un espace vectoriel préhilbertien réel. Soit $F \subset E$ le sous-espace des fonctions nulles sur le segment $[-1, 0]$. Expliciter $F^\perp$. En déduire que $F + F^\perp \neq E$.

Solution. Soit $f \in E$ une fonction nulle sur le segment $[0, 1]$. Elle est clairement orthogonale à F. Inversement, soit $f \in E$ orthogonale à F. On définit une fonction g sur le segment $[-1, 1]$ par : $g(t) := 0$ si $t \in [-1, 0]$ et $g(t) := tf(t)$ si $t \in]0, 1]$. Cette fonction est continue, donc $g \in F$. On en déduit $(f \mid g) = \displaystyle\int_{-1}^{1} f(t)g(t)\,\mathrm{d}t = \int_{0}^{1} tf(t)^2\,\mathrm{d}t = 0$.

La fonction $t \mapsto tf(t)^2$ est continue et positive sur le segment $[0, 1]$, elle est donc nulle sur ce segment ; par suite la fonction f est aussi nulle sur $]0, 1]$ et, par continuité, elle est nulle sur $[0, 1]$. Ainsi $F^\perp$ est l'espace des fonctions de E qui sont nulles sur le segment $[0, 1]$. On vérifie facilement que $F + F^\perp$ est le sous-espace de E des fonctions nulles en 0, c'est un hyperplan de E et $F + F^\perp \neq E$.

Théorème et définition 9. Soient E un espace préhilbertien et F un sous-espace vectoriel de E. S'il existe un sous-espace vectoriel $G \subset E$ tel que $E = F \overset{\perp}{\oplus} G$, alors $G = F^\perp$; on l'appelle *le supplémentaire orthogonal de* F.

Démonstration. Supposons que $E = F \overset{\perp}{\oplus} G$. Alors $G \subset F^\perp$.
Inversement, soit $x \in F^\perp$, il existe $(y, z) \in F \times G$ tel que $x = y + z$, alors $y = x - z \in F \cap F^\perp$, donc $y = 0$ et $F^\perp \subset G$. ∎

Supposons maintenant que E soit un espace *euclidien*, on a alors un résultat *essentiel*.

Théorème 10 (Toute forme linéaire sur un espace euclidien est un produit scalaire). Soit E un espace vectoriel euclidien :
(i) Pour toute forme linéaire f sur E, il *existe* un *unique* vecteur $a \in E$ tel que, pour tout $x \in E$, $f(x) = (\,a \mid x\,)$.
(ii) L'application $a \mapsto a^\flat$ de E dans E^* qui associe au vecteur $a \in E$, la forme linéaire $a^\flat : x \mapsto (\,a \mid x\,)$ est un *isomorphisme* d'espaces vectoriels réels.

Démonstration. L'assertion (i) résulte de l'assertion (ii). Montrons celle-ci. L'application $\flat : a \mapsto a^\flat$ est linéaire : pour tout $(\lambda, a_1, a_2) \in \mathbb{R} \times E^2$, $(\lambda a_1 + a_2)^\flat = \lambda a_1^\flat + a_2^\flat$. Elle est de plus injective : $a^\flat = 0$ si, et seulement si, $(\,a \mid x\,) = 0$ quel que soit $x \in E$, c'est à dire si, et seulement si, a est orthogonal à E, ce qui équivaut à $a = 0$. Par ailleurs $\dim E = \dim E^*$, l'application linéaire injective $\flat : E \to E^*$ est donc un isomorphisme de E sur E^*. L'application inverse $E^* \to E$ sera notée $\sharp : f \mapsto f^\sharp$. ∎

On prendra garde au fait que ce résultat n'est pas vrai en général pour un espace préhilbertien de dimension infinie (*cf.* plus loin, en 1.5.1, l'exemple de la page 162 et aussi l'exercice I.4.12 de la page 209).

Remarque. Les isomorphismes $\flat : E \to E^*$ et $\sharp : E^* \to E$ sont dits *canoniques*. Ils sont indépendants de tout choix de base.

Exemple. Soit E un espace euclidien réel de dimension $n > 0$. Soit $B := B(a; r)$ la boule ouverte de centre $a \in E$ et de rayon $r > 0$ de E. Soit $f : B \to \mathbb{R}$ une fonction différentiable en a. La différentielle $df(a)$ est une forme linéaire sur E. On *définit* un *vecteur* $\overrightarrow{\operatorname{grad}} f(a)$ de E (le *gradient* de la fonction f en a) par $\overrightarrow{\operatorname{grad}} f(a) := \sharp(df(a))$. En utilisant une base orthonormée , un vecteur $h \in E$ se représente par ses coordonnées $(h_1, \ldots, h_n) \in \mathbb{R}^n$, on a :

$$df(a)(h) = \frac{\partial f}{\partial x_1}(a)h_1 + \cdots + \frac{\partial f}{\partial x_n}(a)h_n = (\,\overrightarrow{\operatorname{grad}} f(a) \mid h\,),$$

donc $\overrightarrow{\operatorname{grad}} f(a) = \left(\dfrac{\partial f}{\partial x_1}(a), \ldots, \dfrac{\partial f}{\partial x_n}(a)\right)$. Si f est différentiable sur B, on peut faire la même chose en tout point de B et l'on obtient, à partir de la différentielle df, une fonction vectorielle : $\overrightarrow{\operatorname{grad}}\ f$. En mécanique classique ou en physique, on obtient ainsi un champ de *forces* $\overrightarrow{F} := -\overrightarrow{\operatorname{grad}} V$ à partir d'un *potentiel* V. On prendra bien garde au fait que

le potentiel V étant donné, la force $\overrightarrow{F}$ ne prend de sens qu'après la donnée d'un produit scalaire sur l'espace E (physiquement d'une *métrique*).
Voyons ce qui se passe quand on calcule avec des coordonnées.
Soient $\mathcal{B} := (e_1, \dots, e_n)$ une base de l'espace euclidien E et $\mathcal{B}^* := (e_1^*, \dots, e_n^*)$ la base duale. Notons $A := \big((\, e_i \mid e_j \,)\big)_{i,j\in[\![1,n]\!]}$ la matrice du produit scalaire de E dans la base $\mathcal{B}$. Soit $x = \sum_{j=1}^{n} x_j\, e_j \in E$. On a $\mathrm{e}_i^\flat(x) = (\, e_i \mid x \,) = \sum_{j=1}^{n} (\, e_i \mid e_j \,)\, x_j$, donc $\flat(e_i) = \mathrm{e}_i^\flat = \sum_{j=1}^{n} (\, e_i \mid e_j \,)\, e_j^*$. Ainsi A est la matrice de l'application linéaire $\flat : E \to E^*$ dans les bases $\mathcal{B}$ et $\mathcal{B}^*$. Par suite la matrice A^{-1} est la matrice de l'application $\sharp$ dans les bases $\mathcal{B}^*$ et $\mathcal{B}$. Ainsi, si $f \in E^*$ est une forme linéaire, donnée dans la base duale (e_i^*), pour calculer le vecteur $f^\sharp$ on utilisera en général la matrice A^{-1}, ou l'inversion du système linéaire correspondant.
Si la base $\mathcal{B}$ est orthonormée, on a $A = I$ et $e_i^\flat = e_i^*$, pour tout $i \in [\![1, n]\!]$: *l'image par* $\flat$ *d'une base orthonormée en est la base duale*. Dans ce cas le calcul de $f^\sharp$ est trivial.

Exercice 2.
Soit $\mathcal{B} = (b_1, b_2, b_3)$ une base de $E = \mathbb{R}^3$ muni de son produit scalaire usuel. Soit f la forme linéaire sur $\mathbb{R}^3$ qui à tout $x := x_1b_1 + x_2b_2 + x_3b_3 \in \mathbb{R}^3$ associe $3x_1 - x_2 + 2x_3 \in \mathbb{R}$.
(i) On suppose d'abord que $\mathcal{B}$ est la base canonique (e_1, e_2, e_3) de $\mathbb{R}^3$. Déterminer le vecteur $f^\sharp \in \mathbb{R}^3$.
(ii) On suppose maintenant que $b_1 := e_1 - e_2$, $b_2 := e_1 + e_2 - e_3$, $b_3 := e_1 - 2e_3$. Déterminer le vecteur $f^\sharp \in \mathbb{R}^3$.
Solution. (i) La base $\mathcal{B}$ étant orthonormée, $f(x) := 3x_1 - x_2 + 2x_3 = (\, a \mid x \,)$ si, et seulement si, $a = f^\sharp = 3e_1 - e_2 + 2e_3$.
(ii) On cherche $a := \alpha_1 b_1 + \alpha_2 b_2 + \alpha_3 b_3 \in \mathbb{R}^3$ tel que, pour tout $x \in E$, l'on ait $(\, a \mid x \,) = f(x)$. Ceci est réalisé si, et seulement si, $(\, a \mid b_1 \,) = 3$, $(\, a \mid b_2 \,) = -1$, $(\, a \mid b_3 \,) = 2$, ce qui s'écrit sous forme de système linéaire : $\begin{cases} 2\alpha_1 + \alpha_3 = 3 \\ 3\alpha_2 + 3\alpha_3 = -1 \\ \alpha_1 + 3\alpha_2 + 5\alpha_3 = 2. \end{cases}$
L'unique solution de ce système est $\alpha_1 = 1$, $\alpha_2 = -4/3$, $\alpha_3 = 1$. Ainsi $f^\sharp = a = b_1 - \frac{4}{3}b_2 + b_3$.

1.2.3 Coordonnées covariantes et coordonnées contravariantes

Étant donné un vecteur $x \in E$, il est déterminé par ses coordonnées dans une base $\mathcal{B} = (e_1, \dots, e_n)$. Conformément aux conventions des physiciens, nous les notons (x^i) et les appelons *coordonnées contravariantes*. Par ailleurs, $x^\flat(e_i) = (\, x \mid e_i \,)$, donc $x^\flat = \sum_{i=1}^{n} (\, x \mid e_i \,) e_i^*$, les réels $(\, x \mid e_i \,)$ sont donc les *coordonnées* de la *forme linéaire* $x^\flat$ dans la base duale $\mathcal{B}^*$, ces nombres déterminent le *vecteur* $x = \sharp(x^\flat)$, on dit que ce sont les coordonnées *covariantes* du *vecteur* x ; on les note (x_i). Par exemple, si l'on revient à l'exercice 2 (partie (ii)), le vecteur x de $E = \mathbb{R}^3$ de coordonnées contravariantes $(x^1 = 1, x^2 = -4/3, x^3 = 1)$, admet pour coordonnées covariantes $(x_1 = 3, x_2 = -1, x_3 = 2)$. On a $\flat(x) = \flat(\sum_{i=1}^{n} x^i e_i) = \sum_{i=1}^{n} x_i e_i^*$: par analogie avec la notation musicale $\flat$ baisse les indices et $\sharp$ les monte. On a la formule $\|x\|^2 = \sum_{i=1}^{n} x^i x_i$. Si la base $\mathcal{B}$

est orthonormée, on a $x^i = x_i$ pour tout $i \in [\![1, n]\!]$. En physique ou en géométrie différentielle on note souvent $G = (g_{ij})$, et on appelle *tenseur métrique*, la matrice symétrique $\big((\, e_i \mid e_j \,)\big)_{i,j\in[\![1,n]\!]}$, et $G^{-1} = (g^{ij})$ la matrice inverse. Les indices « en haut » (resp. « en bas ») sont dits contravariants (resp. covariants). Les formules de passage entre coordonnées contravariantes et coordonnées covariantes s'écrivent : $x_i = \sum g_{ij}x^j$, $x^i = \sum g^{ij}x_j$. En physique, on omet souvent le signe $\sum$, et l'on somme *implicitement* sur les paires d'indices identiques figurant à la fois en haut en en bas, suivant une convention due à A. Einstein ; on écrit par exemple : $x_i = g_{ij}x^j$, $(\, x \mid y \,) = x^i y_i = x_i y^i$.

Écrivons les coordonnées *contravariantes* x^i d'un vecteur $x \in E$ sous forme de *vecteur colonne* X et ses coordonnées *covariantes* x_i sous forme d'un *vecteur ligne* $X^\flat$, on a $X^\flat = {}^tXG = {}^t(GX)$. Regardons ce qui se passe si l'on fait un changement de base dans E, passant de la base $\mathcal{B}$ à la base $\mathcal{B}' := (e'_1, \dots, e'_n)$ par la matrice de changement de base $P \in GL_n(\mathbb{R})$, on a (avec des notations évidentes) les formules de passage pour les deux systèmes de coordonnées : $X' = P^{-1}X$ et $X'^\flat = X^\flat P$ (en effet : $G' = {}^tPGP$ et $X'^\flat = {}^t(G'X') = {}^t({}^tPGX) = {}^t(GX)P = X^\flat P$). La terminologie *contravariantes* vs *covariantes* provient de la structure de ces formules de passage (P^{-1} vs P). En notant $P := (p^j_i)$ et $P^{-1} := (q^i_j)$, on a, en utilisant la convention d'Einstein, les formules correspondantes :

$$\begin{aligned} e'_j &= p^i_j e_i & \qquad e_i &= q^j_i e'_j \\ x^j &= p^j_i x'^i & \qquad x'^i &= q^i_j x^j \\ x'_j &= p^i_j x_i & \qquad x_i &= q^j_i x'_j. \end{aligned}$$

Les coordonnées covariantes se transforment « comme les vecteurs de base ».

Théorème 11. Soit E un espace vectoriel *euclidien*. Pour tout sous-espace vectoriel F de E, on a :

$$E = F \overset{\perp}{\oplus} F^\perp, \quad \dim F + \dim F^\perp = \dim E \quad \text{et} \quad (F^\perp)^\perp = F.$$

Démonstration. On prendra bien garde dans cette démonstration à l'utilisation de *deux notions distinctes d'orthogonalité* : l'orthogonalité dans l'espace euclidien E et l'orthogonalité entre E et son dual E^* (*cf.* définition 3 de la page 98). Pour éviter les confusions nous adopterons donc ici (et seulement ici) deux notations différentes : l'orthogonalité dans E sera notée par $\perp$, comme d'habitude ; celle entre E et son dual E^* par $\dagger$. Notons $\dim E = n$, soit $a \in E$, $a \in F^\perp$ si, et seulement si, $a^\flat(x) = 0$ quel que soit $x \in F$. Ainsi $a \in F^\perp$ si, et seulement si, $a^\flat \in F^\dagger$ ($F^\dagger \subset E^*$ est l'orthogonal de F au sens de l'orthogonalité entre E et E^*). L'isomorphisme $\flat$ induit ainsi un isomorphisme entre les deux orthogonaux $F^\perp$ et $F^\dagger$. On a $\dim F^\dagger = n - \dim F$ donc $\dim F^\perp = n - \dim F$ et $\dim F + \dim F^\perp = \dim E$, d'où $F + F^\perp = E$ et $F \overset{\perp}{\oplus} F^\perp = E$.

On a évidemment $F \subset (F^\perp)^\perp$ et $\dim F = \dim (F^\perp)^\perp$ donc $F = (F^\perp)^\perp$. ∎

Corollaire 12. Soient E un espace vectoriel *euclidien* et F un sous-espace vectoriel de E, il existe une base orthonormée de E formée de la réunion d'une base orthonormée de F et d'une base orthonormée de $F^\perp$.

1.2.4 Projecteurs orthogonaux, symétries orthogonales

Rappelons qu'un *projecteur* de l'espace vectoriel E est un endomorphisme p de E tel que $p^2 = p$ et qu'une *symétrie* de E est un endomorphisme involutif $s : s^2 = \mathrm{Id}_E$ (*cf.* [L1], modules II.3 et II.4).

Étant donné un projecteur p, on lui associe la symétrie $s = 2p - \mathrm{Id}$. Inversement s étant une symétrie on lui associe le projecteur $p = \dfrac{s + \mathrm{Id}}{2}$. On sait que $\mathrm{Ker}\, p = \mathrm{Ker}(s + \mathrm{Id})$, $\mathrm{Im}\, p = \mathrm{Ker}(s - \mathrm{Id})$ et $E = \mathrm{Ker}\, p \oplus \mathrm{Im}\, p = \mathrm{Ker}(s + \mathrm{Id}) \oplus \mathrm{Ker}(s - \mathrm{Id})$. L'axe de la symétrie s est l'image du projecteur p et la direction de s est le noyau de p. Une symétrie de E est dite *hyperplane* si, et seulement si, son axe est un hyperplan de E.

Définition 5. Un *projecteur* p de l'espace préhilbertien réel E est dit *orthogonal* si, et seulement si, son image et son noyau sont orthogonaux. Dans ces conditions on dira que p est le projecteur orthogonal de E sur le sous-espace $F = \mathrm{Im}\, p$. De même une symétrie s de E est dite *orthogonale* si, et seulement si, $\mathrm{Ker}(s - \mathrm{Id})$ et $\mathrm{Ker}(s + \mathrm{Id})$ sont orthogonaux.

Remarques. Soit p un un projecteur de E.

(i) Si p est un projecteur orthogonal : $E = \mathrm{Ker}\, p \overset{\perp}{\oplus} \mathrm{Im}\, p$.

(ii) La symétrie $s := 2p - \mathrm{Id}$ est orthogonale si, et seulement si, p l'est. (On a $\mathrm{Ker}(s - \mathrm{Id}) = \mathrm{Im}\, p$ et $\mathrm{Ker}(s + \mathrm{Id}) = \mathrm{Ker}\, p$.)

(iii) L' application $q := \mathrm{Id} - p$ est un projecteur. Il est orthogonal si, et seulement si, p l'est : $\mathrm{Ker}\, q = \mathrm{Im}\, p$ et $\mathrm{Ker}\, p = \mathrm{Im}\, q$.

(iv) Si p est un projecteur orthogonal, le parallélogramme de sommets $0, x, x + s(x), s(x)$ est un *losange* (*cf.* figure I.4.2).

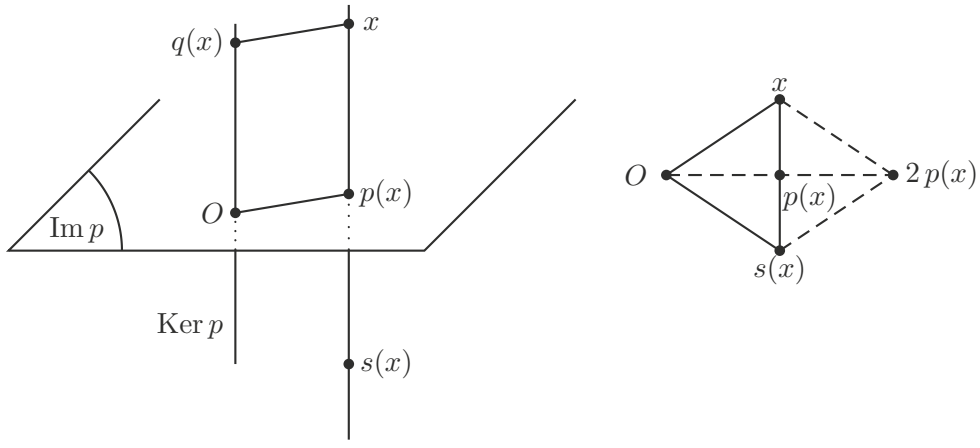


FIGURE I.4.2 – Projection et symétrie orthogonale : $x + s(x) = 2p(x)$ et $x = q(x) + p(x)$.

Exercice 3.
Montrer qu'un *projecteur* non nul de l'espace préhilbertien réel E est orthogonal si, et seulement si, $|||p||| := \sup_{\|x\|\leqslant 1} \|p(x)\| = 1$.

Solution.
– Supposons p orthogonal.
Pour tout $x \in E$, les vecteurs $x - p(x) \in \operatorname{Ker} p$ et $p(x) \in \operatorname{Im} p$ sont orthogonaux, donc, d'après le théorème de Pythagore (proposition 6 de la page 145) :

$$\|x\|^2 = \|x - p(x)\|^2 + \|p(x)\|^2 \quad \text{d'où} \quad \|p(x)\|^2 \leqslant \|x\|^2.$$

L'égalité est atteinte si $x \in \operatorname{Im} p$ ($\operatorname{Im} p \neq \{0\}$).
– Supposons $|||p||| = 1$.
Pour tous $(x, y) \in \operatorname{Ker} p \times \operatorname{Im} p$ et $t \in \mathbb{R}$, on a alors $\|p(x + ty)\|^2 \leqslant \|x + ty\|^2$, donc $t^2\|y\|^2 \leqslant \|x + ty\|^2 = \|x\|^2 + t^2\|y\|^2 + 2t(\,x \mid y\,)$ et $0 \leqslant \|x\|^2 + 2t(\,x \mid y\,)$. Le réel t étant arbitraire, cela impose $(\,x \mid y\,) = 0$.
Les sous-espaces $\operatorname{Ker} p$ et $\operatorname{Im} p$ sont donc orthogonaux et le projecteur p est orthogonal.

1.2.5 Calcul du projecteur orthogonal sur un sous-espace d'un espace euclidien

Soient E un espace euclidien et F un sous-espace vectoriel de E, la somme directe orthogonale $E = F \overset{\perp}{\oplus} F^\perp$ (*cf.* la proposition 11 de la page 149) permet de définir un projecteur orthogonal p sur F, de noyau $F^\perp$, on va le calculer.

Proposition 13. Soient E un espace euclidien et $F \subset E$ un sous-espace vectoriel de dimension $k > 0$, on désigne par p le projecteur orthogonal de E sur F. Soit $(\varepsilon_1, \ldots, \varepsilon_k)$ une base *orthonormée* de F, pour tout $x \in E$, on a :

$$p(x) = \sum_{h=1}^{k} (\,\varepsilon_h \mid x\,)\varepsilon_h.$$

Démonstration. Soit $x \in E$.

On sait qu'il existe $(\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$ tel que $p(x) = \sum_{h=1}^{k} \alpha_h \varepsilon_h$. Comme $x - p(x) \in \operatorname{Ker} p$, pour tout $j \in [\![1, k]\!]$, on a $\big(\,\varepsilon_j \;\big|\; x - p(x)\,\big) = 0$, ainsi :

$$(\,\varepsilon_j \mid x\,) = \big(\,\varepsilon_j \;\big|\; p(x)\,\big) = (\,\varepsilon_j \mid \sum_{h=1}^{k} \alpha_h \varepsilon_h\,) = \sum_{h=1}^{k} \alpha_h (\,\varepsilon_j \mid \varepsilon_h\,) = \alpha_j. \qquad \blacksquare$$

Remarque. Supposons donnée une base quelconque (non nécessairement orthonormée) $\mathcal{B} = (v_1, \ldots, v_k)$ de F. Pour tout $x \in E$, $p(x)$ est déterminé par les conditions :

pour tout $h \in [\![1,k]\!]$, $(v_h \mid x) = (v_h \mid p(x))$. Donc si $p(x) = \sum_{h=1}^{k} \alpha_h v_h$, les réels $\alpha_1, \ldots, \alpha_k$ sont solutions du système linéaire des k équations :

$$\sum_{h=1}^{k} \alpha_h (v_j \mid v_h) = (v_j \mid x) ; \quad j \in [\![1,k]\!].$$

On remarque que la matrice du système est la *matrice associée*, dans la base $\mathcal{B}$, à la restriction à F du produit scalaire sur E. Elle est *définie positive*, donc inversible. Le système admet donc une *unique solution* (théorème de Cramer).

1.3 Distance d'un point à un sous-espace de dimension finie, inégalité de Bessel

1.3.1 Distance à un sous-espace vectoriel

Définition 6. Soient E un espace préhilbertien et $A \subset E$ un sous-ensemble *non vide*. Soit $a \in E$, on appelle distance de a à A et l'on note $d(a,A)$ le nombre réel positif $d(a,A) := \inf\{d(a,x) \mid x \in A\}$.

Cette définition est licite car l'ensemble de nombres réels $\{d(a,x) | x \in A\}$ est non vide et minoré par 0.

Théorème 14. Soient E un espace préhilbertien et F un sous-espace vectoriel de E, *de dimension finie*, le sous-espace F admet un *supplémentaire orthogonal* et, pour tout $a \in E$, $d(a,F) = \|a - p(a)\|$, p étant le projecteur orthogonal sur F ; de plus, pour tout $b \in F$ tel que $b \neq p(a)$, on a $d(a,F) < d(a,b)$.

Démonstration. On raisonne par analogie avec la proposition 13 de la page précédente. Soit $(\varepsilon_1, \ldots, \varepsilon_k)$ une base *orthonormée* de F. Soit $p : E \to F$ l'application qui à tout $x \in E$ associe $p(x) := \sum_{h=1}^{k} (\varepsilon_h \mid x)\, \varepsilon_h$. On vérifie facilement que p est linéaire, puis que $p^2 = p$: p est un projecteur de E. On a $\operatorname{Ker} p = \{\varepsilon_1, \ldots, \varepsilon_k\}^\perp = F^\perp$ et $\operatorname{Im} p = F$, p est donc orthogonal et $E = F \overset{\perp}{\oplus} F^\perp$. Soit $a \in E$. Pour tout $b \in F$, $a - b = a - p(a) + p(a) - b$, d'où $\|a-b\|^2 = \|a-p(a)\|^2 + \|p(a)-b\|^2$ (théorème de Pythagore : proposition 6 de la page 145), et si $b \neq p(a)$:

$$d(a,F)^2 = \|a - p(a)\|^2 < \|a - b\|^2. \qquad \blacksquare$$

Remarque. Pour calculer $d(a,F)$, le calcul est souvent plus rapide si l'on remarque que $\|a - p(a)\|^2 = (a - p(a) \mid a)$.

1.3.2 Inégalité de Bessel

Corollaire 15 (Inégalité de Bessel). Soient E un espace préhilbertien, $F \subset E$ un sous-espace vectoriel de E de dimension *finie* $k \geqslant 1$ et $(\varepsilon_1, \ldots, \varepsilon_k)$ une base *orthonormée* de F. Soit p le projecteur orthogonal sur F.

Pour tout $x \in E$, on a $\|x\|^2 \geqslant \big\|p(x)\big\|^2 = \displaystyle\sum_{h=1}^{k} (\,\varepsilon_h \mid x\,)^2$.

Exercice 4.

Calculer $\quad \delta = \displaystyle\inf_{(a_1,a_2)\in\mathbb{R}^2} \int_0^1 (1 + a_1 t + a_2 t^2)^2 \, \mathrm{d}t$

Solution.

Le $\mathbb{R}$-espace vectoriel $\mathbb{R}_2[T]$ muni du produit scalaire $(\,P \mid Q\,) := \displaystyle\int_0^1 P(t)Q(t)\, \mathrm{d}t$ est un espace euclidien de dimension 3. On a :

$$\int_0^1 (1 + a_1 t + a_2 t^2)^2 \, \mathrm{d}t = \|1 + a_1 t + a_2 t^2\|^2 = \|1 - P(t)^2\|\,,$$

en posant $P(t) := -a_1 t - a_2 t^2$. Donc $\delta = \displaystyle\inf_{(a_1,a_2)\in\mathbb{R}^2} \int_0^1 (1 + a_1 t + a_2 t^2)^2 \, \mathrm{d}t = d^2(1, F)$, $F \subset \mathbb{R}_2[T]$ étant le sous-espace vectoriel $\mathrm{Vect}\{t, t^2\}$ de $\mathbb{R}_2[T]$. On est donc ramené à déterminer le projeté orthogonal $p(1)$ de 1 sur F. Notons $p(1) := b_1 t + b_2 t^2$. Les réels b_1 et b_2 sont solutions du système linéaire : $\begin{cases} \big(\,1 - p(1) \,\big|\, t\,\big) \;=\; 0 \\ \big(\,1 - p(1) \,\big|\, t^2\,\big) \;=\; 0. \end{cases}$

Soit :

$$\begin{cases} (\,1 \mid t\,) \;=\; b_1(\,t \mid t\,) + b_2(\,t^2 \mid t\,) \\ (\,1 \mid t^2\,) \;=\; b_1(\,t \mid t^2\,) + b_2(\,t^2 \mid t^2\,). \end{cases}$$

On a $(\,1 \mid t\,) = 1/2$, $(\,t \mid t\,) = (\,1 \mid t^2\,) = 1/3$, $(\,t \mid t^2\,) = 1/4$ et $(\,t^2 \mid t^2\,) = 1/5$. D'où le système : $\begin{cases} \dfrac{1}{2} \;=\; \dfrac{b_1}{3} + \dfrac{b_2}{4} \\ \dfrac{1}{3} \;=\; \dfrac{b_1}{4} + \dfrac{b_2}{5} \end{cases}$ dont la solution est $\left(b_1 = 4,\ b_2 = -\dfrac{10}{3}\right)$.

Ainsi $p(1) = 4t - \dfrac{10}{3} t^2$. On en déduit :

$$\delta = d^2(1, F) = \|1\|^2 - (\,1 \mid p(1)\,) = 1 - \left(\frac{4}{2} - \frac{10}{9}\right) = \frac{1}{9} \qquad \text{et} \qquad d(1, F) = \frac{1}{3}.$$

Le cas général, i.e. le calcul de $\quad \delta = \displaystyle\inf_{(a_1,\ldots,a_n)\in\mathbb{R}^n} \int_0^1 (1 + a_1 t + \cdots + a_n t^n)^2 \, \mathrm{d}t$ est l'objet de l'exercice I.4.0 de la page 206.

1.3.3 Matrice de Gram

Définition 7. Soient $p \in \mathbb{N}^*$ et $(v) := (v_1, \dots, v_p)$ un système de p vecteurs d'un espace préhilbertien E. On appelle *matrice de Gram* de ce système et l'on note $G(v)$ la matrice $G(v) := \big((\, v_i \mid v_j \,)\big)_{i,j\in[\![1,p]\!]}$.
Son déterminant est appelé *déterminant de Gram* du système (v) et noté $\mathrm{Gram}\,(v)$.

Remarque.
Si le système (v) est *libre*, on considère le sous-espace $F := \mathrm{Vect}\{v_1, \dots, v_p\}$ et on le munit d'un produit scalaire par restriction de celui de E : F est un espace euclidien de dimension p. La matrice de Gram $G(v)$ est alors la matrice du produit scalaire de F dans la base $(v_1, \dots, v_p)$.
Il en résulte, dans ce cas, que la matrice $G(v)$ est *définie positive* et que $\mathrm{Gram}\,(v) > 0$.

Proposition 16. Soient $p \in \mathbb{N}^*$, $(v) = (v_1, \dots, v_p)$ un système de p vecteurs d'un espace préhilbertien et $G = G(v)$ sa matrice de Gram. On a $G \in \mathcal{S}_p^+(\mathbb{R})$ et le rang de la matrice G est celui du système de vecteurs (v). En particulier, on a $\mathrm{Gram}\,(v) = \det G \geqslant 0$ et $\mathrm{Gram}\,(v) = \det G = 0$ si, et seulement si, le système (v) est lié. Si le système (v) est libre, on a $G \in \mathcal{S}_p^{++}(\mathbb{R})$.

Démonstration. Si $F = \mathrm{Vect}\{v_1, \dots, v_p\}$ est réduit à zéro, $G(v) = 0$. Sinon F possède une base orthonormée $(\varepsilon_1, \dots, \varepsilon_r)$, l'entier r étant le rang du système (v). Pour tout $i \in [\![1,p]\!]$, $v_i = \sum_{h=1}^{r} (\, \varepsilon_h \mid v_i \,)\ \varepsilon_h$. Donc pour tout $(i,j) \in [\![1,p]\!]^2$, $(\, v_i \mid v_j \,) = \sum_{h=1}^{r} (\, \varepsilon_h \mid v_i \,)(\, \varepsilon_h \mid v_j \,)$. Soit $M \in \mathcal{M}_{r,p}(\mathbb{R})$ dont le $j^{ème}$ vecteur colonne C_j est formé des coordonnées $(\, \varepsilon_1 \mid v_j \,), \dots, (\, \varepsilon_r \mid v_j \,)$ de v_j dans la base $(\varepsilon_1, \dots, \varepsilon_r)$ ($j \in [\![1,p]\!]$). On a $G(v) = {}^tMM$. On a vu (proposition 48 de la page 134) qu'alors $G \in S_p^+(\mathbb{R})$ et que $\mathrm{rang}\, G = \mathrm{rang}\, M = \mathrm{rang}(v)$. On a $\det G = 0$ si, et seulement si, la famille (v) est liée. Si elle est libre on a donc $G \in \mathcal{S}_p^{++}(\mathbb{R})$ et par suite $\det G > 0$. ∎

Proposition 17. Soit E un espace préhilbertien. Soit $(v_1, \dots, v_n)$ une base d'un sous-espace vectoriel F de dimension finie $n > 0$ de E. Pour tout point $a \in E$, on a $d^2(a, F) = \dfrac{\mathrm{Gram}\,(a, v_1, \dots, v_n)}{\mathrm{Gram}\,(v_1, \dots, v_n)}$.

Démonstration.
On note que $(\, a \mid a \,) = \|a - p(a)\|^2 + \|p(a)\|^2$ et, pour tout $j \in [\![1,n]\!]$, $(\, a \mid v_j \,) = (\, p(a) \mid v_j \,)$.

Ainsi :

$$\operatorname{Gram}(a,v_1,\dots,v_n)=\begin{vmatrix} \|a-p(a)\|^2 & (a\mid v_1) & \dots & (a\mid v_n)\\ 0 & (v_1\mid v_1) & \dots & (v_1\mid v_n)\\ \vdots & \vdots & \dots & \vdots\\ 0 & (v_n\mid v_1) & \dots & (v_n\mid v_n)\end{vmatrix}$$

$$+\begin{vmatrix} \|p(a)\|^2 & (p(a)\mid v_1) & \dots & (p(a)\mid v_n)\\ (v_1\mid p(a)) & (v_1\mid v_1) & \dots & (v_1\mid v_n)\\ \vdots & \vdots & \dots & \vdots\\ (v_n\mid p(a)) & (v_n\mid v_1) & \dots & (v_n\mid v_n)\end{vmatrix}.$$

Donc $\operatorname{Gram}(a,v_1,\dots,v_n)=\|a-p(a)\|^2\operatorname{Gram}(v_1,\dots,v_n)+\operatorname{Gram}(p(a),v_1,\dots,v_n)$. Les $n+1$ vecteurs $p(a),v_1,\dots,v_n$ de l'espace vectoriel F sont liés et leur déterminant de Gram est donc nul. Ainsi $\operatorname{Gram}(a,v_1,\dots,v_n)=d^2(a,F)\operatorname{Gram}(v_1,\dots,v_n)$. ∎

On retrouve ainsi le résultat de l'exercice 4 de la page 153 :

$$d^2(1,F)=\frac{\operatorname{Gram}(1,t,t^2)}{\operatorname{Gram}(t,t^2)}=\frac{240}{2160}=\frac{1}{9}.$$

1.4 Orthogonalisation de Gram-Schmidt

1.4.1 Algorithme de Gram-Schmidt

La méthode d'orthogonalisation dite *de Gram-Schmidt* est en fait due à l'astronome, mathématicien et physicien français Pierre-Simon Laplace (qui la donne dans le *Premier Supplément* de sa *Théorie Analytique des Probabilités*, 1816). Elle a été redécouverte par le mathématicien danois Jorgen Gram, dans le cadre de recherches sur la *Méthode des moindres carrés* (1883), puis présentée par Erhard Schmidt (1907). Cette méthode est un important outil théorique et pratique. Nous consacrerons un module à l'une de ses applications, les *polynômes orthogonaux* dans le livre [L2-L3].

Théorème 18 (Orthogonalisation de Gram-Schmidt).
Soient E un espace préhilbertien réel et $(v_n)_{n\geqslant 0}$ (resp. $(v_n)_{n\in[\![0,N]\!]}$) une famille *libre* de vecteurs de E, il *existe* une *unique* famille orthogonale $(e_n)_{n\geqslant 0}$ (resp. $(e_n)_{n\in[\![0,N]\!]}$) telle que, pour tout $k\in\mathbb{N}$ (resp. $k\in[\![0,N]\!]$) :
(i) $\operatorname{Vect}(v_0,\dots,v_k)=\operatorname{Vect}(e_0,\dots,e_k)$;
(ii) la matrice de passage de $(v_0,\dots,v_k)$ à $(e_0,\dots,e_k)$ est *triangulaire supérieure unipotente*.
On dit alors que la famille $(e_n)_{n\geqslant 0}$ (resp. $(e_n)_{n\in[\![0,N]\!]}$) est *l'orthogonalisée* de la famille $(v_n)_{n\geqslant 0}$ (resp. $(v_n)_{n\in[\![0,N]\!]}$) par le procédé d'orthogonalisation de Gram-Schmidt.

Démonstration. On va *construire* la solution par *récurrence*, ce qui permet de parler de *procédé d'orthogonalisation*. On note $F_k:=\operatorname{Vect}(v_0,\dots,v_k)$. On définit $e_0:=v_0$. Soit $k>0$, supposons avoir construit $e_0,\dots,e_{k-1}$ et construisons e_k. On

cherche $y_k \in F_{k-1}$, tel que $e_k = v_k + y_k$ réponde à la question, le sous-espace $F_{k-1} \subset F_k$ est un *hyperplan*, dont le supplémentaire orthogonal est une *droite* D_k, si e_k existe, il engendre D_k et sa projection orthogonale sur F_{k-1} est nulle, donc $-y_k$ est nécessairement le projeté orthogonal $p(v_k)$ de v_k sur F_{k-1}. Ceci montre l'unicité de y_k, par ailleurs, si l'on définit $y_k := -p(v_k)$, puis $e_k := v_k + y_k$, on a $p(e_k) = 0$ et e_k est orthogonal à F_{k-1}. Ainsi e_k répond à la question (et est unique) : la matrice de passage est clairement triangulaire supérieure et unipotente. ∎

Remarque. On a $(e_k \mid v_k) = \|e_k\|^2$ et $\|v_k\|^2 = \|e_k\|^2 + \|y_k\|^2$, d'où $\|e_k\| \leqslant \|v_k\|$, l'égalité étant vraie si, et seulement si, $v_k \in F_{k-1}^{\perp}$.

Soit $k \geqslant 1$. Connaissant $e_0, \dots, e_{k-1}$ et v_k, on peut calculer e_k en utilisant la base *orthonormée* $\left(\dfrac{e_0}{\|e_0\|}, \dots, \dfrac{e_{k-1}}{\|e_{k-1}\|}\right)$ de F_{k-1} (*cf.* proposition 13 de la page 151). Notons $p_h(w)$ la projection orthogonale d'un vecteur $w \in E$ sur la droite $\mathbb{R}e_h$, on a :

$$e_k = v_k - \sum_{h=0}^{k-1} (e_h \mid v_k) \frac{e_h}{\|e_h\|^2} = v_k - \sum_{h=0}^{k-1} p_h(v_k).$$

Cette formule n'est pas la plus simple pour les calculs pratiques, il vaut mieux procéder comme dans l'exercice 5 de la page suivante.

Corollaire 19. Soient E un espace préhilbertien réel et $(v_n)_{n\geqslant 0}$ (resp. $(v_n)_{n\in[\![0,N]\!]}$) une famille *libre* de vecteurs de E.
Il *existe* une *unique* famille *orthonormée* $(\varepsilon_n)_{n\geqslant 0}$ (resp. $(\varepsilon_n)_{n\in[\![0,N]\!]}$) telle que, pour tout $k \in \mathbb{N}$ (resp. $k \in [\![0, N]\!]$), $\mathrm{Vect}(v_0, \dots, v_k) = \mathrm{Vect}(\varepsilon_0, \dots, \varepsilon_k)$ et $(\varepsilon_k \mid v_k) > 0$.

Démonstration. Pour tout $n \in \mathbb{N}$ (resp. $n \in [\![0, N]\!]$) le vecteur ε_n est nécessairement colinéaire à e_n et de norme 1.

On doit donc avoir $\varepsilon_n = \pm\dfrac{e_n}{\|e_n\|}$, $(\varepsilon_n \mid v_n) = \pm\dfrac{(e_n \mid v_n)}{\|e_n\|} = \pm\dfrac{\|e_n\|^2}{\|e_n\|} = \pm\|e_n\|$. La seule valeur possible pour ε_n est donc $\varepsilon_n = \dfrac{e_n}{\|e_n\|}$. Cette valeur convient évidemment. ∎

Exemple.
Les vecteurs :

$$v_0 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \qquad v_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}$$

forment une base de $\mathbb{R}^3$. Nous allons calculer la base *orthogonalisée* par le procédé de Gram-Schmidt.
Le premier vecteur e_0 de celle-ci est $e_0 := v_0$, le deuxième est :

$$e_1 := v_1 - (e_0 \mid v_1) \frac{e_0}{\|e_0\|^2} = \begin{pmatrix} 4/3 \\ 1/3 \\ 5/3 \end{pmatrix}$$

et le dernier :

$$e_2 := v_2 - (\,e_0 \mid v_2\,)\frac{e_0}{\|e_0\|^2} - (\,e_1 \mid v_2\,)\frac{e_1}{\|e_1\|^2} = \begin{pmatrix} 1 \\ -3/2 \\ -1/2 \end{pmatrix}.$$

La base *orthonormée* correspondante est

$$\varepsilon_0 = \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ -1/\sqrt{3} \end{pmatrix} \qquad \varepsilon_1 = \begin{pmatrix} 4/\sqrt{42} \\ 1/\sqrt{42} \\ 5/\sqrt{42} \end{pmatrix} \qquad \varepsilon_2 = \begin{pmatrix} 2/\sqrt{14} \\ -3/\sqrt{14} \\ -1/\sqrt{14} \end{pmatrix}$$

Polynômes orthogonaux. Soit $E = \mathbb{R}[T]$ muni du produit scalaire :

$$(\,P \mid Q\,) := \int_a^b P(t)Q(t)w(t)\ \mathrm{d}t$$

(*cf.* l'exercice 12 de la page 129), on peut appliquer à la base $(t^n)_{n\in\mathbb{N}}$ de E le procédé d'orthogonalisation de Gram-Schmidt, on obtient une nouvelle base $(P_n)_{n\in\mathbb{N}}$ de E, possédant les propriétés suivantes :
(i) quel que soit $n \in \mathbb{N}$, $\deg P_n = n$;
(ii) le polynôme P_n est *unitaire* ;
(iii) $(\,P_m \mid P_n\,) = 0$ pour tout couple (m, n) d'entiers *distincts*.

Exercice 5.

On choisit $b := -1$, $a := 1$, $w := 1$. Les polynômes P_n sont appelés *polynômes de Legendre*. Calculer P_0, P_1, P_2, P_3.

Solution. On a $(\,t^p \mid t^q\,) = \displaystyle\int_{-1}^{1} t^{p+q}\ \mathrm{d}t$, donc $(\,t^p \mid t^q\,) = 0$ si $p+q$ est impair et $(\,t^p \mid t^q\,) = \dfrac{2}{p+q+1}$ si $p+q$ est pair.

Ainsi $P_0(t) = 1$, $P_1(t) = t$. On pose $P_2(t) = t^2 + \alpha_2 t + \beta_2$ et on détermine α_2 et β_2 grâce au système : $\begin{cases} (\,P_2 \mid P_0\,) = 2\left(\dfrac{1}{3} + \beta_2\right) = 0 \\ (\,P_2 \mid P_1\,) = 2\dfrac{\alpha_2}{3} = 0. \end{cases}$

D'où $\alpha_2 = 0$, $\beta_2 = -\dfrac{1}{3}$ et $P_2(t) = t^2 - \dfrac{1}{3}$.

On pose $P_3 = t^3 + \alpha_3 t^3 + \beta_3 t + \gamma_3$ et l'on détermine α_3, β_3 et γ_3 grâce au système :

$$\begin{cases} (\,P_3 \mid P_0\,) = 2\left(\dfrac{\alpha_3}{3} + \gamma_3\right) = 0 \\ (\,P_3 \mid P_1\,) = 2\left(\dfrac{1}{5} + \dfrac{\beta_3}{3}\right) = 0 \\ (\,P_3 \mid P_2\,) = 2\left(\dfrac{\alpha_3}{5} + \dfrac{\gamma_3}{3}\right) = 0 \end{cases}$$

D'où $\alpha_3 = \gamma_3 = 0$ $\beta_3 = -\dfrac{3}{5}$ et $P_3(t) = t^3 - \dfrac{3}{5}t$.

Nous reviendrons sur les polynômes de Legendre dans le module *Polynômes orthogonaux* de [L2-L3]. Nous montrerons que, pour tout $n \in \mathbb{N}$, $P_n(t) = \dfrac{n!}{(2n)!}\dfrac{\mathrm{d}^n}{\mathrm{d}t^n}[(t^2-1)^n]$.

Gram-Schmidt et Cholesky. Revenons à l'orthogonalisation de Gram-Schmidt. Soient E un espace préhilbertien réel et $(v) := (v_0, \dots, v_n)$ une famille libre de vecteurs de E. On pose, comme plus haut, $F_n := \mathrm{Vect}(v_0, \dots, v_n) \subset E$ et l'on munit ce sous-espace du produit scalaire induit par celui de E. Soit $G := \big((\, v_i \mid v_j \,)\big)_{i,j\in[\![0,n]\!]}$ la matrice de Gram de la famille (v) (de vecteurs de F_n). C'est la matrice associée au produit scalaire de F_n dans la base (v). La matrice associée à ce produit scalaire dans la base orthogonalisée $(e) = (e_0, \dots, e_n)$ est la matrice diagonale $D := \mathrm{Diag}(\|e_0\|^2, \dots, \|e_n\|^2)$ (la base (e) est orthogonale). La matrice $T := \mathrm{Pass}\big((e),(v)\big)$ est triangulaire supérieure unipotente. Posons $\Theta := \mathrm{Diag}(\|e_0\|, \dots, \|e_n\|)T$. On a $G = {}^tTDT = {}^t\Theta\Theta$, ce qui est la *décomposition de Cholesky* de $G \in S_n^{++}(\mathbb{R})$.

Le procédé d'orthogonalisation de Gram-Schmidt fournit une nouvelle démonstration du théorème 36 de la page 120, en effet toute matrice de $S_n^{++}(\mathbb{R})$ peut être considérée comme la matrice associée à un produit scalaire sur $\mathbb{R}^n$.

Proposition 20 (Inégalité d'Hadamard).

Pour toute matrice $A \in M_n(\mathbb{R})$, $|\det A| \leqslant \displaystyle\prod_{j=1}^{n} \|C_j\|$, où $\|C_j\|$ désigne la norme euclidienne usuelle du $j^{\text{ème}}$ vecteur colonne C_j de A.

Cette inégalité est due à J. Hadamard (mathématicien français, 1865-1963), elle généralise des propriétés « bien connues » de l'aire et du volume en dimensions deux et trois.

Démonstration. On peut supposer que la matrice A est inversible (sinon le résultat est évident), ses vecteurs colonnes $(C_1, \dots, C_n)$ forment alors une base de $\mathbb{R}^n$. L'espace $\mathbb{R}^n$ étant muni de sa norme usuelle, la matrice de Gram du système $(C_1, \dots, C_n)$ est ${}^tAA$. Si $(e_1, \dots, e_n)$ est la base orthogonalisée de Schmidt de $(C_1, \dots, C_n)$, on a (*cf.* ci-dessus) : ${}^tAA = {}^tTDT$, avec $D := \mathrm{Diag}(\|e_1\|^2, \dots, \|e_n\|^2)$.

Ainsi : $\big(\mathrm{Det}(A)\big)^2 = \mathrm{Det}(D) = \displaystyle\prod_{j=1}^{n} \|e_j\|^2$. Pour tout $j \in [\![1,n]\!]$, $\|e_j\| \leqslant \|C_j\|$ (*cf.* la remarque de la page 156), d'où le résultat. ∎

Gram-Schmidt, Gauß et Cholesky.

Comme nous l'avons noté (*cf.* 3.2.2 de la page 128), lorsque les mineurs principaux de la matrice A d'une forme quadratique $\Phi \in Q(E)$ dans une base $\mathcal{B} := (v_1, \dots, v_n)$ sont tous non nuls, l'algorithme LU peut être interprété comme une simple « traduction » de l'algorithme de Gauß. Si Φ est *définie positive*, A est la matrice de Gram de $\mathcal{B}$ et $A \in S_n^{++}(\mathbb{R})$. On peut, d'une part, déterminer la base $(e) = (e_1, \dots, e_n)$ orthogonalisée de $\mathcal{B}$, et, d'autre part, appliquer la décomposition de Gauß à la forme quadratique Φ ; on obtient $\Phi = \displaystyle\sum_{i=1}^{n} \alpha_i f_i^2$, les formes

linéaires $(f) := (f_1, \dots, f_n)$ formant une *base* de E^*. Nous allons montrer que cette base n'est autre que *la base duale* de (e), soit $(v^*) := (v_1^*, \dots, v_n^*)$ la base duale de (v). On sait que la matrice de passage $\mathrm{Pass}((v^*), (f))$ est triangulaire unipotente. il en est donc de même de la matrice de passage $T' = \mathrm{Pass}\big((f^*), (v)\big)$ (où (f^*) est la base de E dont (f) est la base duale). On pose $D' := \mathrm{Diag}(\alpha_1, \dots, \alpha_n)$, $T := \mathrm{Pass}\big((e), (v)\big)$. On a $A = {}^tTDT = {}^tT'D'T'$, d'où par unicité de la décompositions LU : $T = T'$ et $(f^*) = (e)$, puis $(f) = (e^*)$.
On a ainsi une nouvelle interprétation des algorithmes LU, de Gauß et de Cholesky sous forme d'orthogonalisation de Schmidt et inversement.

Exercice 6.
Reprenons l'exemple des polynômes de Legendre (*cf.* exercice 5). Soit $E := \mathbb{R}_3[T]$, avec le produit scalaire $(\,P \mid Q\,) := \displaystyle\int_{-1}^{1} P(t)Q(t)\,\mathrm{d}t$. Vérifier le résultat établi ci-dessus.

Solution. On a $n = 4$, $v_1 = 1$, $v_2 = t$, $v_3 = t^2$, $v_4 = t^3$. La matrice de Gram correspondante est $G = \big((\,t^i \mid j^j\,)\big)_{i,j\in[\![0,3]\!]} = \begin{pmatrix} 2 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 2/5 \\ 2/3 & 0 & 2/5 & 0 \\ 0 & 2/5 & 0 & 2/7 \end{pmatrix}$. Posons $P(t) := x_0 + x_1t + x_2t^2 + x_3t^3$.

On a

$$\begin{aligned}(\,P \mid P\,) &= 2\left(x_0^2 + \frac{1}{3}x_1^2 + \frac{1}{5}x_2^2 + \frac{1}{7}x_3^2 + \frac{2}{3}x_0x_2 + \frac{2}{5}x_1x_3\right)\\ &= 2\left((x_0 + \frac{1}{3}x_2)^2 + \frac{1}{3}(x_1 + \frac{3}{5}x_3)^2 + \frac{4}{45}x_2^2 + \frac{4}{175}x_3^2\right).\end{aligned}$$

Avec les notations utilisées ci-dessus, on a :

$$f_1 = x_0 + \frac{1}{3}x_2, \quad f_2 = x_1 + \frac{3}{5}x_3, \quad f_3 = x_2 \quad \text{et} \quad f_4 = x_3.$$

En inversant ces relations, on obtient : $x_0 = f_1 - \frac{1}{3}f_3$, $x_1 = f_2 - \frac{3}{5}f_4$, $x_2 = f_3$, $x_3 = f_4$.
Par ailleurs, on a (*cf.* l'exercice 5 de la page 157) :

$$\mathrm{Pass}\Big((1, t, t^2, t^3), (P_0, P_1, P_2, P_3)\Big) = \begin{pmatrix} 1 & 0 & -1/3 & 0 \\ 0 & 1 & 0 & -3/5 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

d'où le résultat (*cf.* l'exercice 6 de la page 101.

1.4.2 Algorithme de Gram-Schmidt modifié

Quand le procédé d'orthonormalisation de Gram-Schmidt est programmé sur machine, en *calcul flottant*, le système obtenu en sortie n'est, en général, pas exactement orthogonal, à cause des erreurs d'arrondi. Dans certains cas il est même loin de l'être ; pour cette raison, on dit que le procédé est *numériquement instable*. Une légère modification du procédé de Gram-Schmidt permet de réduire ce défaut, le nouvel algorithme est appelé algorithme de Gram-Schmidt modifié (ou stable). L'idée de ce nouveau procédé est de traiter, à chaque pas, *l'ensemble* des vecteurs du système.

Soit $(v_0, \dots, v_n)$ le système libre donné. On pose $(v_0^{(0)} := v_0, \dots, v_n^{(0)} := v_n)$, puis on définit un nouveau système $(\varepsilon_0, v_1^{(1)}, \dots, v_n^{(1)})$, où $\varepsilon_0 := \frac{v_0}{\|v_0\|}$ et les vecteurs $v_1^{(1)}, \dots, v_n^{(1)})$ sont les *projetés orthogonaux* sur $(\varepsilon_0)^\perp$ des vecteurs $(v_1, \dots, v_n)$.

On a, pour tout $h \in [\![1,n]\!]$: $v_h^{(1)} = v_h - (\,v_h \mid \varepsilon_0\,)\varepsilon_0$.

Dans l'étape suivante, on définit, de façon analogue, un troisième système $(\varepsilon_0, \varepsilon_1, v_2^{(2)}, \ldots, v_n^{(2)})$, où $\varepsilon_1 := \frac{v_1^{(1)}}{\|v_1^{(1)}\|}$ et les vecteurs $v_2^{(2)}, \ldots, v_n^{(2)})$ sont les projetés orthogonaux sur $\mathrm{Vect}(\varepsilon_0, \varepsilon_1)^\perp$ des vecteurs $(v_2^{(1)}, \ldots, v_n^{(1)})$. On a, pour tout $h \in [\![1,n]\!]$: $v_h^{(2)} = v_h^{(1)} - (\,v_h^{(1)} \mid \varepsilon_1\,)\varepsilon_1$ On poursuit (s'il y a lieu) la construction de façon itérative, en utilisant, pour $j \in [\![0,n]\!]$ et $h \in [\![j+1,n]\!]$, les formules :

$$\varepsilon_j := \frac{v_j^{(j)}}{\|v_j^{(j)}\|} \quad \text{et} \quad v_h^{(j+1)} := v_h^{(j)} - (\,v_h^{(j)} \mid \varepsilon_j\,)\varepsilon_j = p_j(v_h^{(j)}).$$

Ces formules remplacent les formules $\mathrm{e}_k = v_k - \sum_{h=0}^{k-1} p_h(v_k) = v_k - \sum_{h=0}^{k-1} (\,v_k \mid \varepsilon_h\,)\varepsilon_h$, $\varepsilon_k = \frac{e_k}{\|e_k\|}$ de l'algorithme de Gram-Schmidt initial.

Pour tout $k \in [\![0,n]\!]$, $F_k = (v_0, \ldots, v_k) = (\varepsilon_0, \ldots, \varepsilon_k)$, la matrice de passage de $(v_0, \ldots, v_k)$ à $(\varepsilon_0, \ldots, \varepsilon_k)$ est triangulaire supérieure et l'on vérifie que $(\,\varepsilon_k \mid v_k\,) > 0$. Le système $(\varepsilon_0, \ldots, \varepsilon_n)$ est orthonormé, le procédé modifié fournit donc bien le *même résultat* que le procédé initial (on utilise l'assertion d'unicité dans le corollaire 19 de la page 156).

Exemples.

(i) Reprenons l'exemple de la page 156, traitons le avec l'algorithme de Gram-Schmidt modifié. On a :

$$v_0^{(0)} := v_0 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \qquad v_1^{(0)} := v_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \qquad v_2^{(0)} := v_2 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}.$$

Le premier vecteur de la base orthonormalisée est $\varepsilon_0 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix}$.

On calcule ensuite $v_1^{(1)} = \begin{pmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{5}{3} \end{pmatrix}$ et $v_2^{(1)} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix}$. Le deuxième vecteur de la base orthonormalisée est donc : $\varepsilon_1 = \begin{pmatrix} \frac{4}{\sqrt{42}} \\ \frac{1}{\sqrt{42}} \\ \frac{5}{\sqrt{42}} \end{pmatrix}$, et $v_2^{(2)} = \begin{pmatrix} 1 \\ -\frac{3}{2} \\ -\frac{1}{2} \end{pmatrix}$, ce qui donne le troisième vecteur de la base orthonormée. $\varepsilon_2 = \begin{pmatrix} \frac{2}{\sqrt{14}} \\ -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{14}} \end{pmatrix}$.

(ii) Appliquons l'algorithme de Gram-Schmidt modifié à la base de $\mathbb{R}^4$:

$$v_0 = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \end{pmatrix} \qquad v_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \end{pmatrix} \qquad v_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 1 \end{pmatrix} \qquad v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \end{pmatrix}.$$

Écrivons sous forme matricielle, dans la base canonique de $\mathbb{R}^4$ les systèmes successifs obtenus à chaque étape de l'algorithme :

$$A = A^{(0)} := \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \qquad A^{(1)} := \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{5} & -\frac{2}{5} & 0 \\ \frac{1}{\sqrt{5}} & \frac{6}{5} & \frac{4}{5} & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

$$A^{(2)} := \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & \frac{2}{7} & \frac{3}{14} \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & -\frac{4}{7} & -\frac{3}{7} \\ 0 & \frac{5}{\sqrt{70}} & \frac{6}{7} & \frac{9}{14} \\ 0 & 0 & 1 & 2 \end{pmatrix} \qquad A^{(3)} := \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & \frac{2}{\sqrt{105}} & -\frac{1}{\sqrt{30}} \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & -\frac{4}{\sqrt{105}} & \frac{2}{\sqrt{30}} \\ 0 & \frac{5}{\sqrt{70}} & \frac{6}{\sqrt{105}} & -\frac{3}{\sqrt{30}} \\ 0 & 0 & \frac{7}{\sqrt{105}} & \frac{4}{\sqrt{30}} \end{pmatrix}.$$

Notons $Q := A^{(3)}$. Les calculs effectués au cours de l'algorithme se traduisent par $A = QR$, avec

$$R := \begin{pmatrix} \sqrt{5} & \frac{4}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ 0 & \frac{\sqrt{14}}{\sqrt{5}} & \frac{16}{\sqrt{70}} & \frac{5}{\sqrt{70}} \\ 0 & 0 & \frac{\sqrt{15}}{\sqrt{7}} & \frac{20}{\sqrt{105}} \\ 0 & 0 & 0 & \frac{\sqrt{5}}{\sqrt{6}} \end{pmatrix}.$$

On obtient ce que l'on appelle une *décomposition* QR de la matrice A. Nous étudierons plus loin cette question.

1.5 Adjoint d'un endomorphisme d'un espace préhilbertien réel

1.5.1 Généralités

Théorème et définition 21. Soient E un espace vectoriel préhilbertien réel et u un endomorphisme de E. S'il existe un endomorphisme v de E tel que, pour tout $(x, y) \in E^2$, $\big(u(x) \mid y\big) = \big(x \mid v(y)\big)$, cet endomorphisme est unique. On l'appelle *l'adjoint* de u ; on le note u^*. Si v est l'adjoint de u, v admet u pour adjoint : $v^* = (u^*)^* = u$.

Démonstration.

Soient v_1 et v_2 deux endomorphismes de E tels que, pour tout $(x, y) \in E^2$:

$$\big(u(x) \mid y\big) = \big(x \mid v_1(y)\big) \qquad \text{et} \qquad \big(u(x) \mid y\big) = \big(x \mid v_2(y)\big).$$

Pour $y \in E$ fixé, on a, pour tout $x \in E$:

$$\big(x \mid (v_1 - v_2)(y)\big) = \big(x \mid v_1(y)\big) - \big(x \mid v_2(y)\big) = 0,$$

donc $(v_1 - v_2)(y) = 0$. Cela étant vrai pour tout $y \in E$, on a $v_1 = v_2$. ∎

Soit $u \in \mathcal{L}(E)$ admettant un adjoint u^*. Si F est un sous-espace vectoriel de E stable par u, F n'est pas en général stable par u^* : il n'y a pas d'opération naturelle de restriction pour l'adjoint.

Nous détaillerons en 5.5.2 le cas de la dimension finie. Voici des exemples dans le cas préhilbertien général.

Exemples.
(i) Soit $(a,b) \in E^2$. Soit $u \in \mathcal{L}(E)$ l'endomorphisme défini par $u : x \mapsto (\,a \mid x\,)\,b$. Pour tout $(x,y) \in E^2$, $\big(\,u(x) \;\big|\; y\,\big) = (\,a \mid x\,)(\,b \mid y\,) = \big(\,x \;\big|\; (\,b \mid y\,)a\,\big)$. L'endomorphisme u admet donc pour adjoint l'endomorphisme u^* défini par $u^* : y \mapsto (\,b \mid y\,)a$.
(ii) Soit $E := \ell^2(\mathbb{R})$ muni de son produit scalaire.
Si $x := (x_n)_{n\in\mathbb{N}} \in \ell^2(\mathbb{R})$, on pose $s^+(x) := (0, x_0, x_1, x_2, \ldots)$ $(s^+(x)_n := x_{n-1}$, avec la convention $x_{-1} = 0)$ et $s^-(x) := (x_1, x_2, x_3, \ldots)$ $(s^-(x)_n := x_{n+1})$.
On vérifie que s^+ et s^- sont des endomorphismes de $\ell^2(\mathbb{R})$: s^+ est le *décalage* (ou « shift ») *à droite*, s^- le *décalage* (ou « shift ») *à gauche*.
On a $\big(\,s^+(x) \;\big|\; y\,\big) = \big(\,x \;\big|\; s^-(y)\,\big) = x_0y_1 + x_1y_2 + \cdots + x_ny_{n+1} + \cdots$, les endomorphismes s^+ et s^- sont donc adjoints.

Un endomorphisme $u \in \mathcal{L}(E)$ n'a pas toujours un adjoint, comme le montrent les deux exemples suivants (pour un autre exemple, *cf.* l'exercice I.4.12 de la page 209). Le lecteur pourra revenir sur ces deux exemples après la lecture du module II.1 et étudier la continuité des endomorphismes considérés dans les deux cas.

Exemples.
(i) Soit $E := \mathbb{R}[X]$ muni du produit scalaire défini par $(\,x^m \mid x^n\,) = \delta_{mn}$, pour tous $m,n \in \mathbb{N}$. Soit u l'unique endomorphisme de E tel que $u(x^n) = 1$ pour tout $n \in \mathbb{N}$. Cet endomorphisme n'a pas d'adjoint. En effet s'il en avait un, on aurait, d'une part : $1 = \big(\,u(x^n) \;\big|\; 1\,\big) = \big(\,x^n \;\big|\; u^*(1)\,\big)$, pour tout $n \in \mathbb{N}$, et, d'autre part : $u^*(1) \in \mathbb{R}[X]$, donc $\big(\,x^n \;\big|\; u^*(1)\,\big) = 0$, pour tout $n > \deg u^*(1)$, ce qui est impossible.
(ii) Soit $E = \ell^2(\mathbb{R})$, on note $e_n \in E$ la suite dont tous les termes sont nuls sauf le $(n+1)^{\text{ème}}$ qui est égal à 1. Soit $a := \left(\dfrac{1}{n+1}\right)_{n\in\mathbb{N}} \in E$, on définit une forme linéaire f sur E par $f(x) := (\,a \mid x\,) = \displaystyle\sum_{n\in\mathbb{N}} \frac{x_n}{n+1}$, pour tout $x \in E$. On définit un endomorphisme $u \in \mathcal{L}(E)$ par $u\big((x_n)_{n\in\mathbb{N}}\big) := \left(\dfrac{x_n}{n+1}\right)_{n\in\mathbb{N}}$, u est injectif mais n'est pas surjectif : $a \notin \operatorname{Im} u$ (sinon la suite $(1,1,\ldots,1,\ldots)$ appartiendrait à $E = \ell^2(\mathbb{R})$, ce qui est faux), on a toutefois l'inclusion $\{e_0,\ldots,e_n,\ldots\} \subset \operatorname{Im} u$.
a) On note $F := \operatorname{Im} u$, on le munit du produit scalaire induit par celui de E, et l'on désigne par g la restriction de f à F. L'application linéaire $g : F \to \mathbb{R}$ *n'est pas un produit scalaire* : il n'existe pas d'élément $b := (b_n) \in F$ tel que $g(x) = (\,b \mid x\,)$, pour tout $x \in F$, sinon l'on aurait $b_n = ((\,b \mid x\,))_n = g(e_n) = f(e_n) = \dfrac{1}{n+1}$, pour tout $n \in \mathbb{N}$, donc $b = a$, ce qui est impossible puisque $a \notin F$.
b) On définit un endomorphisme v de E par $v(x) := f(x)e_0$, le sous-espace F de E est v-stable $(e_0 \in F)$, on note w la restriction de v à F. Nous allons montrer que $w \in \mathcal{L}(F)$

n'a pas d'adjoint. S'il en avait un, on aurait, pour tout $x \in F$:

$$\big(\, w(x) \;\big|\; e_0 \,\big) = \big(\, x \;\big|\; w^*(e_0) \,\big) = \big(\, f(x)e_0 \;\big|\; e_0 \,\big) = f(x),$$

on aurait en particulier $\big(\, e_n \;\big|\; w^*(e_0) \,\big) = f(e_n) = \dfrac{1}{n+1}$, pour tout $n \in \mathbb{N}$, d'où $w^*(e_0) = a$, ce qui est impossible puisque $a \notin F$.

Soit E un espace vectoriel réel, on rappelle que les endomorphismes de E forment une $\mathbb{R}$-algèbre $\mathcal{L}_{\mathbb{R}}(E)$. Pour simplifier, dans ce paragraphe, nous la noterons $\mathcal{L}(E)$.

Proposition 22. Soit E un espace préhilbertien réel.
L'ensemble des endomorphismes de E ayant un adjoint est une sous-algèbre de $\mathcal{L}(E)$. Plus précisément, si $u, v \in \mathcal{L}(E)$ ont des adjoints u^* et v^* : quel que soit $\alpha \in \mathbb{R}$, $(\alpha u + v)^* = \alpha u^* + v^*$, $(uv)^* = v^*u^*$ et $(\mathrm{Id}_E)^* = \mathrm{Id}_E$.
Si u est un *automorphisme* de E, si u et son inverse u^{-1} admettent des adjoints, alors $(u^*)^{-1} = (u^{-1})^*$.

Démonstration. Pour tout $(x, y) \in E^2$:

$$\big(\, (\alpha u + v)(x) \;\big|\; y \,\big) = \big(\, \alpha x \;\big|\; u^*(y) \,\big) + \big(\, x \;\big|\; v^*(y) \,\big) = \big(\, x \;\big|\; (\alpha u^* + v^*)(y) \,\big).$$

On a, par ailleurs, $\big(\, uv(x) \;\big|\; y \,\big) = \big(\, v(x) \;\big|\; u^*(y) \,\big) = \big(\, x \;\big|\; v^*u^*(y) \,\big)$.
Le reste de la démonstration est immédiat. ∎

Proposition 23. Soit u un endomorphisme de l'espace préhilbertien réel E admettant un adjoint u^*. Alors $\operatorname{Ker} u^* = (\operatorname{Im} u)^\perp$ et $\operatorname{Im} u^* \subset (\operatorname{Ker} u)^\perp$. Si E est euclidien, alors $\operatorname{Im} u^* = (\operatorname{Ker} u)^\perp$.
Si $F \subset E$ est un sous-espace vectoriel *stable* par u, alors $F^\perp \subset E$ est un sous-espace vectoriel stable par u^*.

Démonstration.
On a $u^*(x) = 0$ si, et seulement si, $\big(\, u^*(x) \;\big|\; y \,\big) = 0$, pour tout $y \in E$. D'où $u^*(x) = 0$ si, et seulement si, $\big(\, x \;\big|\; u(y) \,\big) = 0$, pour tout $y \in E$, ainsi $\operatorname{Ker} u^* = (\operatorname{Im} u)^\perp$.
Soit $z \in \operatorname{Im} u^*$, il existe $x \in E$ tel que $z = u^*(x)$. Soit alors $y \in \operatorname{Ker} u$, on a $(\, z \mid y \,) = \big(\, u^*(x) \;\big|\; y \,\big) = \big(\, x \;\big|\; u(y) \,\big) = 0$. Ainsi $\operatorname{Im} u^* \subset (\operatorname{Ker} u)^\perp$. L'égalité dans le cas euclidien résulte du théorème du rang.
Soit $F \subset E$ un sous-espace u-stable. Soit $y \in F^\perp$, montrons que $u^*(y) \in F^\perp$. Pour tout $x \in F$, on a $\big(\, u^*(y) \;\big|\; x \,\big) = \big(\, y \;\big|\; u(x) \,\big) = 0$, car $u(x) \in F$. ∎

Exemple. Revenons à l'exemple des décalages de $\ell^2(\mathbb{R})$ (*cf.* l'exemple (ii) page 162). On a $(s^+)^* = s^-$, $\operatorname{Ker} s^- = \{x \in \ell^2(\mathbb{R}) \mid x_n = 0, \text{ pour tout } n \geqslant 1\} = (\operatorname{Im} s^+)^\perp$, et $\operatorname{Im} s^- = \ell^2(\mathbb{R}) = (\operatorname{Ker} s^+)^\perp$. On a aussi :

$$\operatorname{Ker} s^+ = \{0\} = (\operatorname{Im} s^-)^\perp \quad \text{et} \quad \operatorname{Im} s^+ = \{x \in \ell^2(\mathbb{R}) |\, x_0 = 0\} = (\operatorname{Ker} s^-)^\perp.$$

Si l'espace préhilbertien E est de dimension *infinie*, l'inclusion $\operatorname{Im} u^* \subset (\operatorname{Ker} u)^\perp$ de la proposition 23 peut être *stricte*. Considérons, par exemple, $E = \ell^2(\mathbb{R})$ et l'endomorphisme de E défini par $u : (x_n)_{n\in\mathbb{N}} \mapsto \left(\frac{1}{n+1}x_n\right)_{n\in\mathbb{N}}$. Cet endomorphisme est injectif, mais n'est pas surjectif : $\left(\frac{1}{n+1}\right)_{n\in\mathbb{N}} \notin \operatorname{Im} u$. On vérifie facilement que u admet un adjoint et que $u = u^*$. Ainsi $\operatorname{Im} u^* = \operatorname{Im} u \neq E$, $(\operatorname{Ker} u)^\perp = (\{0\})^\perp = E$ et l'inclusion $\operatorname{Im} u^* \subset (\operatorname{Ker} u)^\perp$ est stricte.

1.5.2 Adjoint d'un endomorphisme d'un espace euclidien

Proposition 24. Tout endomorphisme u d'un espace euclidien E admet un adjoint.

Démonstration. Soit $y \in E$ fixé, soit $f : E \to \mathbb{R}$ définie par $f(x) = \big(y \mid u(x)\big)$; c'est une *forme linéaire* sur E, il existe donc un unique vecteur $v(y) \in E$ tel que $f(x) = \big(v(y) \mid x\big) = \big(y \mid u(x)\big)$. Le vecteur $v(y)$ dépend de y, l'application $y \mapsto v(y)$ est *linéaire*, elle s'écrit en effet comme composée d'applications linéaires : $v = \iota^{-1}{}^t u\, \iota$, en utilisant la notion de transposée (proposition 6 de la page 99) et l'injection canonique ι (*cf.* proposition 3 de la page 97), qui est ici un isomorphisme (proposition 10 de la page 102). Nous conseillons au lecteur de *vérifier directement* la linéarité de v.
Finalement, pour tout $(x, y) \in E^2$, $\big(y \mid u(x)\big) = \big(v(y) \mid x\big)$ et v est l'adjoint de u. ∎

Proposition 25. Soit u un endomorphisme de l'espace euclidien E.
Soit $\mathcal{B}$ une base *orthonormée* de E. Si $\operatorname{Mat}_{\mathcal{B}}(u) := M$, alors $\operatorname{Mat}_{\mathcal{B}}(u^*) = {}^tM$.
On a $\operatorname{rang} u = \operatorname{rang} u^*$, $\det u = \det u^*$.

Démonstration.
Soit $M = (m_{ij}) = \operatorname{Mat}_{\mathcal{B}}(u)$. On a $m_{ij} = \big(e_i \mid u(e_j)\big) = \big(u^*(e_j) \mid e_i\big)$. ∎

Exercice 7.
On suppose que la base $\mathcal{B}$ de l'espace euclidien E n'est plus orthonormée. Quelle relation lie les matrices $\operatorname{Mat}_{\mathcal{B}}(u) = M$ et $\operatorname{Mat}_{\mathcal{B}}(u^*) = M'$?
Solution. On note $A := \operatorname{Mat}((\,|\,), \mathcal{B})$ la matrice du produit scalaire dans la base $\mathcal{B}$. Pour tout $(x, y) \in E^2$, $\big(u(x) \mid y\big) = {}^tX{}^tMAY$ et $\big(x \mid u^*(y)\big) = {}^tXAM'Y$ (avec nos conventions habituelles). d'où ${}^tMA = AM'$ et $M' = A^{-1}{}^tMA$.

Proposition 26. Soit u un endomorphisme d'un espace euclidien E.
On a $\operatorname{Ker} u^* = (\operatorname{Im} u)^\perp$ et $\operatorname{Im} u^* = (\operatorname{Ker} u)^\perp$. Un sous-espace $F \subset E$ est u-stable si, et seulement si, $F^\perp$ est u^*-stable.

Démonstration. On sait déjà : $\operatorname{Ker} u^* = (\operatorname{Im} u)^\perp$ et $\operatorname{Im} u^* \subset (\operatorname{Ker} u)^\perp$.
On a vu que $\operatorname{rang} u = \operatorname{rang} u^*$, mais $\operatorname{rang} u = \dim(\operatorname{Ker} u)^\perp$, d'où $\operatorname{Im} u^* = (\operatorname{Ker} u)^\perp$.
On sait que $F^\perp$ est u^*-stable. On a $(F^\perp)^\perp = F$ et $(u^*)^* = u$.
Si $F^\perp$ est u^*-stable, alors F est u-stable. ∎

En appliquant la proposition ci-dessus à un endomorphisme u de $\mathbb{R}^n$, on obtient :

Corollaire 27. Soit $M \in M_n(\mathbb{R})$. On a $\operatorname{Ker} {}^tM = (\operatorname{Im} M)^\perp$ et $\operatorname{Im} {}^tM = (\operatorname{Ker} M)^\perp$.

On utilise ce corollaire pour résoudre un système d'équations linéaires *au sens des moindres carrés* par la méthode dite *de l'équation normale*.

Equation normale. On considère un système linéaire $MX = Y$, où $y \in \mathbb{R}^n$ et $M \in M_{n,p}$ (avec $n > p$) *n'ayant pas de solution*. On se propose de le « résoudre au sens des moindres carrés » : on cherche $x \in \mathbb{R}^p$ tel que $\|Y - MX\|$ soit *minimal*. Si x est une solution de ce problème, $\|Y - MX\| = d(y, \operatorname{Im} M)$ et le vecteur de $\mathbb{R}^n$ défini par la matrice unicolonne $Y - MX$ est orthogonal à $\operatorname{Im} M$. Comme $(\operatorname{Im} M)^\perp = \operatorname{Ker} {}^tM$, ${}^tMMX = {}^tMY$, ce qui conduit à un nouveau système $AX = B$, avec $A = {}^tMM$ et $B = {}^tMY$ (l'équation normale). Dans ce nouveau système, la matrice A est maintenant *symétrique* et positive, définie positive si M est de rang p ; dans ce dernier cas on peut le résoudre en utilisant l'algorithme de Cholesky : on l'écrit : $AX = {}^tTTX = B$ et l'on résout le système *triangulaire* ${}^tTU = B$, puis le système *triangulaire* $TX = U$.

1.6 Groupe orthogonal

1.6.1 Généralités

Théorème et définition 28. Soit E un espace vectoriel préhilbertien réel. Pour tout automorphisme $u \in \mathcal{GL}(E)$ les assertions suivantes sont équivalentes.
(i) Pour tout $x \in E$, $\|u(x)\| = \|x\|$ (on dit que u conserve la norme).
(ii) Pour tout $(x, y) \in E^2$, $\big(\, u(x) \;\big|\; u(y) \,\big) = (\, x \mid y \,)$ (on dit que u conserve le produit scalaire).
(iii) L'automorphisme u a un adjoint u^* et $u^* = u^{-1}$.
Si u vérifie l'une quelconque de ces assertions, on dit que u est un automorphisme *orthogonal* de E. On note $\mathcal{O}(E)$ l'ensemble des automorphismes orthogonaux de E.

Démonstration. On va montrer que (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (i).

(i) $\Rightarrow$ (ii). C'est une application immédiate de l'égalité de polarisation. On suppose que (i) est vraie. Pour tout $(x, y) \in E^2$:

$$\begin{aligned}(\, x \mid y \,) &= \frac{1}{2}\big(\|x+y\|^2 - \|x\|^2 - \|y\|^2\big) \\ &= \frac{1}{2}\big(\big\|u(x+y)\big\|^2 - \big\|u(x)\big\|^2 - \big\|u(y)\big\|^2\big) = \big(\, u(x) \;\big|\; u(y) \,\big).\end{aligned}$$

(ii) $\Rightarrow$ (iii). Pour tout $(x, y) \in E^2$: $\big(\, u(x) \;\big|\; y \,\big) = \big(\, u(x) \;\big|\; u\big(u^{-1}(y)\big) \,\big)$. On suppose que (ii) est vraie, alors $\big(\, u(x) \;\big|\; u(u^{-1}(y)) \,\big) = \big(\, x \;\big|\; u^{-1}(y) \,\big)$ et $\big(\, u(x) \;\big|\; y \,\big) = \big(\, x \;\big|\; u^{-1}(y) \,\big)$, ce qui montre que u^{-1} est l'adjoint de u.

(iii) $\Rightarrow$ (i). On suppose que (iii) est vraie. Pour tout $x \in E$:

$$\big\|u(x)\big\|^2 = \big(\, u(x) \;\big|\; u(x) \,\big) = \big(\, x \;\big|\; u^*(u(x)) \,\big) = \big(\, x \;\big|\; u^{-1}(u(x)) \,\big) = (\, x \mid x \,) = \|x\|^2.$$

Remarque. Si $u \in \mathcal{O}(E)$, on a :

$$\begin{aligned}\|x-y\|^2 &= \|x\|^2 + \|y\|^2 - 2(\, x \mid y \,) \\ &= \big\|u(x)\big\|^2 + \big\|u(y)\big\|^2 - 2\big(\, u(x) \;\big|\; u(y) \,\big) = \big\|u(x)-u(y)\big\|^2.\end{aligned}$$

Si d est la distance euclidienne de E : $d\big(u(x),u(y)\big) = d(x,y)$, u est donc une *isométrie* de E (elle admet 0 pour point fixe).

Exercice 8.

Montrer que toute *bijection* u de l'espace préhilbertien E qui conserve le produit scalaire est un automorphisme orthogonal de E.

Solution. Il s'agit de montrer que l'application u est *linéaire*. Soit $(\alpha, x, y) \in \mathbb{R} \times E^2$. Considérons $z = u(\alpha x + y) - \alpha u(x) - u(y)$. Pour tout vecteur $t \in E$:

$$\begin{aligned}(\, z \mid t \,) &= \big(\, u(\alpha x + y) \;\big|\; t \,\big) - \alpha\big(\, u(x) \;\big|\; t \,\big) - \big(\, u(y) \;\big|\; t \,\big) \\ &= \big(\, \alpha x + y \;\big|\; u^{-1}(t) \,\big) - \alpha\big(\, x \;\big|\; u^{-1}(t) \,\big) - \big(\, y \;\big|\; u^{-1}(t) \,\big) = 0.\end{aligned}$$

Donc $z = 0$ et u est linéaire.

Proposition 29. Soient E un espace préhilbertien réel et $u \in \mathcal{O}(E)$. Pour tout sous-espace $F \subset E$ *invariant* par u, $F^\perp$ est *stable* par u.

Démonstration. On rappelle que F est *invariant* par u si, et seulement si, $u(F) = F$. Si F est invariant par u, F est u^{-1}-stable, donc $F^\perp$ est $(u^{-1})^*$-stable, donc u-stable, car $u = (u^{-1})^*$.

Remarque. Soient E un espace préhilbertien réel , $u \in \mathcal{O}(E)$ et F un sous-espace de E u-stable. Si F est de dimension *finie*, il est alors u-*invariant* (on utilise le théorème du rang).

Proposition 30. L'ensemble $\mathcal{O}(E)$ des automorphismes orthogonaux de l'espace préhilbertien réel E est un *groupe*. On l'appelle *le groupe orthogonal* de E.

Démonstration. On montre que $\mathcal{O}(E)$ est un sous-groupe du groupe $\mathcal{GL}(E)$: $\mathcal{O}(E)$ contient l'identité, par ailleurs, pour tous $u_1, u_2 \in \mathcal{O}(E)$, $u_1^{-1}u_2 \in \mathcal{O}(E)$. En effet u_1^{-1} admet u_1 pour adjoint, donc $u_1^{-1}u_2$ a un adjoint et :

$$\big(u_1^{-1}u_2\big)^* = u_2^*\big(u_1^{-1}\big)^* = u_2^*u_1 = \big(u_1^{-1}u_2\big)^{-1}.$$

1.6.2 Groupe orthogonal d'un espace euclidien

Un espace euclidien étant de dimension finie, on peut simplifier la définition 28 de la page 165.

Proposition 31. Soit u un *endomorphisme* de l'espace *euclidien* E, u est un automorphisme orthogonal si et seulement s'il conserve la norme.

Démonstration. Supposons que u conserve la norme.
Pour tout $x \in \operatorname{Ker} u$, $\|x\| = \big\|u(x)\big\| = 0$, u est donc injective et par suite bijective. ∎

Proposition 32. Soit u un *endomorphisme* de l'espace euclidien E, les assertions suivantes sont équivalentes :
(i) $u \in \mathcal{O}(E)$;
(ii) l'image par u de toute base orthonormée de E est une base orthonormée ;
(iii) il existe une base orthonormée de E dont l'image par u est une base orthonormée.

Démonstration. (i) $\Rightarrow$ (ii). Soit $u \in \mathcal{O}(E)$, soit $\mathcal{B} = (e_1, \dots, e_n)$ une base orthonormée de E. Pour tout $(i,j) \in [\![1,n]\!]^2$, le produit scalaire $\big(\, u(e_i) \;\big|\; u(e_j) \,\big) = (\, e_i \mid e_j \,) = \delta_{ij}$. La base $\big(u(e_1), \dots, u(e_n)\big)$ est donc orthonormée.
(ii) $\Rightarrow$ (iii) est immédiat. Montrons (iii) $\Rightarrow$ (i). Soit $\mathcal{B} := (e_1, \dots, e_n)$ une base orthonormée de E dont l'image est orthonormée. Pour tout $x \in E$, $x = \sum_{i=1}^{n} (\, e_i \mid x \,)\, e_i$, donc $u(x) = \sum_{i=1}^{n} (\, e_i \mid x \,)\; u(e_i)$; $\big(u(e_1), \dots, u(e_n)\big)$ étant une base orthonormée, on a : $\big\|u(x)\big\|^2 = \sum_{i=1}^{n} (\, e_i \mid x \,)^2 = \|x\|^2$. ∎

Théorème et définition 33. Soit E un espace euclidien de dimension strictement positive, l'application qui associe à tout élément de $\mathcal{O}(E)$ son déterminant est un morphisme de $\mathcal{O}(E)$ dans le sous groupe multiplicatif $\{1, -1\}$ de $\mathbb{R}^*$. Le noyau de ce morphisme est un sous-groupe de $\mathcal{O}(E)$ dit *groupe spécial orthogonal* de E, ou groupe des rotations de E. Il est noté $\mathcal{SO}(E)$ ou $\mathcal{O}^+(E)$. On note $\mathcal{O}^-(E)$ l'ensemble des éléments de E de déterminant égal à -1, $\mathcal{O}^-(E)$ est dit *ensemble des isométries négatives* de $\mathcal{O}(E)$.

Démonstration. Soit $u \in \mathcal{O}(E)$, on a $uu^* = \mathrm{Id}$ et $\det u = \det u^*$, d'où $|\det u|^2 = 1$ et $\det u = \pm 1$ puisque $\det u \in \mathbb{R}$. ∎

Exemples.
(i) Si l'espace vectoriel euclidien E est de dimension paire (respectivement impaire) : $-\mathrm{Id}_E \in \mathcal{SO}(E)$ $\big($resp. $-\mathrm{Id}_E \in \mathcal{O}^-(E)\big)$. Si E est de dimension impaire, l'application $u \mapsto -u$ est une bijection de $\mathcal{SO}(E)$ sur $\mathcal{O}^-(E)$.

(ii) Si l'espace vectoriel euclidien E est de dimension 3, toute symétrie orthogonale par rapport à une droite vectorielle (on dit aussi *retournement*) appartient à $\mathcal{SO}(E)$. (La matrice de cette symétrie dans une base orthonormée convenable, que l'on laisse expliciter au lecteur, est $\mathrm{Diag}(1,-1,-1)$.

(iii) Toute symétrie orthogonale hyperplane d'un espace euclidien E (de dimension strictement positive) est un élément de $\mathcal{O}^-(E)$. (Une symétrie orthogonale hyperplane est aussi appelée *réflexion*.)
Le résultat est vrai si $\dim E = 1$, on peut donc supposer $n = \dim E > 1$.

Soit $E = H \overset{\perp}{\oplus} \mathbb{R}v$, H étant un hyperplan de E et v un vecteur unitaire, orthogonal à H. Soit $(e_1, \ldots, e_{n-1})$ une base orthonormée de H ; $\mathcal{B} := (e_1, \ldots, e_{n-1}, v)$ est une base orthonormée de E. Soit s la symétrie orthogonale par rapport à H, elle transforme la base orthonormée $\mathcal{B}$ en la base orthonormée $(e_1, \ldots, e_{n-1}, -v)$, elle est donc orthogonale et son déterminant est égal à -1.

Proposition 34. Soit u un automorphisme orthogonal de l'espace euclidien E. Si $\lambda \in \mathbb{R}$ est une valeur propre de u, alors $\lambda = \pm 1$.

Démonstration. Soit $x \in E$ un vecteur propre associé à λ : $\big\|u(x)\big\| = |\lambda|\|x\| = \|x\|$. Donc $|\lambda| = 1$, car $x \neq 0$ ∎

1.6.3 Matrices orthogonales

Nous nous intéressons ici au groupe orthogonal $\mathcal{O}(\mathbb{R}^n)$, l'espace $\mathbb{R}^n$ $(n > 0)$ étant muni de son produit scalaire canonique. C'est à un isomorphisme (non canonique) près le cas général : si E est un espace euclidien de dimension n, le choix d'une base orthonormée de E donne un isomorphisme de groupes entre $\mathcal{O}(E)$ et $\mathcal{O}(\mathbb{R}^n)$. Cette étude sera approfondie dans le module I.6.

Théorème et définition 35. Soit $\mathcal{B}$ la base canonique de $\mathbb{R}^n$, l'application qui, à tout automorphisme orthogonal de $\mathbb{R}^n$, associe sa matrice dans la base $\mathcal{B}$ est un homomorphisme injectif de groupes de $\mathcal{O}(\mathbb{R}^n)$ dans $GL_n(\mathbb{R})$. Son image est appelée *groupe des matrices orthogonales* de $M_n(\mathbb{R})$, et notée $O_n(\mathbb{R})$. Le groupe $O_n(\mathbb{R})$ est isomorphe au groupe $\mathcal{O}(\mathbb{R}^n)$. L'image de $\mathcal{O}^+(\mathbb{R}^n)$ est notée $SO_n(\mathbb{R})$ ou $O_n^+(\mathbb{R})$, de même l'image de $\mathcal{O}^-(\mathbb{R}^n)$ est notée $O_n^-(\mathbb{R})$; $SO_n(\mathbb{R})$ est un sous-groupe du groupe spécial linéaire $SL_n(\mathbb{R})$.

La démonstration est laissée au lecteur, de même que celle de la proposition suivante.

Proposition 36. Soit $P \in M_n(\mathbb{R})$, les conditions suivantes sont équivalentes :
(i) la matrice P est orthogonale ;
(ii) $P^tP = I_n$;
(iii) ${}^tPP = I_n$;
(iv) les vecteurs colonnes de P forment une base orthonormée de $\mathbb{R}^n$;
(v) les vecteurs lignes de P forment une base orthonormée de $\mathbb{R}^n$.

Exemple. Toute *matrice de permutation* est une matrice orthogonale. Soit $s \in \mathfrak{S}_n$, soit $\mathcal{B} = (e_1, \dots, e_n)$ la base canonique de $\mathbb{R}^n$. On définit $u \in \mathcal{L}(\mathbb{R}^n)$ par $u(e_j) = e_{s(j)}$, pour tout $j \in [\![1, n]\!]$. On a $u \in \mathcal{O}(\mathbb{R}^n)$, en effet quelque soit $(i, j) \in [\![1, n]\!]^2$:

$$\big(\, u(e_i) \;\big|\; u(e_j) \,\big) = \big(\, e_{s(i)} \;\big|\; e_{s(j)} \,\big) = \delta_{ij}.$$

La matrice P associée à u (i. e. la matrice de permutation associée à la permutation s) est une matrice orthogonale. À titre d'exercice, le lecteur en déduira, en utilisant le théorème de Cayley (*cf.* module I.2, théorème 2 de la page 61), que tout groupe *fini* G est isomorphe à un sous-groupe de $O_n(\mathbb{R})$, pour un entier n convenable (dépendant de G).

Étude de $\mathcal{O}_2(\mathbb{R})$

La matrice $P := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ est orthogonale si, et seulement si, $\begin{cases} a^2 + c^2 &= 1 \\ b^2 + d^2 &= 1 \\ ab + cd &= 0. \end{cases}$

Il existe donc $\theta, \varphi \in \mathbb{R}$ tels que $a = \cos\theta$, $c = \sin\theta$, $b = \sin\varphi$, $d = \cos\varphi$.
On a $0 = \cos\theta\sin\varphi + \sin\theta\cos\varphi = \sin(\varphi + \theta)$, *i.e.* $\varphi + \theta = 0$ (mod. π).
Si $\varphi = -\theta$ (mod. 2π), on note $R(\theta) := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \in \mathcal{O}_2^+(\mathbb{R})$.
Si $\varphi = \pi - \theta$ (mod. 2π), on note $S(\theta) := \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \in \mathcal{O}_2^-(\mathbb{R})$.
Ainsi $\mathcal{O}_2^+(\mathbb{R}) = \{R(\theta) | \theta \in \mathbb{R}\}$ et $\mathcal{O}_2^-(\mathbb{R}) = \{S(\theta) | \theta \in \mathbb{R}\}$.
Le groupe $\mathcal{O}_2^+(\mathbb{R})$ est le groupe des rotations de $\mathbb{R}^2$, $R(\theta)$ est la matrice associée à la *rotation* r dont un angle est θ. On a $S^2(\theta) = I$, la matrice S est donc la matrice d'une symétrie s, et de ${}^tS(\theta) = S(\theta) = S^{-1}(\theta)$ on déduit que s est une *symétrie orthogonale*.
Le lecteur vérifiera que :

$$S(\theta) = \begin{pmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \cos\frac{\theta}{2} & \sin\frac{\theta}{2} \\ -\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}.$$

1.7 Endomorphismes symétriques et applications

1.7.1 Généralités

Théorème et définition 37. Soit E un espace préhilbertien réel, pour tout endomorphisme u de E, les deux assertions suivantes sont équivalentes :
(i) pour tout $(x, y) \in E^2$, $\big(\, u(x) \;\big|\; y \,\big) = \big(\, x \;\big|\; u(y) \,\big)$ (resp. $\big(\, u(x) \;\big|\; y \,\big) = -\big(\, x \;\big|\; u(y) \,\big)$) ;
(ii) u admet un adjoint u^* et $u^* = u$ (resp. $u^* = -u$).
Si u vérifie l'une de ces deux propriétés, on dit que u est un endomorphisme *symétrique* ou *auto-adjoint* (resp. *antisymétrique*).

La vérification est laissée au lecteur.

Proposition 38. L'endomorphisme u de l'espace préhilbertien E est antisymétrique si, et seulement si, pour tout $x \in E$, $\big(\, u(x) \;\big|\; x \,\big) = 0$.

Démonstration. Si u est antisymétrique $\big(\, u(x) \;\big|\; x \,\big) = -\big(\, x \;\big|\; u(x) \,\big) = -\big(\, u(x) \;\big|\; x \,\big)$, donc $2\big(\, u(x) \;\big|\; x \,\big) = 0$ et $\big(\, u(x) \;\big|\; x \,\big) = 0$.
Inversement, si, pour tout $x \in E$, $\big(\, u(x) \;\big|\; x \,\big) = 0$, alors, pour tout $(x, y) \in E^2$, on a :

$$\big(\, u(x+y) \;\big|\; x+y \,\big) = 0 = \big(\, u(x) \;\big|\; x \,\big) + \big(\, u(x) \;\big|\; y \,\big) + \big(\, u(y) \;\big|\; x \,\big) + \big(\, u(y) \;\big|\; y \,\big)$$

d'où $\big(\, u(x) \;\big|\; y \,\big) + \big(\, u(y) \;\big|\; x \,\big) = 0$. ∎

Proposition 39. L'ensemble $\mathbf{S}(E)$ des endomorphismes symétriques de l'espace préhilbertien E et l'ensemble $\mathbf{A}(E)$ des endomorphismes antisymétriques de E sont des sous-espaces vectoriels supplémentaires de l'espace vectoriel des endomorphismes de E ayant un adjoint.

Démonstration. On vérifie sans difficulté que les ensembles $\mathbf{S}(E)$ et $\mathbf{A}(E)$ sont des sous-espaces vectoriels de l'espace vectoriel des endomorphismes de E ayant un adjoint. Pour tout $u \in \mathcal{L}(E)$ ayant un adjoint, on a $u = \dfrac{u+u^*}{2} + \dfrac{u-u^*}{2}$; l'endomorphisme $\dfrac{u+u^*}{2}$ est symétrique et l'endomorphisme $\dfrac{u-u^*}{2}$ est antisymétrique. De plus $\mathbf{S}(E) \cap \mathbf{A}(E) = \{0\}$. En effet $u \in \mathbf{S}(E) \cap \mathbf{A}(E)$ si, et seulement si, pour tous $x, y \in E$, $(\, u(x) \mid y \,) = (\, x \mid u(y) \,) = -(\, x \mid u(y) \,)$, donc $(\, u(x) \mid y \,) = 0$. On en déduit $u = 0$. ∎

Exercice 9.

Toute *application* u de l'espace préhilbertien E dans lui même, telle que :

$$\forall (x, y) \in E^2 \,,\; (\, u(x) \mid y \,) = (\, x \mid u(y) \,),$$

est un *endomorphisme* symétrique de E.
Solution. Il suffit de montrer que l'application u est *linéaire*.
Montrons que, pour tout $(\alpha, x, y) \in \mathbb{R} \times E^2$, $u(\alpha x + y) - \alpha u(x) - u(y) = 0$. Ceci équivaut à $\big(\, u(\alpha x + y) - \alpha u(x) - u(y) \;\big|\; z \,\big) = 0$, pour tout $z \in E$, or :

$$\begin{aligned} \big(\, u(\alpha x + y) \;\big|\; z \,\big) &= \big(\, \alpha x + y \;\big|\; u(z) \,\big) = \alpha \big(\, x \;\big|\; u(z) \,\big) + \big(\, y \;\big|\; u(z) \,\big) \\ &= \big(\, \alpha u(x) \;\big|\; z \,\big) + \big(\, u(y) \;\big|\; z \,\big) = \big(\, \alpha u(x) + u(y) \;\big|\; z \,\big), \end{aligned}$$

d'où le résultat.

Exercice 10.

Soient $u, v \in \mathbf{S}(E)$, l'endomorphisme uv est-il symétrique ?
Solution. On a $(uv)^* = v^* u^* = vu$. Ainsi $(uv)^* = uv$ si, et seulement si, $uv = vu$, c'est à dire si, et seulement si, u et v *commutent.*

Endomorphismes symétriques positifs, définis positifs, d'un espace préhilbertien réel

Définition 8. Soit E un espace préhilbertien réel, un endomorphisme $u \in \mathbf{S}(E)$ est dit *positif* (resp. *défini positif*) si, pour tout $x \in E$ (resp. $x \in E \setminus \{0\}$), $(\,u(x) \mid x\,) \geqslant 0$ (resp. $(\,u(x) \mid x\,) > 0$). On note $\mathbf{S}^+(E)$ l'ensemble des endomorphismes symétriques positifs, $\mathbf{S}^{++}(E)$ l'ensemble des endomorphismes symétriques définis positifs.

On montre sans difficulté le résultat suivant.

Proposition 40. Soit E un espace préhilbertien, pour tout $u \in \mathbf{S}(E)$, l'applicationv de E^2 dans $\mathbb{R}$, qui associe à (x, y) le réel $\big(\,u(x) \;\big|\; y\,\big)$ est une forme bilinéaire symétrique sur E, *i.e.* un élément de $\mathcal{S}(E)$.
Elle est positive (resp. définie positive) si, et seulement si, u l'est.

Proposition 41. Soit E un espace préhilbertien, pour tout $u \in \mathcal{L}(E)$ ayant un adjoint u^*, les endomorphismes u^*u et uu^* sont symétriques *positifs*, u^*u est défini positif si, et seulement si, u est *injectif*, de même uu^* est défini positif si, et seulement si, u^* est *injectif*.

Démonstration. Les endomorphismes u^*u et uu^* sont évidemment symétriques, de plus $\big(\,u^*u(x) \;\big|\; x\,\big) = \big(\,u(x) \;\big|\; u(x)\,\big) = \big\|u(x)\big\|^2 \geqslant 0$. Ainsi u^*u est positif ; il est défini positif si, et seulement si, $u(x) \neq 0$, pour tout $x \in E$ non nul, c'est à dire si, et seulement si, u est injectif. On termine en échangeant u et u^*. ∎

Si $u \in \mathbf{A}(E)$, $u^2 = -uu^*$, donc $-u^2$ est un endomorphisme positif. D'où le :

Corollaire 42. Pour tout $u \in \mathbf{A}(E)$, $-u^2 \in \mathbf{S}^+(E)$.

1.7.2 Projecteurs orthogonaux — Symétries orthogonales

Proposition 43. Soit p un projecteur de l'espace préhilbertien E, p est un projecteur *orthogonal* si, et seulement si, p est *symétrique*.
Soit s une symétrie de l'espace préhilbertien E. Les assertions suivantes sont équivalentes :
(i) s est une symétrie *orthogonale* ;
(ii) s est *symétrique* ;
(iii) s est un *automorphisme orthogonal* de E.

On prendra garde au fait qu'un projecteur différent de l'identité n'est pas un automorphisme orthogonal de E : il n'est pas injectif (il contracte la norme, mais ne la conserve pas).

Démonstration. Soit p un projecteur de E. Si p est orthogonal, pour tous $x, y \in E$, $\big(\,p(x) \;\big|\; y\,\big) = \big(\,p(x) \;\big|\; y - p(y) + p(y)\,\big) = \big(\,p(x) \;\big|\; p(y)\,\big) = \big(\,x \;\big|\; p(y)\,\big)$, p est donc sy-

métrique. Inversement, supposons que p soit symétrique. Pour tout $(x,y) \in \operatorname{Im} p \times \operatorname{Ker} p$: $(\,x \mid y\,) = \big(\,p(x) \;\big|\; y\,\big) = \big(\,x \;\big|\; p(y)\,\big) = 0$, ainsi p est orthogonal.
Soit s une symétrie de E.

- Montrons (i) $\Rightarrow$ (ii). Supposons s orthogonale, on sait qu'elle s'écrit $s = 2p - \mathrm{Id}$, p étant un projecteur orthogonal (*cf.* remarques de la page 150). Celui-ci étant symétrique, s l'est aussi (*cf.* proposition 39 de la page 170).
- Montrons (ii) $\Rightarrow$ (iii). Supposons s symétrique. On a $s^{-1} = s = s^*$ donc $s \in \mathcal{O}(E)$.
- Montrons (iii) $\Rightarrow$ (i). Supposons s orthogonale.
  Pour tous $x \in \operatorname{Ker}(s - \mathrm{Id})$ et $y \in \operatorname{Ker}(s + \mathrm{Id})$, on a :
  $$(\,x \mid y\,) = \big(\,s(x) \;\big|\; s(y)\,\big) = (\,x \mid -y\,) = -(\,x \mid y\,),$$
  donc $(\,x \mid y\,) = 0$; s est donc orthogonale. ∎

Symétries orthogonales hyperplanes d'un espace préhilbertien — Algorithme QR

Proposition 44. Soit v un vecteur *non nul* de l'espace préhilbertien E. Soit s la symétrie orthogonale par rapport à *l'hyperplan* H orthogonal à v.
Pour tout $x \in E$, on a : $s(x) = x - \dfrac{2(\,v \mid x\,)}{\|v\|^2} v$.

Démonstration. Soit p la projection orthogonale sur l'hyperplan H ($\operatorname{Ker} p = \mathbb{R}v$). Soit q la projection orthogonale sur la droite $\mathbb{R}v$ ($\operatorname{Ker} q = H$). On a $p + q = \mathrm{Id}$ et $s = 2p - \mathrm{Id}$, d'où $s = \mathrm{Id} - 2q$. Les points $0, x, x + s(x), s(x)$ forment un *losange* de centre $p(x)$.
On a : $q(x) = \dfrac{(\,v \mid x\,)}{(\,v \mid v\,)} v$, d'où $s(x) = x - 2q(x) = x - \dfrac{2(\,v \mid x\,)}{\|v\|^2} v$. ∎

Proposition 45.
Soient E un espace vectoriel euclidien de dimension $n > 0$ et $\mathcal{B} := (e_1, \ldots, e_n)$ une base orthonormée de E. Un endomorphisme s de E est une *symétrie orthogonale hyperplane* si, et seulement s'il existe une *matrice unicolonne* $V \in \mathcal{M}_{n,1}(\mathbb{R})$ telle que $\operatorname{Mat}_{\mathcal{B}}(s) = I_n - 2\dfrac{V\,{}^tV}{{}^tVV}$.

Démonstration. Soit s une symétrie orthogonale hyperplane par rapport à H et v un vecteur non nul orthogonal à H : $v := \sum_{i=1}^{n} v_i e_i$. On note, comme d'habitude, X, Y, V les matrices unicolonnes des coordonnées de $x \in E$, $y := s(x)$ et v, on a : $Y = X - 2\dfrac{{}^tVX}{{}^tVV} V$.
Le *nombre réel* ${}^tVX$ commute avec V, d'où : $Y = X - 2\dfrac{V\,{}^tVX}{{}^tVV} = \left(I_n - 2\dfrac{V\,{}^tV}{{}^tVV}\right) X$
(on considère le réel ${}^tVX$ comme une matrice carrée d'ordre un et on utilise l'associativité de la multiplication des matrices). La réciproque est immédiate. ∎

Définition 9 (Matrices de Householder). Soient $v \in \mathbb{R}^n$ un vecteur non nul et $V \in \mathcal{M}_{n,1}(\mathbb{R})$ la matrice unicolonne associée. On appelle *matrice de Householder* d'ordre n associée à v la matrice $H_v = I_n - 2\dfrac{V\,{}^tV}{{}^tVV}$.

On convient de considérer la matrice identité I_n comme la matrice de Householder associée à $v = 0$: $H_0 = I_n$.

Proposition 46. Soit (a, b) un couple de vecteurs d'un espace préhilbertienréel E, *distincts*, et *de même longueur* . Il existe une unique symétrie orthogonale hyperplane s de E telle que $s(a) = b$. L'axe de s est l'hyperplan $(b-a)^\perp$.

Démonstration.

Si s existe, pour tout $x \in E$, $s(x) = x - \dfrac{2(\,v \mid x\,)}{\|v\|^2}v$ et $b = s(a) = a - \dfrac{2(\,v \mid a\,)}{\|v\|^2}v$, ainsi v et $b-a$ sont *colinéaires*. Les vecteurs a et b étant distincts, l'axe de la symétrie s est $(b-a)^\perp$; s est donc unique. Vérifions que la symétrie s d'axe $(b-a)^\perp$ est bien solution. On a : $s(x) = x - \dfrac{2(\,b-a \mid x\,)}{\|b-a\|^2}(b-a)$, donc $s(a) = a - \dfrac{2(\,b-a \mid a\,)}{\|b-a\|^2}(b-a)$.

On remarque :

$$(b-a, b-a) = \|a\|^2 + \|b\|^2 - 2(\,a \mid b\,) = 2(\|a\|^2 - (\,a \mid b\,)) = -2(\,b-a \mid a\,),$$

d'où $s(a) = a + b - a = b$.

Géométriquement $(b-a)^\perp$ est *l'hyperplan médiateur* du segment joignant a à b. ∎

Décomposition QR d'une matrice carrée à coefficients réels.

Notations : $(e_1, \ldots, e_n)$ est la base canonique de $\mathbb{R}^n$, $F_k := \mathrm{Vect}(e_k, \ldots, e_n)$, pour $k \in [\![1, n]\!]$.

Théorème 47 (Décomposition QR et méthode de Householder).
Soit $A \in M_n(\mathbb{R})$.
(i) Il existe un couple (Q, R) formé d'une matrice *orthogonale* Q et d'une matrice *triangulaire supérieure* R tel que $A = QR$. Si A est *inversible* et si l'on impose de plus aux éléments diagonaux de R d'être *positifs*, une telle décomposition est *unique*.
(ii) Il existe des vecteurs $v_1, \ldots, v_{n-1} \in \mathbb{R}^n$ tels que $H_{v_{n-1}} \ldots H_{v_1} A$ soit une matrice triangulaire supérieure. De plus on peut choisir les v_k tels que $v_k \in F_k$, pour $k \in [\![1, n-1]\!]$.

Démonstration. Montrons l'assertion (ii) (l'existence dans l'assertion (i) s'en déduit). Remarquons tout d'abord que toute symétrie orthogonale hyperplane s de $\mathbb{R}^n = F_1$ échangeant deux vecteurs distincts a et b de F_{k+1} laisse *invariants* les vecteurs $e_1, \ldots, e_k$ (ces vecteurs sont orthogonaux à F_{k+1}, donc à $b-a$). La matrice associée à s dans la base canonique est donc de la forme $\begin{pmatrix} I_k & 0_{k,n-k} \\ 0_{n-k,k} & H'_k \end{pmatrix}$, où H'_k est la matrice associée à la

restriction de s à F_{k+1} dans la base $e_{k+1}, \dots, e_n$. L'idée de base de la démonstration de (ii) est similaire à celle que nous avons employée dans la décomposition LU : on multiplie *à gauche* la matrice donnée par une matrice *inversible* H de façon à *remplacer* le premier vecteur colonne par un nouveau vecteur dont les $n-1$ dernières coordonnées sont nulles. Dans le cas de LU, H était une *matrice de transvection*, ici elle sera une *matrice de Householder* (l'algorithme que nous allons décrire est dû à Householder). Soit $C_1 \in \mathbb{R}^n = F_1$ le premier vecteur colonne de A. S'il est colinéaire à e_1, on pose $v_1 := 0$, $H_{v_1} := H_0$ ou $v_1 := e_1$, $H_{v_1} := H_{e_1}$ si l'on souhaite changer le signe. Sinon, l'on sait qu'il existe une symétrie hyperplane échangeant C_1 et, au choix, l'un des deux vecteurs $\pm\|C_1\|e_1$. On note H_{v_1} la matrice de Householder correspondante (remarquons que l'on peut faire le choix systématique du signe $+$). On pose $\delta_1 = \|C_1\|$. On a donc $H_{v_1}A = \begin{pmatrix} \pm\delta_1 & \star & \dots & \star \\ 0 & & & \\ \vdots & & A_1 & \\ 0 & & & \end{pmatrix}$.

Soit C_2 le premier vecteur colonne de A_1, que nous interprétons comme un vecteur de F_2. S'il est colinéaire à e_2, on pose $v_2 = 0$, $H_{v_2} = H_0$ ou $v_2 = e_2$, $H_{v_2} = H_{e_2}$ si l'on souhaite changer le signe. Sinon l'on sait qu'il existe une une symétrie hyperplane échangeant C_2 et, au choix, l'un des deux vecteurs $\pm\|C_2\|e_2$. Le vecteur e_1 est invariant par cette symétrie (*cf.* remarque initiale). On note H_{v_2} la matrice de Householder correspondante. On pose $\delta_2 = \|C_2\|$. On a alors : $H_{v_2}H_{v_1}A = \begin{pmatrix} \pm\delta_1 & \star & \star & \dots & \star \\ 0 & \pm\delta_2 & \star & \dots & \star \\ 0 & 0 & & & \\ \vdots & \vdots & & A_2 & \\ 0 & 0 & & & \end{pmatrix}$.

Le lecteur vérifiera que la première ligne de $H_{v_1}A$ n'a pas été modifiée.

En itérant la méthode, on peut ainsi déterminer successivement $H_{v_3}, \dots, H_{v_{n-1}}$ de telle sorte que la matrice $H_{v_{n-1}} \dots H_{v_1}A$ soit une matrice triangulaire supérieure, dont la diagonale est $(\pm\delta_1, \dots, \pm\delta_{n-1}, \star)$ (nous laissons la formalisation de la récurrence au lecteur). Cela termine la démonstration de (ii). On peut de plus changer le signe du dernier terme diagonal $\star$, si ce terme n'est pas nul, en multipliant à gauche par la matrice de Householder H_{e_n}. L'existence de la décomposition (i) en résulte aussitôt.

Il reste à prouver l'unicité dans le cas où la matrice A est inversible. Soient $A = QR$ et $A = Q'R'$ deux décompositions de A, les termes diagonaux de R et R' étant tous positifs. On en déduit $RR'^{-1} = Q^{-1}Q'$, ainsi la matrice triangulaire supérieure RR'^{-1} est orthogonale, donc *diagonale* à termes ± 1 (*cf.* l'exercice ci-dessous). Ces termes étant tous positifs, RR'^{-1} est l'identité et les deux décompositions sont égales. ∎

Exercice 11.

Montrer que toute matrice triangulaire supérieure et orthogonale est *diagonale* à termes ± 1.

Solution. Soit $T_n := (t_{ij})_{i,j\in[\![1,n]\!]}$ une matrice triangulaire supérieure et orthogonale. Le premier vecteur colonne étant unitaire, on a nécessairement $t_{11} = \pm 1 = \varepsilon$. Si $n := 1$, la démonstration est finie, sinon l'on écrit $T_n = \begin{pmatrix} \varepsilon & {}^tV \\ 0 & T_{n-1} \end{pmatrix}$, où V est une matrice colonne à $(n-1)$ éléments et $T_{n-1} \in M_{n-1}$ est une matrice triangulaire supérieure. On a :

$$I_n = T_n {}^tT_n = \begin{pmatrix} 1 + {}^tVV & {}^tV{}^tT_{n-1} \\ T_{n-1}V & T_{n-1}{}^tT_{n-1} \end{pmatrix}.$$

Ainsi $V = 0$ et T_{n-1} est triangulaire supérieure et orthogonale. On termine la preuve par récurrence.

Corollaire 48. Soit E un espace euclidien de dimension $n \geqslant 2$. Tout automorphisme *orthogonal* u de E peut s'écrire comme le produit d'au plus n symétries orthogonales hyperplanes. Si n est pair (resp. impair) et si $u \in \mathcal{O}^-(E)$ (resp. $u \in \mathcal{O}^+(E)$), il peut s'écrire comme le produit d'au plus $n-1$ symétries orthogonales hyperplanes.

Démonstration. Le choix d'une base orthonormée de E permet de se ramener au problème matriciel analogue, les symétries orthogonales étant remplacées par des matrices de Householder $\neq I_n$. Soit A une matrice orthogonale, on lui applique l'algorithme de Householder, en choisissant le signe $+$ à chaque étape. On obtient ainsi une décomposition $A = H_{v_1} \dots H_{v_{n-1}} R$. La matrice R est triangulaire supérieure et orthogonale, donc diagonale, ses termes étant tous égaux à ± 1. D'après nos choix, les $n-1$ premiers termes sont tous positifs et donc égaux à 1 : $R = \mathrm{Diag}(1, 1, \dots 1, \varepsilon)$. Si $\varepsilon := 1$, A est le produit d'au plus $n-1$ matrices de Householder différentes de l'identité. Si $\varepsilon := -1$, $R = H_{e_n}$ et A est le produit d'au plus n matrices de Householder différentes de l'identité. Si A est le produit de k matrices de Householder différentes de l'identité, $\det A = (-1)^k$. Ainsi, si $k = n$, $\det A = (-1)^n$ et $A \in \mathcal{O}_n^-$ si n est impair, $A \in \mathcal{O}_n^+$ si n est pair. ∎

Exemple. Appliquons la méthode de Householder à la matrice :

$$A := \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}.$$

Nous ne donnons que les résultats intermédiaires, laissant au lecteur les détails élémentaires (mais laborieux…) du calcul.

Écrivons les matrices successives obtenues à chaque étape de l'algorithme (nous n'explicitons pas les matrices de Householder H_{v_i} $(i = 1, 2, 3)$ utilisées, mais seulement les vecteurs v_i correspondants) :

$$v_1 := \begin{pmatrix} 2 - \sqrt{5} \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad v_2 := \begin{pmatrix} 0 \\ -\frac{3+\sqrt{14}}{\sqrt{5}} \\ 1 \\ 0 \end{pmatrix} \quad v_3 := \begin{pmatrix} 0 \\ 0 \\ \frac{4-\sqrt{30}}{\sqrt{14}} \\ 1 \end{pmatrix}$$

$$H_{v_1}A=\begin{pmatrix}\sqrt{5}&\frac{4}{\sqrt{5}}&\frac{1}{\sqrt{5}}&0\\0&-\frac{3}{\sqrt{5}}&-\frac{2}{\sqrt{5}}&0\\0&1&2&1\\0&0&1&2\end{pmatrix}\qquad H_{v_2}H_{v_1}A=\begin{pmatrix}\sqrt{5}&\frac{4}{\sqrt{5}}&\frac{1}{\sqrt{5}}&0\\0&\sqrt{\frac{14}{5}}&\frac{16}{\sqrt{70}}&\frac{5}{\sqrt{70}}\\0&0&\frac{4}{\sqrt{14}}&\frac{3}{\sqrt{14}}\\0&0&1&2\end{pmatrix}$$

$$H_{v_3}H_{v_2}H_{v_1}A=\begin{pmatrix}\sqrt{5}&\frac{4}{\sqrt{5}}&\frac{1}{\sqrt{5}}&0\\0&\sqrt{\frac{14}{5}}&\frac{16}{\sqrt{70}}&\frac{5}{\sqrt{70}}\\0&0&\sqrt{\frac{15}{7}}&\frac{20}{\sqrt{105}}\\0&0&0&-\sqrt{\frac{5}{6}}\end{pmatrix}.$$

Finalement, en posant $H_{e_4}H_{v_3}H_{v_2}H_{v_1}$ on obtient $A=QR$ avec :

$$Q:=\begin{pmatrix}\frac{2}{\sqrt{5}}&-\frac{3}{\sqrt{70}}&\frac{2}{\sqrt{105}}&-\frac{1}{\sqrt{30}}\\\frac{1}{\sqrt{5}}&\frac{6}{\sqrt{70}}&-\frac{4}{\sqrt{105}}&\frac{2}{\sqrt{30}}\\0&\frac{5}{\sqrt{70}}&\frac{6}{\sqrt{105}}&-\frac{3}{\sqrt{30}}\\0&0&\frac{7}{\sqrt{105}}&\frac{4}{\sqrt{30}}\end{pmatrix}\quad\text{et}\quad R:=\begin{pmatrix}\sqrt{5}&\frac{4}{\sqrt{5}}&\frac{1}{\sqrt{5}}&0\\0&\frac{\sqrt{14}}{\sqrt{5}}&\frac{16}{\sqrt{70}}&\frac{5}{\sqrt{70}}\\0&0&\frac{\sqrt{15}}{\sqrt{7}}&\frac{20}{\sqrt{105}}\\0&0&0&\frac{\sqrt{5}}{\sqrt{6}}\end{pmatrix}.$$

Orthogonalisation de Gram-Schmidt et décomposition QR.
Nous avons établi l'existence d'une décomposition QR, les termes diagonaux de R étant tous *positifs*, en utilisant l'algorithme de Householder. Nous allons voir que, dans le cas d'une matrice *inversible* l'algorithme de Gram-Schmidt conduit au *même résultat*. Il est toutefois important de remarquer que, si le résultat est le même, les deux façons d'y parvenir sont différentes. Cela a une conséquence pour les calculs effectifs : la méthode de Householder est numériquement plus efficace, plus « stable », que celle de Gram-Schmidt (et même que celle de Gram-Schmidt modifiée). Nous n'étudierons pas ici cette question.
Soient $A\in GL_n(\mathbb{R})$ et $(C_1,\dots,C_n)$ le système de ses vecteurs colonnes. L'espace $\mathbb{R}^n$ est muni de son produit scalaire et de sa base $\mathcal{B}:=(e_1,\dots,e_n)$ canoniques. Considérons la base $(\varepsilon_1,\dots,\varepsilon_n)$ orthonormalisée de $(C_1,\dots,C_n)$ par le procédé de Gram-Schmidt. La matrice de passage $R=\mathrm{Pass}\big((\varepsilon_1,\dots,\varepsilon_n),(C_1,\dots,C_n)\big)$ est une matrice *triangulaire supérieure* dont la diagonale est à coefficients positifs. La matrice de passage $\mathrm{Pass}\big((\mathcal{B},(\varepsilon_1,\dots,\varepsilon_n)\big)$ est *orthogonale*. La matrice initiale A est la matrice de passage $\mathrm{Pass}\big(\mathcal{B},(C_1,\dots,C_n)\big)$. La règle de produit des matrices de passage $\big(\mathcal{B}\to(\varepsilon_1,\dots,\varepsilon_n)\to(C_1,\dots,C_n)\big)$ donne $A=QR$ (*cf.* [L1], module II.2).

Exemple. Il est aisé de vérifier ce résultat si $A:=\begin{pmatrix}4&2\\3&-1\end{pmatrix}$. On a $\varepsilon_1=\frac{1}{5}(4,3)$, $\varepsilon_2=\frac{1}{5}(3,-4)$. On calcule $C_1=5\varepsilon_1$ et $C_2=\varepsilon_1+2\varepsilon_2$. D'où $Q=\frac{1}{5}\begin{pmatrix}4&3\\3&-4\end{pmatrix}$ et $R=\begin{pmatrix}5&1\\0&2\end{pmatrix}$. On vérifie que $A=QR$ et que $Q=H_v$ où $v=C_1-5e_1$

Le lecteur intéressé pourra expliciter, de façon similaire, la relation entre la décomposition QR et l'algorithme de Gram-Schmidt modifié (*cf.* l'exemple de la page 160).

1.7.3 Endomorphismes symétriques et antisymétriques d'un espace euclidien

Proposition 49. Soient E un espace vectoriel euclidien et $\mathcal{B} := (e_1, \dots, e_n)$ une base *orthonormée* de E.
(i) L'application qui à un endomorphisme $u \in \mathbf{S}(E)$ (resp. $u \in \mathbf{A}(E)$) associe sa matrice M dans la base $\mathcal{B}$ est un *isomorphisme* d'espaces vectoriels de $\mathbf{S}(E)$ (resp. $\mathbf{A}(E)$) sur $S_n(\mathbb{R})$ $\big($resp. $A_n(\mathbb{R})\big)$.
(ii) L'endomorphisme symétrique $u \in \mathbf{S}(E)$ est *positif* (resp. *défini positif*) si, et seulement si, $M \in S_n^+(\mathbb{R})$ $\big($resp. $M \in S_n^{++}(\mathbb{R})\big)$.

Démonstration. Montrons (i). Si $M = \mathrm{Mat}_{\mathcal{B}}(u)$, on sait que ${}^tM = \mathrm{Mat}_{\mathcal{B}}(u^*)$. Ainsi $u \in \mathbf{S}(E)$ si, et seulement si, ${}^tM = M$ et $u \in \mathbf{A}(E)$ si, et seulement si, ${}^tM = -M$.
Par ailleurs, pour tout $x \in E$, $(\,u(x) \mid x\,) = {}^tXMX$, d'où l'assertion (ii). ∎

Proposition 50. Soit E un espace vectoriel *euclidien*. Pour tout $u \in \mathcal{L}(E)$, les endomorphismes auto-adjoints positifs u^*u et uu^* ont même rang : celui de u.

Démonstration. On a vu (*cf.* proposition 41 de la page 171) que les endomorphismes u^*u et uu^* sont positifs. Soient $\mathcal{B}$ une base orthonormée de E et $M = \mathrm{Mat}_{\mathcal{B}}(u)$, on a ${}^tM = \mathrm{Mat}_{\mathcal{B}}(u^*)$, d'où $\mathrm{Mat}_{\mathcal{B}}(u^*u) = {}^tMM$ et $\mathrm{Mat}_{\mathcal{B}}(uu^*) = M{}^tM$. D'après la proposition 41 de la page 171, on a ${}^tMM \in S_n^+(\mathbb{R})$ et $\operatorname{rang}{}^tMM = \operatorname{rang} M = \operatorname{rang} M{}^tM$. ∎

Proposition 51. Soit E un espace vectoriel *euclidien*. L'application qui à tout *endomorphisme* $u \in \mathbf{S}(E)$ associe la *forme bilinéaire* symétrique ψ définie par $\psi(x, y) = \big(\,u(x) \,\big|\, y\,\big)$ est un *isomorphisme* d'espaces vectoriels de $\mathbf{S}(E)$ sur $\mathcal{S}(E)$.
Les *matrices* de l'endomorphisme u et de la forme bilinéaire ψ correspondante dans une *base orthonormée* sont les *mêmes*. Cette matrice commune reste *semblable* à elle même quand on change de base orthonormée.
De plus l'endomorphisme u est *positif* (resp. *défini positif*) si, et seulement si, la forme bilinéaire symétrique ψ est *positive* (resp. *définie positive*).

Le lecteur retiendra en particulier que *les matrices d'une forme quadratique donnée sur un espace vectoriel euclidien dans les bases orthonormées de cet espace sont toutes semblables*. En particulier, elles ont toutes le *même spectre*.

Démonstration. On vérifie sans difficulté que l'application $\theta : u \mapsto \psi$ de $\mathbf{S}(E)$ dans $\mathcal{S}(E)$ (*cf.* la proposition 40 de la page 171) est linéaire. On a $\psi = 0$ si, et seulement si, pour tout $(x, y) \in E^2$, $\big(\,u(x) \,\big|\, y\,\big) = 0$; s'il en est ainsi, $u(x) = 0$ et par suite $u = 0$: l'application θ est *injective*. Par ailleurs, $\dim \mathbf{S}(E) = \dim S_n(\mathbb{R}) = \dim \mathcal{S}(E)$, l'application linéaire injective θ est donc un isomorphisme de $\mathbf{S}(E)$ sur $\mathcal{S}(E)$.

Choisissons une *base orthonormée* $\mathcal{B} := (e_1, \dots, e_n)$ de E, soient $u \in \mathbf{S}(E)$ et $M := \mathrm{Mat}_{\mathcal{B}}(u) \in S_n(\mathbb{R})$. Soit $\psi := \theta(u)$. Pour tous $x, y \in E$,

$$\psi(x, y) = (\, u(x) \mid y \,) = {}^tX {}^tMY = {}^tXMY,$$

d'où $M = \mathrm{Mat}(\psi; \mathcal{B})$. Dans un changement de *base orthonormée*, la matrice M est transformée en la matrice *semblable* $P^{-1}MP$ en tant que matrice d'un *endomorphisme* ; elle est transformée en la matrice *congruente* ${}^tPMP$ en tant que matrice d'une *application bilinéaire*. Il n'y a pas de contradiction, car la matrice de changement de base P est *orthogonale* : $P^{-1} = {}^tP$.

Pour tout $x \in E$, $\psi(x, x) = (\, u(x) \mid x \,)$, donc u est positif (resp. *défini positif*) si, et seulement si, ψ l'est. ∎

Exercice 12.

On suppose maintenant que $\mathcal{B} := (\varepsilon_1, \dots, \varepsilon_n)$ est une base quelconque (non orthonormée) de E. Soit $M := \mathrm{Mat}_{\mathcal{B}}(u)$. Quelle est la matrice associée à $\psi := \theta(u)$ dans la base $\mathcal{B}$?

Solution. Soit $G := \mathrm{Mat}((\,|\,), \mathcal{B})$. Pour tout $(x, y) \in E^2$, $\psi(x, y) = (\, u(x) \mid y \,) = {}^tX {}^tMGY$. Donc $\mathrm{Mat}(\psi; \mathcal{B}) = {}^tMG$. Remarquons que :

$$\psi(x, y) = \psi(y, x) = {}^tX {}^tMGY = {}^tY {}^tMGX = {}^tX {}^tGMX,$$

donc ${}^tMG = {}^tGM$.

Exercice 13.

Cet exercice porte sur les théorèmes d'Apollonius (ils sont dus à Apollonius de Perge, on les trouve dans le septième livre de son *Traité des Coniques*).

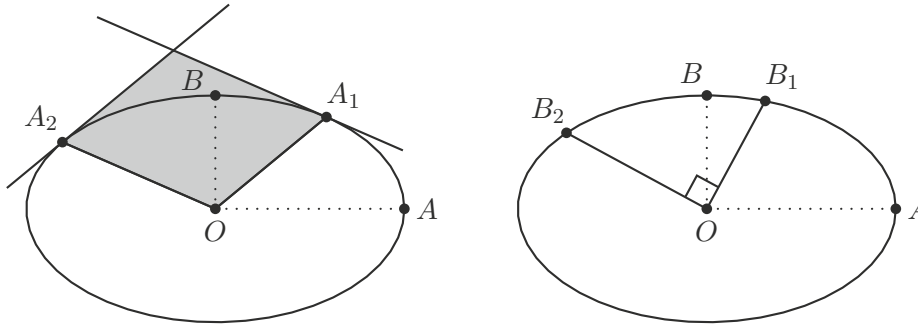


FIGURE I.4.3 – Théorèmes d'Apollonius.

On munit $\mathbb{R}^2$ de sa structure canonique d'espace affine euclidien, et l'on désigne par $\mathcal{R} = (O, \vec{\imath}, \vec{\jmath})$ le repère canonique d'origine $O = (0, 0)$. On considère *l'ellipse* $\mathcal{E}$ dont une équation dans le repère $\mathcal{R}$ est $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$ $(a, b > 0)$ et la forme quadratique $\Phi(x, y) = \dfrac{x^2}{a^2} + \dfrac{y^2}{b^2}$.

Soient $\mathcal{D}_1$ et $\mathcal{D}_2$ deux droites passant par l'origine, elles coupent l'ellipse $\mathcal{E}$ respectivement en A_1, A_1' et A_2, A_2'. On dit que la droite D_2 est *conjuguée* de la droite D_1 par rapport à $\mathcal{E}$

si D_2 est parallèle aux tangentes à l'ellipse en A_1 et A_1' (qui sont évidemment parallèles entre elles, par symétrie).
(i) Montrer que, si D_2 est conjuguée de D_1, $\mathcal{D}_1$ contient les milieux des segments découpés par l'ellipse sur les droites parallèles à $\mathcal{D}_2$.
(ii) Montrer que $\mathcal{D}_2$ est conjuguée de $\mathcal{D}_1$ si, et seulement si, $(\overrightarrow{OA_1}, \overrightarrow{OA_2})$ est une base Φ-conjuguée de $\mathbb{R}^2$. En déduire qu'alors D_1 est conjuguée de D_2. On dit, dans ces conditions, que $\overrightarrow{OA_1}$ et $\overrightarrow{OA_2}$ sont deux *demi-diamètres conjugués* de l'ellipse (*cf.* figure I.4.3 de la page précédente).
(iii) Montrer que si $\overrightarrow{OA_1}$ et $\overrightarrow{OA_2}$ sont deux *demi-diamètres conjugués* de $\mathcal{E}$, alors :
- $OA_1^2 + OA_2^2 = a^2 + b^2$,
- l'aire du parallélogramme construit sur $\overrightarrow{OA_1}$ et $\overrightarrow{OA_2}$ est égale à ab

(premier et troisième théorème d'Apollonius).
(iv) Soient OB_1 et OB_2 deux demi-diamètres *orthogonaux* de $\mathcal{E}$.
Montrer que $\dfrac{1}{OB_1^2} + \dfrac{1}{OB_2^2} = \dfrac{1}{a^2} + \dfrac{1}{b^2}$ (deuxième théorème d'Apollonius).

Solution. La question (i) peut être résolue en faisant une transformation affine (les propriétés considérées : tangentes, milieux…, sont affines). On utilise $(x, y) \mapsto (x, \frac{a}{b}y)$ qui transforme l'ellipse en le cercle de centre O et de rayon a. Nous laissons les détails au lecteur.
Montrons (ii). Soient $\overrightarrow{OA_1} := x_1\vec{\imath} + y_1\vec{\jmath}$ et $\overrightarrow{OA_2} = x_2\vec{\imath} + y_2\vec{\jmath}$ deux vecteurs, avec $A_1, A_2 \in \mathcal{E}$. Une équation de la tangente à l'ellipse en A_1 est $\dfrac{x_1}{a^2}(x-x_1)+\dfrac{y_1}{b^2}(y-y_1) = 0$, la direction de $\overrightarrow{OA_2}$ est donc conjuguée de celle de $\overrightarrow{OA_2}$ par rapport à $\mathcal{E}$ si, et seulement si, $\dfrac{x_1x_2}{a^2} + \dfrac{y_1y_2}{b^2} = 0$, c'est-à-dire si, et seulement si, les deux vecteurs $\overrightarrow{OA_1}$ et $\overrightarrow{OA_2}$ sont conjugués par rapport à la forme quadratique Φ. Cette dernière condition est symétrique, il en est donc de même pour la relation de conjugaison par rapport à $\mathcal{E}$.
Montrons (iii). Considérons $(\mathbb{R}^2, \Phi)$, c'est un espace vectoriel euclidien (Φ est définie positive). Si $\overrightarrow{OA_1}$ et $\overrightarrow{OA_2}$ sont deux vecteurs *conjugués*, ils forment une base orthonormée de cet espace euclidien. La direction conjuguée de l'axe des abscisses est l'axe des ordonnées, on a alors $A_1 = A$, $A_2 = B$, $\overrightarrow{OA} = a\vec{\imath}$ et $\overrightarrow{OB} = b\vec{\jmath}$.
Considérons la forme bilinéaire symétrique ψ définie sur cet espace par le produit scalaire canonique $(\,|\,)$ de $\mathbb{R}^2$: $\psi\big((x_1, y_1), (x_2, y_2)\big) = \big(\,(x_1, y_1) \;\big|\; (x_2, y_2)\,\big)$.
La matrice $M = \mathrm{Mat}(\psi, (\overrightarrow{OA_1}, \overrightarrow{OA_2})$ de cette forme dans une base orthonormée (*i.e.* Φ-orthonormée) $(\overrightarrow{OA_1}, \overrightarrow{OA_2})$ peut être interprétée comme la matrice d'un endomorphisme (*cf.* proposition 51 de la page 177) et reste donc *semblable* à elle-même quand cette base varie en restant *orthonormée*. Ainsi les deux matrices $M := \mathrm{Mat}(\psi, (\overrightarrow{OA_1}, \overrightarrow{OA_2}))$ et $M_0 := \mathrm{Mat}(\psi, (\overrightarrow{OA}, \overrightarrow{OB})$ sont semblables.
On en déduit $\operatorname{Tr} M = \operatorname{Tr} M_0$ et $\det M = \det M_0$.
La matrice M est la matrice de Gram de $(\overrightarrow{OA_1} = (x_1, y_1), \overrightarrow{OA_2} = (x_2, y_2))$; on a :

$M = \begin{pmatrix} \overrightarrow{OA_1}^2 & \overrightarrow{OA_1}.\overrightarrow{OA_2} \\ \overrightarrow{OA_1}.\overrightarrow{OA_2} & \overrightarrow{OA_2}^2 \end{pmatrix}$ et, en particulier, $M_0 = \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix}$.

On en déduit $\operatorname{Tr} M = \operatorname{Tr} M_0 = OA_1^2 + OA_2^2 = a^2 + b^2$ et :

$$\det M = \det M_0 = \operatorname{Gram}(M) = \begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix}^2 = a^2 b^2,$$

ce qui termine la démonstration (la valeur absolue de $\begin{vmatrix} x_1 & x_2 \\ y_1 & y_2 \end{vmatrix}$ est l'aire du parallélogramme construit sur les diamètres conjugués).

Montrons (iv). Soit $\theta \in \mathbb{R}$ tel que :

$$\overrightarrow{OB_1} = OB_1(\cos\theta\vec{\imath} + \sin\theta\vec{\jmath}) = OB_1\vec{I} \quad \text{et} \quad \overrightarrow{OB_2} = OB_2(-\sin\theta\vec{\imath} + \cos\theta\vec{\jmath}) = OB_2\vec{J}.$$

On en déduit : $OB_1^2\left(\dfrac{\cos^2\theta}{a^2} + \dfrac{\sin^2\theta}{b^2}\right) = 1$, $\Phi(\vec{I}) = \dfrac{1}{OB_1^2}$ et, de même $\Phi(\vec{J}) = \dfrac{1}{OB_2^2}$.

Les matrices associées à Φ dans les bases orthonormées $(\vec{\imath}, \vec{\jmath})$ et $(\vec{I}, \vec{J})$ sont *semblables*, donc leurs traces sont égales : $\Phi(\vec{\imath}) + \Phi(\vec{\jmath}) = \Phi(\vec{I}) + \Phi(\vec{J})$. On a donc :

$$\frac{1}{a^2} + \frac{1}{b^2} = \frac{1}{OB_1^2} + \frac{1}{OB_2^2}.$$

Le lecteur vérifiera que les trois théorèmes d'Apollonius s'étendent facilement au cas des quadriques (dans l'espace $\mathbb{R}^3$, *cf.* le module « Surfaces » de [L2-L3]) et plus généralement en dimension finie quelconque.

Actions de groupes sur des espaces vectoriels de matrices.

Nous utilisons ici les notions introduites dans le module I.2. Il y a deux actions « intéressantes » du groupe linéaire $GL_n(\mathbb{R})$ sur l'espace vectoriel des matrices carrées $M_n(\mathbb{R})$: pour tout $(P, M) \in GL_n(\mathbb{R}) \times M_n(\mathbb{R})$, $(P, M) \mapsto P^{-1}MP$ et $(P, M) \mapsto {}^tPMP$. On sait que les *orbites* d'une action de groupe sont des *classes d'équivalence* pour une relation d'équivalence convenable. Dans le premier cas cette relation est la *similitude* des matrices, dans le second cas c'est la *congruence*. Ces deux relations sont distinctes. Considérons le sous-espace $S_n(\mathbb{R}) \subset M_n(\mathbb{R})$ des matrices *symétriques* ; les deux actions se comportent de façon très différente vis à vis de ce sous-espace : il est invariant pour la seconde, l'orbite d'une matrice symétrique est formée de matrices symétriques, donc « reste dans $S_n(\mathbb{R})$ », pour la première, par contre, l'orbite d'une matrice symétrique contient en général des matrices non symétriques, elle « sort de $S_n(\mathbb{R})$ ». Le problème disparaît si l'on remplace les actions du groupe $GL_n(\mathbb{R})$, par celles (induites) de son sous-groupe $O_n(\mathbb{R})$: $(P, M) \mapsto P^{-1}MP$ et $(P, M) \mapsto {}^tPMP$ pour $(P, M) \in O_n(\mathbb{R}) \times M_n(\mathbb{R})$, en effet maintenant les deux actions *coïncident*, puisque $P^{-1} = {}^tP$. En particulier $S_n(\mathbb{R})$ est invariant par cette action commune. Ceci permet un « double regard » sur les matrices symétriques : matrices d'applications ou matrices de formes bilinéaires, tant que l'on travaille *exclusivement avec des bases orthonormées*.

1.7.4 Diagonalisation d'un endomorphisme symétrique d'un espace euclidien

Théorème 52 (Théorème spectral pour les endomorphismes symétriques d'un espace euclidien). Soit E un espace vectoriel euclidien de dimension $n > 0$. Tout endomorphisme *symétrique* de E est *diagonalisable* et ses sous-espaces propres sont deux à deux *orthogonaux* :

$$E = \overset{\perp}{\bigoplus_{\lambda \in \operatorname{sp} u}} E_\lambda(u).$$

Conséquence. Il existe une base orthonormée $\mathcal{B}$ formée de vecteurs propres de u, donc $\operatorname{Mat}_{\mathcal{B}}(u) = \operatorname{Diag}(\lambda_1, \dots, \lambda_n)$, les valeurs propres étant répétées avec leurs multiplicités. Notons que, si les valeurs propres sont *toutes simples*, la base $\mathcal{B} = (e_1, \dots, e_n)$ est déterminée de manière unique, à l'ordre et au « signe » des e_i près. On a $(\, u(x) \mid y \,) = \sum_{i=1}^{n} \lambda_i x_i y_i$ et $\|u(x)\|^2 = \sum_{i=1}^{n} \lambda_i^2 x_i^2$.

Démonstration. Nous montrerons d'abord que l'endomorphisme u est *scindé*, puis que ses sous-espaces propres sont deux à deux *orthogonaux*, ce qui nous permettra de conclure.
a) Pour prouver que u est scindé, nous choisissons une base orthonormée de E et nous considérons la matrice *symétrique* $A = \operatorname{Mat}_{\mathcal{B}}(u)$. Montrons que le polynôme caractéristique de A est scindé (sur $\mathbb{R}$). Nous utiliserons pour cela les remarques suivantes :

- Si $M \in M_n(\mathbb{C}) := (m_{ij})$, on note $\overline{M} = (\overline{m_{ij}})$ la matrice de $M_n(\mathbb{C})$ dont les coefficients sont les *conjugués* de ceux de M. La conjugaison est un automorphisme de corps de $\mathbb{C}$, on en déduit que $\overline{MN} = \overline{M}\,\overline{N}$ et que $({}^t\overline{M}) = {}^t\overline{M}$.
- Soit $x \in \mathbb{C}^n$. Si X est la matrice unicolonne correspondante (des coordonnées de x dans la base canonique de $\mathbb{C}^n$), on a ${}^tX\overline{X} = \sum_{h=1}^{n} |x_h|^2 = 0$ si, et seulement si, $x = 0$.

Considérons la matrice réelle A comme une matrice de $M_n(\mathbb{C})$. Soient $\lambda \in \operatorname{sp}_{\mathbb{C}} A$ et $x \in \mathbb{C}^n \setminus \{0\}$ tels que $AX = \lambda X$. On a ${}^t\overline{X}AX = \lambda {}^t\overline{X}X$. La matrice A étant réelle, on en déduit $A = \overline{A}$, puis $A\overline{X} = \overline{\lambda}\overline{X}$ et $\overline{{}^t\overline{X}AX} = {}^tXA\overline{X} = \overline{\lambda}{}^tX\overline{X}$. La matrice A étant symétrique, on a par ailleurs ${}^tXA\overline{X} = {}^tX{}^tA\overline{X} = {}^t(AX)\overline{X} = \lambda {}^tX\overline{X}$. Ainsi ${}^tXA\overline{X} = \overline{\lambda}{}^tX\overline{X} = \lambda {}^tX\overline{X}$ et $\lambda = \overline{\lambda}$, puisque ${}^tX\overline{X} \neq 0$, les valeurs propres de A sont toutes *réelles*.
b) Soit (λ, μ) un couple de valeurs propres *distinctes* de u. Montrons que les sous-espaces propres associés $E_\lambda(u)$ et $E_\mu(u)$ sont orthogonaux (si u n'a qu'une seule valeur propre le problème ne se pose pas). Soient $x \in E_\lambda(u)$ et $y \in E_\mu(u)$, on a $u(x) = \lambda x$ et $u(y) = \mu y$, de $(\, u(x) \mid y \,) = (\, x \mid u(y) \,)$ l'on déduit $\lambda(\, x \mid y \,) = \mu(\, x \mid y \,)$, puis $(\, x \mid y \,) = 0$.
c) Désignons par $F = \overset{\perp}{\bigoplus_{\lambda \in \operatorname{sp} u}} E_\lambda(u)$ la somme des sous-espaces propres de u. Le sous-espace $F \subset E$ est évidemment u-stable ; ou bien $F^\perp = \{0\}$ et la propriété est démontrée ($F = E$), ou bien $F^\perp \neq \{0\}$, alors, F étant u stable, $F^\perp$ est aussi u-stable (*cf.* proposition 26 de la page 164), la restriction $u_{|F^\perp}$ est un endomorphisme *symétrique* et par

suite admet un sous-espace propre de dimension au moins 1 (d'après a)), qui est aussi un sous-espace propre de u, ce qui est contradictoire. Finalement $F^\perp = \{0\}$ et u est diagonalisable.

Si $A \in S_n(\mathbb{R})$, en appliquant le théorème ci-dessus à l'endomorphisme u de $E = \mathbb{R}^n$ dont A est la matrice dans la base canonique, on obtient le théorème suivant :

Théorème 53. Soit $A \in S_n(\mathbb{R})$. Il existe un couple (D, P) formé d'une matrice *diagonale* D et d'une matrice *orthogonale* P tel que $A = PDP^{-1} = PD{}^tP$.

Le lecteur prendra garde au fait que ce résultat n'est pas toujours vrai pour une matrice symétrique $A \in S_n(K)$ à coefficients dans un corps quelconque K. Soient par exemple $K = \mathbb{C}$ et $A = \begin{pmatrix} 1 & j & j^2 \\ j & j^2 & 1 \\ j^2 & 1 & j \end{pmatrix}$. On a $A^2 = 0$ (A est nilpotente), donc si elle était diagonalisable, elle serait nulle.

Exercice 14.

Soit $A := \begin{pmatrix} 1 & -2 & 3 \\ -2 & 1 & -3 \\ 3 & -3 & 6 \end{pmatrix}$. Déterminer les valeurs propres et les sous-espaces propres de A. En déduire un couple de matrices $(D, P) \in \mathcal{D}_n(\mathbb{R}) \times O_n(\mathbb{R})$ tel que $A = PDP^{-1}$.

Solution. Les valeurs propres de A sont -1, 0 et 9. On a $E_0(A) = \mathbb{R}v_1$, $E_{-1}(A) = \mathbb{R}v_2$ et $E_9(A) = \mathbb{R}v_3$, avec $v_1 = (-1, 1, 1)$, $v_2 = (1, 1, 0)$ et $v_3 = (1, -1, 2)$. On vérifie que $(v_1 \mid v_2) = 0$, $(v_1 \mid v_3) = 0$, $(v_2 \mid v_3) = 0$ (la matrice A est symétrique). On peut donc prendre $D = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 9 \end{pmatrix}$ et $P = \begin{pmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{pmatrix}$.

1.7.5 Applications

Signature d'une forme quadratique réelle

Proposition 54. Soit E un espace vectoriel réel de dimension $n > 0$. Soient Φ une forme quadratique sur E, de signature (p, q), $\mathcal{B}$ une base de E et $A := \mathrm{Mat}(\Phi; \mathcal{B})$. Alors p est égal au nombre de racines strictement positives du polynôme caractéristique de A et q est égal au nombre de racines strictement négatives de ce polynôme.

Démonstration.

On sait qu'il existe $(D, P) \in D_n(\mathbb{R}) \times O_n(\mathbb{R})$ tel que $D = {}^tPAP = P^{-1}AP$.

Soit $\mathcal{B}' = (e'_1, \ldots, e'_n)$ la base de E telle que $\mathrm{Pass}(\mathcal{B}, \mathcal{B}') = P$; alors $D = \mathrm{Mat}(\Phi; \mathcal{B}')$ et $\mathcal{B}'$ est donc une base Φ-conjuguée de E. Par ailleurs les matrices A et D sont *semblables*.

Exercice 15.

Soit l'espace vectoriel $\mathbb{R}^n$ muni de sa base canonique. Soit $\alpha \in \mathbb{R}$. Quelle est la signature de la forme quadratique qui à $x := (x_1, \dots, x_n) \in \mathbb{R}^n$ associe $Q(x) := \sum_{i=1}^{n} x_i^2 + \alpha(\sum_{i=1}^{n} x_i)^2$?

Solution. La matrice M associée à Q est $M = I_n + \alpha A$, A étant la matrice de $M_n(\mathbb{R})$ dont tous les coefficients sont égaux à 1. La matrice A est de rang 1 ($\mathrm{Ker}\, A$ est l'hyperplan de $\mathbb{R}^n$ d'équation $\sum_{i=1}^{n} x_i = 0$) et sa trace est égale à n, ainsi 0 est valeur propre d'ordre $n-1$ et n est valeur propre d'ordre 1. On en déduit que 1 est valeur propre d'ordre $n-1$ de M et que $1+\alpha n$ est valeur propre d'ordre 1. La signature de Q est donc : $(n,0)$ si $1+\alpha n > 0$, $(n-1,0)$ si $1+\alpha n = 0$ et $(n-1,1)$ si $1+\alpha n < 0$.

Problèmes d'extremum

Proposition 55. Soient E un espace vectoriel euclidien de dimension strictement positive et u un endomorphisme symétrique de E, on a :

$$\sup_{\|x\|=1} \big(\, u(x) \,\big|\, x \,\big) = \sup_{x \in E\setminus\{0\}} \frac{\big(\, u(x) \,\big|\, x \,\big)}{\|x\|^2} = \sup\big(\mathrm{sp}(u)\big),$$

$$\inf_{\|x\|=1} \big(\, u(x) \,\big|\, x \,\big) = \inf_{x \in E\setminus\{0\}} \frac{\big(\, u(x) \,\big|\, x \,\big)}{\|x\|^2} = \inf\big(\mathrm{sp}(u)\big).$$

De plus $u \in \mathbf{S}^+(E)$ $\big($resp. $u \in \mathbf{S}^{++}(E)\big)$ si, et seulement si, toutes les valeurs propres de u sont positives (resp. strictement positives).

Démonstration. Soit $\mathcal{B} := (e_1, \dots, e_n)$ une base orthonormée de E formée de vecteurs propres de u. Pour tout $i \in [\![1, n]\!]$, $u(e_i) = \lambda_i e_i$ et, quitte à permuter les vecteurs de base, on peut supposer que $\lambda_1 \leqslant \lambda_2 \leqslant \dots \leqslant \lambda_n$. Pour tout $x := \sum_{i=1}^{n} x_i e_i \in E$, $\big(\, u(x) \,\big|\, x \,\big) = \sum_{i=1}^{n} \lambda_i x_i^2$, donc $\lambda_1 \sum_{i=1}^{n} x_i^2 \leqslant \big(\, u(x) \,\big|\, x \,\big) \leqslant \lambda_n \sum_{i=1}^{n} x_i^2$. Si $x \neq 0$, $\lambda_1 \leqslant \frac{\big(\, u(x) \,\big|\, x \,\big)}{\|x\|^2} \leqslant \lambda_n$. De plus $\big(\, u(e_1) \,\big|\, e_1 \,\big) = \lambda_1$ et $\big(\, u(e_n) \,\big|\, e_n \,\big) = \lambda_n$. Pour tout $x \neq 0$, $\dfrac{x}{\|x\|}$ est de norme 1, le résultat s'en déduit facilement. ∎

Les résultats suivants se déduisent aisément de la proposition 55.

Corollaire 56. Soit $A \in S_n(\mathbb{R})$.

(i) On a $\sup\limits_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \dfrac{{}^tXAX}{{}^tXX} = \sup\big(\mathrm{sp}(A)\big)$ et $\inf\limits_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \dfrac{{}^tXAX}{{}^tXX} = \inf\big(\mathrm{sp}(A)\big)$.

(ii) La matrice A est positive (resp. définie positive) si, et seulement si, toutes ses valeurs propres sont positives (resp. strictement positives).

Corollaire 57. Soient Φ une forme quadratique sur l'espace vectoriel euclidien E, $\mathcal{B}$ une base orthonormée de E et $A := \mathrm{Mat}(\Phi;\mathcal{B})$.
On a $\sup\limits_{\|x\|=1} \Phi(x) = \sup\big(\mathrm{sp}(A)\big)$ et $\inf\limits_{\|x\|=1} \Phi(x) = \inf\big(\mathrm{sp}(A)\big)$.

Quotient de Rayleigh. Soit $A \in S_n(\mathbb{R})$, on appelle *quotient de Rayleigh* de A l'application $R_A : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ définie par $R_A(x) = \dfrac{{}^tXAX}{{}^tXX}$.

Diagonalisation simultanée de deux formes quadratiques

Le résultat suivant a de très nombreuses applications : problèmes d'extremum (*cf.* exercice 16 ci-dessous), axes des coniques et des quadriques, étude de la courbure des surfaces (voir [MPA1]), « petits mouvements » et problèmes de stabilité en mécanique…

Théorème 58. Soit E un espace vectoriel réel de dimension $n > 0$.
Soient Φ, Ψ, deux formes quadratiques sur E, Φ étant *définie positive*.
Il existe une base $\mathcal{B} = (e_1, \ldots, e_n)$ de E à la fois Φ-*orthonormée* et Ψ-*conjuguée*.

Démonstration. La forme Φ est *définie positive*, elle permet donc de munir E d'une *structure d'espace euclidien*, considérons l'espace vectoriel euclidien (E, Φ). Soient ψ la forme polaire de Ψ et $u \in \mathbf{S}(E)$ tel que, pour tous $x, y \in E$: $\psi(x,y) = \big(\, u(x) \;\big|\; y \,\big)$ (*cf.* proposition 51 de la page 177), on sait qu'il existe une base orthonormée $\mathcal{B} = (e_1, \ldots, e_n)$ de E formée de *vecteurs propres* de u (*cf.* théorème 52 de la page 181). Pour tous :

$$x := \sum_{i=1}^{n} x_i\, e_i, \quad y := \sum_{i=1}^{n} y_i\, e_i \;\in E,$$

on a $u(x) = \sum\limits_{i=1}^{n} \lambda_i\, x_i\, e_i$ et $\psi(x,y) = \big(\, u(x) \;\big|\; y \,\big) = \sum\limits_{i=1}^{n} \lambda_i\, x_i\, y_i$, la base $\mathcal{B}$ est donc Ψ-conjuguée. ∎

Corollaire 59. Soient $A \in S_n^{++}(\mathbb{R})$ et $B \in S_n(\mathbb{R})$. Il existe un couple de matrices $(D, P) \in \mathcal{D}_n(\mathbb{R}) \times GL_n(\mathbb{R})$ tel que $A = {}^tPP$ et $B = {}^tPDP$.

Démonstration. Il suffit d'appliquer le théorème aux formes quadratiques Φ et Ψ respectivement définies par les matrices A et B dans la base canonique $\mathcal{B}_0$ de $E = \mathbb{R}^n$. On obtient une base $\mathcal{B}$ qui est Φ-orthonormée et Ψ-conjuguée : $\mathrm{Mat}(\Phi, \mathcal{B}) = I$ et $\mathrm{Mat}(\Psi, \mathcal{B}) = D$. Posant $P := \mathrm{Pass}(\mathcal{B}, \mathcal{B}_0)$, on obtient $A = {}^tPP$ et $B = {}^tPDP$. ∎

Remarque. On a $A^{-1}B = P^{-1}({}^tP)^{-1}{}^tPDP = P^{-1}DP$.
La matrice $A^{-1}B$ est donc diagonalisable. On a ainsi montré que, pour tout couple $(A, B) \in S_n^{++}(\mathbb{R}) \times S_n(\mathbb{R})$, la matrice $A^{-1}B$ est diagonalisable.
Par conséquent, si E est un espace vectoriel euclidien, pour tout couple $(u, v) \in \mathbf{S}^{++}(E) \times \mathbf{S}(E)$, l'endomorphisme $u^{-1}v$ est diagonalisable.

Remarque pratique. Les matrices A et B étant données, comment déterminer pratiquement les matrices D et P ? (Dans les applications l'objet important est le plus souvent D). On considère pour cela la matrice $A^{-1}B$: nous venons de voir qu'elle est diagonalisable et que P^{-1} est *une* matrice de passage de la base canonique de $\mathbb{R}^n$ à une base propre de $A^{-1}B$; on a $\lambda \in \operatorname{sp} A^{-1}B$ si et seulement s'il existe $v \in \mathbb{R}^n \setminus \{0\}$ tel que $A^{-1}BV = \lambda V$, ou de façon équivalente $(B - \lambda A)V = 0$, les valeurs propres de $A^{-1}B$ (i.e. les termes diagonaux de D) sont donc les zéros du polynôme $\det(B - \lambda A)$. Ceci permet de calculer D sans calculer A^{-1}. En fait une base *propre* de $A^{-1}B$ peut être choisie arbitrairement, pourvu qu'elle soit Φ*-orthonormée* : si Q est la matrice de passage d'une telle base à la base canonique de $\mathbb{R}^n$, on a $A = {}^tQQ$ et $A^{-1}B = Q^{-1}DQ$ (en ordonnant convenablement la base), d'où $B = AQ^{-1}DQ = {}^tQDQ$. On obtient ainsi une solution au problème de diagonalisation simultanée. Notons que si les zéros λ_i ($i \in [\![1, n]\!]$) du polynôme $\det(B - \lambda A)$ sont *simples*, la résolution de notre problème est facilitée : on détermine un système v_i ($i \in [\![1, n]\!]$) de vecteurs solutions respectives des équations $(B - \lambda_i A)V = 0$ et on le Φ-unitarise. Les v_i étant uniques à une homothétie près, la base ainsi obtenue est *automatiquement* Φ-orthonormée.

Exercice 16.

Soient Φ et Ψ les deux formes quadratiques définies sur $\mathbb{R}^3$ par :

$$\Phi(x_1, x_2, x_3) := 5x_1^2 + 6x_2^2 + 3x_3^2 + 6x_1x_2 - 2x_1x_3 + 4x_2x_3$$

$$\Psi(x_1, x_2, x_3) := 3x_1^2 + 8x_2^2 + 2x_3^2 + 2x_1x_2 - 6x_1x_3 + 4x_2x_3.$$

Déterminer une base de $\mathbb{R}^3$ Φ-orthonormée.

Calculer les extrema de la fonction $\dfrac{\Psi(x_1, x_2, x_3)}{\Phi(x_1, x_2, x_3)}$ lorsque $(x_1, x_2, x_3) \neq (0, 0, 0)$.

Solution. On montre d'abord que la forme Φ est définie positive, par exemple en utilisant la méthode de Gauß :

$$\Phi(x_1, x_2, x_3) = 5\left(x_1 + \frac{1}{5}(3x_2 - x_3)\right)^2 + \frac{21}{5}\left(x_2 + \frac{13}{21}x_3\right)^2 + \frac{25}{21}x_3^2.$$

Soient $A := \begin{pmatrix} 5 & 3 & -1 \\ 3 & 6 & 2 \\ -1 & 2 & 3 \end{pmatrix}$ et $B := \begin{pmatrix} 3 & 1 & -3 \\ 1 & 8 & 2 \\ -3 & 2 & 2 \end{pmatrix}$. On vérifie facilement que :

(i) pour tout $\lambda \in \mathbb{R}$: $\det(B - \lambda A) = -24(\lambda - 1)(\lambda + 1)(\lambda - 2)$;

(ii) une base Φ-orthonormée formée de vecteurs propres de $A^{-1}B$ est :

$$v_1 = \frac{1}{5}(1, 1, -2), v_2 = \frac{1}{5}(3, -2, 4), v_3 = \frac{1}{5}(2, -3, 1).$$

On a donc $P^{-1} = \frac{1}{5}\begin{pmatrix} 1 & 3 & 2 \\ 1 & -2 & -3 \\ -2 & 4 & 1 \end{pmatrix}$, $P = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & 1 \\ 0 & -2 & -1 \end{pmatrix}$, $D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$

(on remarque que, les trois valeurs propres étant distinctes, la base (v_1, v_2, v_3) est unique au changement de chaque v_i par $-v_i$ près). Par suite :

$$\inf_{(x_1,x_2,x_3)\neq(0,0,0)} \frac{\Psi(x_1, x_2, x_3)}{\Phi(x_1, x_2, x_3)} = -1 \quad \text{et} \quad \sup_{(x_1,x_2,x_3)\neq(0,0,0)} \frac{\Psi(x_1, x_2, x_3)}{\Phi(x_1, x_2, x_3)} = 2.$$

Racine carrée d'un endomorphisme symétrique positif d'un espace euclidien

Proposition 60. Soient E un espace vectoriel euclidien et $u \in \mathbf{S}^+(E)$. Il existe un unique endomorphisme $v \in \mathbf{S}^+(E)$ tel que $v^2 = u$, v est dit *racine* carrée de u.

Démonstration. Montrons que v *existe*. Soit $\mathcal{B} = (e_1, \dots, e_n)$ une base orthonormée de vecteurs propres de u. Pour tout $i \in [\![1, n]\!]$, il existe $\lambda_i \in \mathbb{R}^+$ tel que $u(e_i) = \lambda_i e_i$ et on a $\mathrm{Mat}_{\mathcal{B}}(u) = \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$. Soit $v \in \mathcal{L}(E)$ l'endomorphisme défini par $\mathrm{Mat}_{\mathcal{B}}(v) = \mathrm{Diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$. On vérifie aisément que $v \in \mathbf{S}^+(E)$ et que $v^2 = u$.
Montrons que v est *unique*. Cela résulte immédiatement du lemme suivant. ∎

Lemme 61. Soient $v, w \in \mathbf{S}^+(E)$ tels que $v^2 = w^2$, alors $v = w$.

Démonstration. Soit $\mu \in \mathrm{sp}\, v^2$, on a $\mu \geqslant 0$. Si $\mu = 0$, v étant *diagonalisable*, on a $\mathrm{Ker}\, v = \mathrm{Ker}\, v^2$. Si $\mu > 0$, on a $\mathrm{Ker}(v^2 - \mu\ \mathrm{Id}) = \mathrm{Ker}(v - \sqrt{\mu}\ \mathrm{Id}) \oplus \mathrm{Ker}(v + \sqrt{\mu}\ \mathrm{Id})$ (« lemme des noyaux »), mais $\mathrm{Ker}(v + \sqrt{\mu}\ \mathrm{Id}) = \{0\}$ puisque $\mathrm{sp}\, v \subset \mathbb{R}^+$, donc :

$$\mathrm{Ker}(v^2 - \mu\ \mathrm{Id}) = \mathrm{Ker}(v - \sqrt{\mu}\ \mathrm{Id}).$$

Si $v^2 = w^2$, on a, pour tout $\mu \in \mathrm{sp}\, v^2$,

$$\mathrm{Ker}(v^2 - \mu\ \mathrm{Id}) = \mathrm{Ker}(v - \sqrt{\mu}\ \mathrm{Id}) = \mathrm{Ker}(w - \sqrt{\mu}\ \mathrm{Id}).$$

Ainsi les restrictions de v et w aux sous-espaces propres de $v^2 = w^2$ sont égales, E étant somme directe de ces sous-espaces, on en déduit $v = w$. ∎

Corollaire 62.
Soit $A \in S_n^+(\mathbb{R})$. Il existe une unique matrice $B \in S_n^+(\mathbb{R})$ telle que $B^2 = A$.

2 Formes sesquilinéaires Formes hermitiennes

Dans toute cette section, le corps de base est le corps $\mathbb{C}$ des nombres complexes. Nous allons présenter une variante de la théorie des formes bilinéaires symétriques et des formes quadratiques généralisant la version réelle de cette théorie. *L'automorphisme de corps* de $\mathbb{C}$ défini par la *conjugaison* $z = x+iy \mapsto \bar{z} = x-iy$ va jouer un rôle essentiel. Notons que cet automorphisme est *involutif* : $(\bar{\bar{z}}) = z$ et que l'ensemble de ses points fixes est le corps $\mathbb{R}$ des nombres réels. Dans la version réelle de la théorie, la positivité de x^2, quand $x \in \mathbb{R}$, est, comme nous l'avons vu, fondamentale ; dans le cas complexe, si $z \in \mathbb{C}$, z^2 n'est pas en général réel et la notion de forme quadratique complexe n'est donc pas satisfaisante. On remarque que, pour $x \in \mathbb{R}$, $x^2 = |x|^2$, ainsi, pour $z \in \mathbb{C}$, on peut considérer $|z|^2 = z\bar{z}$ qui est réel et positif, on est amené à définir une norme sur $\mathbb{C}^n$ par $\|(z_1, \dots, z_n)\|^2 = |z_1|^2 + \cdots + |z_n|^2 = z_1\bar{z}_1 + \cdots + z_2\bar{z}_2$. À partir de ces remarques,

on peut définir une notion de « produit scalaire » sur un $\mathbb{C}$-espace vectoriel, qui étend celle introduite dans le cas d'un $\mathbb{R}$-espace vectoriel, puis développer les notions associées. Cette généralisation est motivée par diverses applications importantes ; nous verrons l'une d'entre-elles à propos des séries de Fourier (*cf.* module II.4), une autre, qui sort du cadre de cet ouvrage, est la formulation hilbertienne de la mécanique quantique (bien que les grandeurs physiques soient réelles, le concept central de fonction d'onde oblige à utiliser les nombres complexes).
Les définitions, les résultats fondamentaux et leurs preuves sont très voisins de ceux du cas réel, en conséquence, nous n'énoncerons que les résultats centraux et ne détaillerons que les points délicats. Le lecteur est néanmoins invité à faire soigneusement toutes les vérifications pour se familiariser avec ce nouveau formalisme.

2.1 Formes sesquilinéaires

2.1.1 Généralités

Définition 10. Soit $A := (a_{hk}) \in M_{np}(\mathbb{C})$, on appelle *matrice conjuguée* de A la matrice $\overline{A} := (\overline{a_{hk}}) \in M_{np}(\mathbb{C})$ dont les coefficients sont conjugués de ceux de A ; on appelle *matrice adjointe* (ou transconjuguée) de M la matrice $M^* := {}^t\overline{M}$ transposée de $\bar{M}$.

Règles de calcul

On vérifie les règles de calcul suivantes :

(i) Pour toutes matrices $A, B \in M_{np}(\mathbb{C})$: $\overline{\overline{A}} = A$, $(A^*)^* = A$, $(\overline{A+B}) = \overline{A} + \overline{B}$, $(A+B)^* = A^* + B^*$; les applications $A \mapsto \overline{A}$ et $A \mapsto A^*$ sont des automorphismes involutifs du groupe additif $M_{np}(\mathbb{C})$. Pour tous $\alpha \in \mathbb{C}$ et $A \in M_{np}(\mathbb{C})$, on a : $(\overline{\alpha A}) = \overline{\alpha}\overline{A}$ et $(\alpha A)^* = \overline{\alpha} A^*$: on dit que les deux automorphismes sont *semi-linéaires* pour la structure de $\mathbb{C}$-espace vectoriel sur $M_{np}(\mathbb{C})$. (Ces automorphismes sont $\mathbb{R}$-linéaires.)
(ii) Pour tout $(A, B) \in M_{mn}(\mathbb{C}) \times M_{np}(\mathbb{C})$: $(\overline{AB}) = \overline{A}\,\overline{B}$ et $(AB)^* = B^* A^*$.
(iii) Pour toute matrice carrée $A \in M_n(\mathbb{C})$: $\det\overline{A} = \det A^* = \overline{\det A}$, si A est inversible : $\overline{A^{-1}} = (\overline{A})^{-1}$, $(A^{-1})^* = (A^*)^{-1}$.
(iv) Pour toute matrice $A \in M_{np}(\mathbb{C})$: $\operatorname{rang} A = \operatorname{rang}\overline{A} = \operatorname{rang} A^*$.
Notons que si A est à coefficients réels : $\overline{A} = A$ et $A^* = {}^tA$.

Matrices hermitiennes

L'adjectif *hermitien(ne)* vient du nom de Charles Hermite (mathématicien français 1822-1901).

Définition 11. Soit $n \in \mathbb{N}^*$, une matrice $A \in M_n(\mathbb{C})$ est dite *hermitienne* (resp. antihermitienne) si $A^* = A$ (resp. $A^* = -A$).

Si $n = 1$, la matrice A est hermitienne si, et seulement si, son unique coefficient est réel. Notons que si A est à coefficients réels, elle est hermitienne si, et seulement si, elle est symétrique.

Proposition 63. Soit $n \in \mathbb{N}^*$, l'ensemble noté $H_n(\mathbb{C})$ des matrices hermitiennes de $M_n(\mathbb{C})$ est un espace vectoriel *réel* de dimension n^2. Pour toute matrice $M \in H_n(\mathbb{C})$, il existe un unique couple $(A, B) \in S_n(\mathbb{R}) \times A_n(\mathbb{R})$ tel que $M = A + iB$ et le déterminant de M est *réel*.

On prendra garde au fait que $H_n(\mathbb{C})$ n'est pas un sous-espace vectoriel complexe de $M_n(\mathbb{C})$, mais seulement un sous-espace réel de $M_n(\mathbb{C})$ pour sa structure sous-jacente de $\mathbb{R}$-espace vectoriel (de dimension $2n^2$). Considérer le cas $n = 1$!

Démonstration. Soit $M := (m_{hk})_{h,k\in[\![1,n]\!]}$, $M \in H_n(\mathbb{C})$ si, et seulement si, $M = M^*$, i.e. si, et seulement si, $m_{hk} = \overline{m_{kh}}$, pour tous $k, h \in [\![1, n]\!]$. Soit $(a_{hk}, b_{hk})) \in \mathbb{R}^2$ tel que $m_{hk} = a_{hk} + ib_{hk}$, on a : $a_{hk} = a_{kh}$ et $b_{hk} = -b_{kh}$. Si $A := (a_{hk})_{h,k\in[\![1,n]\!]}$ et $B := (b_{hk})_{h,k\in[\![1,n]\!]}$, on a : $M = A + iB$, $A \in S_n(\mathbb{R})$ et $B \in A_n(\mathbb{R})$. On vérifie sans difficulté l'unicité d'une telle décomposition. Ainsi $H_n(\mathbb{C})$ est un $\mathbb{R}$-espace vectoriel isomorphe au $\mathbb{R}$-espace vectoriel $S_n(\mathbb{R}) \times A_n(\mathbb{R})$, sa dimension est donc n^2. Si $M \in H_n(\mathbb{C})$, $\det M^* = \overline{\det M} = \det M$, donc $\det M \in \mathbb{R}$. ∎

Nous n'étudierons pas ici les matrices antihermitiennes. Nous verrons dans le module I.6 qu'elles sont reliées au groupe unitaire (que nous définirons un peu plus loin), comme les matrices antisymétriques le sont au groupe orthogonal, via *l'exponentielle de matrices*.

2.1.2 Formes sesquilinéaires

Définition 12 (Application semi-linéaire). Soient E et F deux $\mathbb{C}$-espaces vectoriels , une application $u : E \to F$ est dite *semi-linéaire* si, et seulement si, pour tout $(\alpha, x, y) \in \mathbb{C} \times E \times E : u(\alpha x + y) = \bar{\alpha}u(x) + u(y)$. Si $F = \mathbb{C}$, on dit que u est une *forme semi-linéaire* sur E.

Une application semi-linéaire est évidemment $\mathbb{R}$-linéaire pour les structures d'espace vectoriel réel sous-jacentes.
On vérifie que l'ensemble des applications semi-linéaires de E dans F est un sous-espace vectoriel du $\mathbb{C}$-espace vectoriel F^E. Une application $u : E \to \mathbb{C}$ est une forme semi-linéaire si, et seulement si, l'application $\bar{u} : E \to \mathbb{C}$ définie par $x \mapsto \overline{u(x)}$ est une forme linéaire.

Définition 13 (Formes sesquilinéaires). Soient E et F deux $\mathbb{C}$-espaces vectoriels, une application $\varphi : E \times F \to \mathbb{C}$ est dite *sesquilinéaire* si :
(i) pour tout $x \in E$, l'application $y \in F \mapsto \varphi(x, y)$ est une forme *linéaire* sur F ;
(ii) pour tout $y \in F$, l'application $x \in E \mapsto \varphi(x, y)$ est une forme *semi-linéaire* sur E.

Sesqui veut dire un et demi (ici une fois linéaire et une fois semi-linéaire).

On prendra garde au fait que certains auteurs adoptent la convention opposée à la nôtre : linéarité par rapport au premier facteur et semi-linéarité par rapport au second. Notre convention s'impose aujourd'hui ; elle est, en particulier, conforme aux règles de calcul en mécanique quantique.
On a les règles de calcul suivantes :
(i) pour tous $x, x' \in E$, $y \in F$, $\alpha \in \mathbb{C}$: $\varphi(\alpha x + x', y) = \bar{\alpha}\,\varphi(x, y) + \varphi(x', y)$;
(ii) pour tous $x \in E$, $y, y' \in F$, $\beta \in \mathbb{C}$: $\varphi(x, \beta y + y') = \beta\,\varphi(x, y) + \varphi(x, y')$.
On prouve facilement le résultat suivant.

Proposition 64. Soient E et F deux $\mathbb{C}$-espaces vectoriels, l'ensemble des formes sesquilinéaires sur $E \times F$ est un sous-espace vectoriel du $\mathbb{C}$-espace vectoriel $\mathbb{C}^{E\times F}$. Si φ est une forme sesquilinéaire sur $E \times F$, pour tous sous-espaces vectoriels E' de E et F' de F, la restriction de φ à $E' \times F'$ est une forme sesquilinéaire sur $E' \times F'$.

2.1.3 Formes sesquilinéaires en dimension finie

Dans cette section, tous les espaces vectoriels considérés sont de dimension finie.

Représentation matricielle d'une forme sesquilinéaire

Théorème et définition 65 (Représentation matricielle). Soient E et F deux $\mathbb{C}$-espaces vectoriels de dimension finie (non nulle) et φ une forme sesquilinéaire sur $E \times F$. Soient $\mathcal{B} := (e_1, \ldots, e_n)$ une base de E et $\mathcal{C} := (\varepsilon_1, \ldots, \varepsilon_p)$ une base de F, on appelle *matrice associée à φ dans les bases $\mathcal{B}$ de E et $\mathcal{C}$ de F* la matrice $M := (m_{hk})_{(h,k)\in[\![1,n]\!]\times[\![1,p]\!]}$ où, pour tout $(h,k) \in [\![1,n]\!] \times [\![1,p]\!]$, $m_{hk} := \varphi(e_h, \varepsilon_k)$. On note $M := \mathrm{Mat}(\varphi; \mathcal{B}, \mathcal{C})$. Pour tous $x := \sum_{h=1}^{n} x_h e_h \in E$ et $y := \sum_{k=1}^{p} y_k \varepsilon_k \in F$, on a $\varphi(x, y) = {}^t\overline{X} M Y$. L'application $\varphi \mapsto M$ est un isomorphisme de $\mathbb{C}$-espaces vectoriels de l'espace vectoriel des formes sesquilinéaires sur $E \times F$ sur $M_{n,p}(\mathbb{C})$.

Démonstration. Pour tous $x := \sum_{h=1}^{n} x_h\, e_h \in E$ et $y := \sum_{k=1}^{p} y_k\, \varepsilon_k \in F$:

$$\varphi(x, y) = \sum_{h=1}^{n} \overline{x_h}\, \varphi(e_h, y) = \sum_{h=1}^{n} \overline{x_h} \sum_{k=1}^{p} y_k\, \varphi(e_h, \varepsilon_k) = {}^t\overline{X} M Y.$$

Inversement, la matrice $M \in M_{n,p}(\mathbb{C})$ étant donnée, l'application $\varphi : (x, y) \mapsto {}^t\overline{X} M Y$ est une forme sesquilinéaire sur $E \times F$. ∎

Changement de base

Proposition 66. Soient $\mathcal{B}$ et $\mathcal{B}'$ deux bases du $\mathbb{C}$-espace vectoriel E, soient $\mathcal{C}$ et $\mathcal{C}'$ deux bases du $\mathbb{C}$-espace vectoriel F, on pose $P := \mathrm{Pass}(\mathcal{B}, \mathcal{B}')$ et $Q := \mathrm{Pass}(\mathcal{C}, \mathcal{C}')$. Pour toute forme sesquilinéaire φ sur $E \times F$, soient $M := \mathrm{Mat}(\varphi; \mathcal{B}, \mathcal{C})$ et $M' := \mathrm{Mat}(\varphi; \mathcal{B}', \mathcal{C}')$, on a : $M' = P^* M Q$. Les matrices M et M' ont le même rang.

Rang d'une forme sesquilinéaire

On appelle *rang* d'une forme sesquilinéaire φ le rang d'une matrice associée à φ. Cette définition est légitime car deux matrices associées à φ, étant équivalentes, ont le même rang.

2.1.4 Formes sesquilinéaires hermitiennes

Dans cette section, $E = F$. On parle alors de forme sesquilinéaire sur E (au lieu de sur $E \times E$).

Définition 14 (Formes sesquilinéaires hermitiennes).
Soit E un $\mathbb{C}$-espace vectoriel, une forme sesquilinéaire φ sur E est dite *hermitienne*, si, pour tous $x, y \in E$: $\varphi(x, y) = \overline{\varphi(y, x)}$.

Remarque. L'ensemble $\mathcal{SH}(E)$ des formes sesquilinéaires hermitiennes sur E n'est pas un sous-espace vectoriel du $\mathbb{C}$-espace vectoriel des formes sesquilinéaires sur E, c'est par contre un sous-espace vectoriel pour la structure sous-jacente d'espace vectoriel réel de E. Pour toute forme $\varphi \in \mathcal{SH}(E)$ et tout $x \in E$, on a $\varphi(x, x) \in \mathbb{R}$.

Proposition 67. Soit E un $\mathbb{C}$-espace vectoriel de dimension $n > 0$.
Soit $\mathcal{B} := (e_1, \ldots, e_n)$ une base de E, une forme φ sesquilinéaire sur E est hermitienne si, et seulement si, sa matrice associée $M := \mathrm{Mat}(\varphi; \mathcal{B}, \mathcal{B})$ est hermitienne.

Dans les conditions ci-dessus, on note $M := \mathrm{Mat}(\varphi; \mathcal{B})$, au lieu de $M = \mathrm{Mat}(\varphi; \mathcal{B}, \mathcal{B})$, la matrice associée.
Soient E un $\mathbb{C}$-espace vectoriel et $\varphi \in \mathcal{SH}(E)$, on peut définir, de la même façon qu'à la page 113, la notion de vecteurs φ-conjugués, d'orthogonalité, de famille φ-orthonormée (*cf.* définition 10 de la page 114), la notion de noyau de φ et celle de vecteurs φ-isotropes. Il suffit de remplacer l'expression *forme bilinéaire symétrique* par l'expression *forme sesquilinéaire hermitienne*. La proposition 28 de la page 114 reste valable pour les formes sesquilinéaires hermitiennes, le théorème 29 de la page 115 peut être reformulé de la façon suivante (la démonstration étant identique) :

Théorème 68 (Existence d'une base de vecteurs φ-conjugués).
Soit E un $\mathbb{C}$-espace vectoriel de dimension $n > 0$. Pour toute forme φ sesquilinéaire hermitienne sur E, il existe une base $\mathcal{B} = (e_1, \ldots, e_n)$ de E formée de vecteurs φ-conjugués. La matrice de φ dans cette base est *diagonale* : $D = \mathrm{Diag}\big(\varphi(e_1, e_1), \ldots, \varphi(e_n, e_n)\big)$. Si φ est de rang r, la base $\mathcal{B}$ contient $n-r$ vecteurs *isotropes* et ceux-ci forment une base du noyau $E^\perp$ de φ.

Espaces vectoriels conjugués

La notion suivante n'est pas indispensable ici, mais elle permet de bien mettre en évidence les parallélismes entre le cas bilinéaire et le cas sesquilinéaire.
Soit E un $\mathbb{C}$-espace vectoriel , c'est un ensemble $\mathcal{E}$, muni d'une loi interne $(w, w') \mapsto w + w'$ et d'une loi externe $(\alpha, w) \mapsto \alpha w$ (vérifiant les conditions qu'il faut). On garde le groupe additif sous-jacent $(\mathcal{E}, +)$ et l'on change la loi externe : la nouvelle loi externe $\star$ est définie, pour tout $(\alpha, w) \in \mathbb{C} \times E$, par $(\alpha, w) \mapsto \alpha \star w := \bar{\alpha} w$. On obtient ainsi un nouvel espace vecto-

riel sur $\mathbb{C}$, il est dit *espace vectoriel conjugué* de E, on le note $\bar{E}$. On prendra garde au fait que l'identité de $\mathcal{E}$ induit un isomorphisme de $\mathbb{R}$-espaces vectoriels entre E et $\bar{E}$, mais que ce n'est pas un isomorphisme de $\mathbb{C}$-espaces vectoriels : il est *semi-linéaire*.

Le passage à l'espace vectoriel conjugué est involutif : $(\overline{E}) = E$. Le lecteur vérifiera que l'identité de $\mathcal{E}$ induit des bijections entre l'ensemble des systèmes libres (respectivement des systèmes générateurs, respectivement des bases) des $\mathbb{C}$-espaces vectoriels E et $\bar{E}$. Les deux espaces sont simultanément de dimension finie et leurs dimensions sont alors égales.

Considérons $\mathbb{C}$ comme $\mathbb{C}$-espace vectoriel, c'est le groupe additif $\mathbb{R}^2$ muni de la loi externe classique, on a : $1(x,y) = (x,y)$ et $i \star (x,y) = (-y,x)$. L'espace vectoriel conjugué de $\mathbb{C}$ est le groupe additif $\mathbb{R}^2$ muni de la loi externe $\star$, on a : $1 \star (x,y) = (x,y)$ et $i \star (x,y) = (y,-x)$, donc $(x,0) + i \star (y,0) = x + i \star y = x - iy$ (la conjugaison revient à changer le choix de $\sqrt{-1}$).

Soient E et F deux $\mathbb{C}$-espaces vectoriels, on vérifie facilement qu'une application $u : E \to F$ est semi-linéaire si, et seulement si, elle induit une application linéaire de E dans $\bar{F}$ (resp. de $\bar{E}$ dans F).

Soit $\varphi : E \times F \to \mathbb{C}$ une forme sesquilinéaire. Pour tout $y \in F$, l'application $f_y : x \in E \mapsto \overline{\varphi(x,y)}$ est une forme *linéaire* sur E, c'est à dire un élément de E^*. De même, pour tout $x \in E$, l'application $h_x : y \in F \mapsto \varphi(x,y)$ est une forme *linéaire* sur F, c'est à dire un élément de F^*. On définit ainsi une application $d_\varphi : F \to E^*$ par $d_\varphi : y \mapsto f_y$ et une application $g_\varphi : E \to F^*$ par $g_\varphi : x \mapsto h_x$. Ces deux applications sont *semi-linéaires* et l'on a, pour tout $(x,y) \in E \times F$: $\overline{d_\varphi(y)(x)} = g_\varphi(x)(y) = \varphi(x,y)$.

Si E et F sont de dimension finie, le rang de φ est égal au rang commun de d_φ et g_φ.

2.2 Formes hermitiennes

2.2.1 Généralités

Proposition et définition 69. Une application Φ, du $\mathbb{C}$-espace vectoriel E dans le corps $\mathbb{R}$ des nombres *réels*, est appelée *forme hermitienne* sur E s'il existe une forme sesquilinéaire hermitienne $\varphi \in \mathcal{SH}(E)$ telle que, pour tout $x \in E$: $\Phi(x) = \varphi(x,x)$. S'il existe une telle forme, elle est unique et appelée *forme polaire* de Φ.

Le résultat de cette proposition est une conséquence des identités dites de « polarisation » ci-dessous.

L'ensemble des formes hermitiennes sur E est noté $\mathcal{H}(E)$, c'est un $\mathbb{R}$-espace vectoriel isomorphe à $\mathcal{SH}(E)$.

Identités dites de polarisation

Soit $\varphi \in \mathcal{SH}(E)$ telle que, pour tout $x \in E$: $\Phi(x) = \varphi(x,x)$.

Pour tout $(x,y) \in E^2$, on a :

$$\begin{aligned}
\Phi(x+y) &= \Phi(x) + \Phi(y) + 2\,\mathrm{Re}\,\varphi(x,y),\\
\Phi(x+iy) &= \Phi(x) + \Phi(y) - 2\,\mathrm{Im}\,\varphi(x,y),\\
\mathrm{Re}\ \varphi(x,y) &= \frac{1}{4}\big(\Phi(x+y) - \Phi(x-y)\big),\\
\mathrm{Im}\ \varphi(x,y) &= -\frac{1}{4}\big(\Phi(x+iy) - \Phi(x-iy)\big).
\end{aligned}$$

Les preuves sont laissées au lecteur (on utilise $\varphi(x,y) = \overline{\varphi(y,x)}$).

Exemple. Le carré du module d'une forme linéaire sur un $\mathbb{C}$-espace vectoriel E.
Soit $f \in E^*$, l'application $\Phi : x \mapsto |f(x)|^2 = \overline{f(x)}f(x)$ est une forme hermitienne de forme polaire $(x,y) \mapsto \varphi(x,y) = \overline{f(x)}f(y)$.

Règle dite de dédoublement

Soit $\mathcal{B} := (e_h)_{h\in I}$ une base du $\mathbb{C}$-espace vectoriel E. Soit $x := \sum_{h\in I} x_h\, e_h$. Pour tout $h \in I$, l'application $x \mapsto |x_h|^2 = \overline{x_h}\, x_h$ est une forme hermitienne de forme polaire $(x,y) \mapsto \overline{x_h}\, y_h$ Pour tout $h,k \in I$, l'application $x \mapsto \overline{x_h}\, x_k + x_h\, \overline{x_k}$ est une forme hermitienne de forme polaire $(x,y) \mapsto (\overline{x_h}\, y_k + y_h\, \overline{x_k})$.

Caractérisation d'une forme hermitienne

Proposition 70. Soit E un $\mathbb{C}$-espace vectoriel. Une application $\Phi : E \to \mathbb{C}$ est une forme hermitienne sur E si, et seulement si, elle vérifie les conditions :
(i) pour tout $(\alpha, x) \in \mathbb{C} \times E) : \Phi(\alpha x) = |\alpha|^2\Phi(x)$;
(ii) l'application φ de $E \times E$ dans $\mathbb{C}$ définie par :

$$(x,y) \mapsto \frac{1}{2}\big(\Phi(x+y) - \Phi(x) - \Phi(y)\big) - \frac{i}{2}\big(\Phi(x+iy) - \Phi(x) - \Phi(y)\big)$$

est sesquilinéaire hermitienne.

Démonstration. Il suffit de vérifier que $\varphi(x,x) = \Phi(x)$. Pour tout $x \in E$:

$$\varphi(x,x) = \frac{1}{2}\big(\Phi(2x) - 2\Phi(x)\big) - i\big(\Phi((1+i)x) - 2\Phi(x)\big)\big) = \Phi(x),$$

car $\Phi(2x) = 4\Phi(x)$ et $\Phi((1+i)x) = |1+i|^2\Phi(x) = 2\Phi(x)$. ∎

2.2.2 Formes hermitiennes en dimension finie

Dans cette section, les espaces vectoriels sont supposés de dimension finie.

Matrice associée à une forme hermitienne

Définition 15. Soient E un $\mathbb{C}$-espace vectoriel de dimension $n > 0$ et $\mathcal{B} := (e_1, \ldots, e_n)$ une base de E. Soit $\Phi \in \mathcal{H}(E)$, on appelle matrice de la forme hermitienne Φ dans la base $\mathcal{B}$, la matrice de la forme polaire φ de Φ dans la base $\mathcal{B}$. On note $\mathrm{Mat}(\Phi;\mathcal{B}) := \mathrm{Mat}(\varphi;\mathcal{B})$

Proposition 71. Soient E un $\mathbb{C}$-espace vectoriel de dimension $n > 0$.
Si $\mathcal{B} := (e_1, \ldots, e_n)$ est une base de E, l'application qui associe à une forme hermitienne Φ sur E sa matrice dans la base $\mathcal{B}$ est un isomorphisme d'espaces vectoriels *réels* de $\mathcal{H}(E)$ sur $H_n(\mathbb{C})$.

Décomposition en carrés

Le théorème suivant découle immédiatement du théorème 68 de la page 190.

Théorème 72 (Décomposition en carrés). Soit Φ une forme hermitienne de rang r sur le $\mathbb{C}$-espace vectoriel E de dimension $n > 0$.
(i) Il existe r formes linéaires *indépendantes* sur E : $f_1, \dots, f_r$ et r nombres réels *non nuls* : $\alpha_1, \dots, \alpha_r$ tels que $\Phi = \sum_{i=1}^{r} \alpha_i |f_i|^2$.
(ii) Si la forme hermitienne Φ est combinaison linéaire à coefficients réels non nuls des carrés des *modules* de s formes linéaires indépendantes sur E, alors $s = r$.

2.2.3 Formes hermitiennes positives, définies positives

On rappelle que si Φ est une forme hermitienne sur un $\mathbb{C}$-espace vectoriel E, alors $\Phi(x)$ est *réel* pour tout $x \in E$.

Définition 16. Soient E un $\mathbb{C}$-espace vectoriel, Φ une forme hermitienne sur E, φ sa forme polaire. On dit que Φ et φ sont *positives* si $\forall x \in E$, $\Phi(x) \geqslant 0$. On dit que Φ et φ sont *définies positives* si $\forall x \in E - \{0\}$, $\Phi(x) > 0$. On dit que Φ et φ sont *négatives* (resp. *définies négatives*) si $-\Phi$ et $-\varphi$ sont positives (resp. définies positives).

On notera respectivement $\mathcal{H}^+(E)$ et $\mathcal{H}^{++}(E)$ *l'ensemble* des formes hermitiennes positives et l'ensemble des formes hermitiennes définies positives sur l'espace vectoriel complexe E.

Exemples.
(i) *Norme de Schur*. Soit $\varphi : M_n(\mathbb{C}) \times M_n(\mathbb{C}) \to \mathbb{C}$ définie par $\varphi(A, B) = \operatorname{Tr} A^* B$. L'application φ est sesquilinéaire. Elle est *définie positive*, pour tout $A \in M_n(\mathbb{C})$:

$$\operatorname{Tr} A^*A = \operatorname{Tr} {}^t\overline{A}A = \sum_{h,k \in [\![1,n]\!]^2} |a_{hk}|^2 \geqslant 0$$

et $\operatorname{Tr} A^*A = 0$ si, et seulement si, les nombres *complexes* a_{hk} sont nuls, quel que soit $(h, k) \in [\![1, n]\!]^2$, c'est-à-dire si, et seulement si, $A = 0$. On appelle $\sqrt{\operatorname{Tr} A^*A}$ la *norme de Schur* de la matrice A.
(ii) Soit $(a, b) \in \mathbb{R}^2$, avec $a < b$. On note $E = \mathcal{C}([a, b], \mathbb{C})$ le $\mathbb{C}$ espace vectoriel des fonctions continues sur $[a, b]$ à valeurs complexes. L'application $\varphi : E \times E \to \mathbb{C}$ définie par $(f, g) \mapsto \int_a^b \overline{f(t)} g(t) \, \mathrm{d}t$ est sesquilinéaire et *définie positive*.

Proposition 73. Soit Φ une forme hermitienne *définie* sur le $\mathbb{C}$-espace E.
La forme Φ est *définie positive* ou *définie négative*.

Démonstration. Supposons qu'il existe $(x, y) \in E^2$ tel que $\Phi(x) > 0$ et $\Phi(y) < 0$. On

considère le *trinôme* (en $t \in R$) :

$$h(t) = \Phi(x+ty) = \Phi(y)\,t^2 + 2\big(\operatorname{Re}\varphi(x,y)\big)t + \Phi(x),$$

son *discriminant* (réduit) $\big(\operatorname{Re}\varphi(x,y)\big)^2 - \Phi(x)\Phi(y)$ est strictement positif, on termine comme dans le cas réel (*cf.* la proposition 40 de la page 130). ∎

Inégalité de Cauchy-Schwarz

Théorème 74. Soit Φ une forme hermitienne positive sur le $\mathbb{C}$-espace vectoriel E. Pour tout $(x,y) \in E^2$ on a :

$$|\varphi(x,y)|^2 \leqslant \Phi(x)\Phi(y).$$

Si de plus Φ est *définie positive*, l'égalité $|\varphi(x,y)|^2 = \Phi(x)\Phi(y)$ est vraie si, et seulement si, x et y sont *colinéaires*.

Démonstration. Soit $(x,y) \in E^2$ fixés, pour tout $(\theta,t) \in \mathbb{R}^2$, on considère :

$$h(\theta,t) := \Phi(x+te^{i\theta}y) = \Phi(y)t^2 + t\big(e^{i\theta}\varphi(x,y) + e^{-i\theta}\overline{\varphi(x,y)}\big) + \Phi(x).$$

Puisque Φ est *positive*, on a $h(\theta,t) \geqslant 0$.
Il existe $\alpha \in \mathbb{R}$ tel que $\varphi(x,y) = e^{i\alpha}|\varphi(x,y)|$, on considère alors le trinôme (en t) : $h(t) := h(-\alpha,t) = \Phi(y)t^2 + 2t|\varphi(x,y)| + \Phi(x)$. On a $h(t) \geqslant 0$ et l'on peut donc conclure comme dans la preuve du théorème 41 de la page 130. ∎

Corollaire 75. Soit Φ une forme hermitienne *positive* sur le $\mathbb{C}$-espace vectoriel E. Son *cône isotrope* et le *noyau* de sa forme polaire φ sont égaux. Ainsi Φ est définie positive si, et seulement si, elle est non dégénérée.

Corollaire 76 (Inégalité de Minkowski). Soit Φ une forme hermitienne positive sur le $\mathbb{C}$-espace vectoriel E, pour tous $x, y \in E$: $\sqrt{\Phi(x+y)} \leqslant \sqrt{\Phi(x)} + \sqrt{\Phi(y)}$.

Démonstration. Les termes considérés étant positifs, l'inégalité ci-dessus est vraie si, et seulement si, $\Phi(x+y) \leqslant \Phi(x) + \Phi(y) + 2\sqrt{\Phi(x)\Phi(y)}$, soit $\operatorname{Re}\varphi(x,y) \leqslant \sqrt{\Phi(x)\Phi(y)}$.
On sait d'après l'inégalité de Cauchy-Schwarz (théorème 74) que :

$$|\operatorname{Re}\varphi(x,y)| \leqslant |\varphi(x,y)| \leqslant \sqrt{\Phi(x)\Phi(y)},$$

d'où le résultat. ∎

Signature d'une forme hermitienne

En reprenant la définition 17 de la page 132 où l'on remplace *forme quadratique* par *forme hermitienne*, on définit la *signature* d'une forme hermitienne.
On a une version hermitienne du théorème d'inertie de Sylvester. La preuve est analogue, en remplaçant x_h^2 par $|x_h|^2$.

Théorème 77 (Théorème d'inertie de Sylvester). Soient E un $\mathbb{C}$-espace vectoriel de dimension $n > 0$ et Φ une forme hermitienne sur E de signature (p, q). Pour toute base $\mathcal{B}$ de E formée de vecteurs deux à deux φ-conjugués, la matrice diagonale $\mathrm{Mat}(\varphi; \mathcal{B}) = \mathrm{Diag}(\alpha_1, \ldots, \alpha_n)$ est formée de p réels strictement positifs, de q réels strictement négatifs et de $n - p - q$ zéros. Le rang de Φ est $p + q$.

Corollaire 78. Soient E un $\mathbb{C}$-espace vectoriel de dimension $n > 0$ et $\Phi \in \mathcal{H}(E)$ de signature (p, q). Il existe une base $\mathcal{B}_0$ de E telle que : $\mathrm{Mat}(\varphi; \mathcal{B}_0) = \mathrm{Diag}(\underbrace{1, \ldots, 1}_{p \text{ fois}}, \underbrace{-1, \ldots, -1}_{q \text{ fois}}, \underbrace{0, \ldots, 0}_{n-p-q \text{ fois}})$. *Si Φ est définie positive, $p = n$ et la base $\mathcal{B}_0$ est φ-orthonormée.*

Matrices hermitiennes définies positives. Décomposition de Cholesky

Définition 17. Une matrice $A \in H_n(\mathbb{C})$ est dite *positive* si, pour tout $x \in \mathbb{C}^n$, on a $X^*AX \geqslant 0$. Elle est dite *définie positive* si, pour tout $x \in \mathbb{C}^n \setminus \{0\}$, on a $X^*AX > 0$.

On notera $H_n^+(\mathbb{C})$ (resp. $H_n^{++}(\mathbb{C})$) *l'ensemble* des matrices positives (resp. définies positives) de $H_n(\mathbb{C})$.

Proposition 79. Soit $A \in H_n(\mathbb{C})$. Elle est *positive* si, et seulement si, il existe une matrice inversible $P \in GL_n(\mathbb{C})$ et un entier $p \in [\![0, n]\!]$ tels que $A = P^*DP$, avec $D := \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}$. La matrice A est *définie positive* si, et seulement s'il existe une matrice $P \in GL_n(\mathbb{C})$ telle que $A = P^*P$.

Proposition 80. Soit $M \in M_{m,n}(\mathbb{C})$. La matrice M^*M appartient à $H_n^+(\mathbb{C})$, $\mathrm{Ker}\, M = \mathrm{Ker}\, M^*M$ et $\mathrm{rang}\, M = \mathrm{rang}\, M^*M$.

En reprenant la preuve du corollaire 39 de la page 124 on montre la proposition suivante (l'écriture $A = LD\,{}^tL$ est changée en l'écriture $A = LDL^*$).

Proposition 81. Une matrice hermitienne est *définie positive* si, et seulement si, tous ses *mineurs principaux* sont *strictement positifs*.

On en déduit la décomposition dite *de Cholesky* d'une matrice de $H_n^{++}(\mathbb{C})$.

Proposition 82 (Décomposition de Cholesky).
Pour toute matrice $A \in H_n^{++}(\mathbb{C})$, il existe une *unique* matrice $T \in M_n(\mathbb{C})$ *triangulaire supérieure*, d'éléments diagonaux réels *strictement positifs* telle que $A = T^*T$.

3 Espaces vectoriels préhilbertiens complexes, espaces hermitiens

3.1 Produit scalaire, norme hermitienne

Définition 18. On appelle *produit scalaire* sur le $\mathbb{C}$-espace vectoriel E, toute forme sesquilinéaire hermitienne définie positive.

Pour tous $x, y \in E$, on notera en général $(\,x \mid y\,)$ leur produit scalaire. On appelle *espace vectoriel préhilbertien complexe* le couple $\big(E, (\,|\,)\big)$ formé du $\mathbb{C}$-espace vectoriel E et d'un produit scalaire sur E. Si la dimension de E est *finie*, l'espace vectoriel préhilbertien complexe E est dit *espace hermitien*. Lorsqu'il n'y aura pas d'ambiguïté, on allégera la notation $\big(E, (\,|\,)\big)$ en E.

Exemples.

(i) L'espace $\mathbb{C}^n$ muni du produit scalaire canonique $(x, y) \mapsto \sum_{h=1}^{n} \overline{x}_h\, y_h$ est un espace hermitien. C'est la structure hermitienne canonique de $\mathbb{C}^n$.
(ii) Soit E l'ensemble des suites complexes $x := (x_n)_{n\geqslant 0} \in \mathbb{C}^{\mathbb{N}}$, *de carré sommable*, c'est à dire telles que $\sum_{n\geqslant 0} |x_n|^2 < +\infty$. Comme dans le cas réel, on vérifie que E est un sous-$\mathbb{C}$-espace vectoriel de $\mathbb{C}^{\mathbb{N}}$ et que l'application $(x, y) \mapsto (\,x \mid y\,) := \sum_{n=0}^{+\infty} \overline{x_n}\, y_n$ est un produit scalaire sur E.
L'espace E muni du produit scalaire $(\,|\,)$est un espace préhilbertien complexe, c'est, par définition, l'espace $\ell^2(\mathbb{C})$.

La restriction d'un produit scalaire à un sous-espace définit un produit scalaire sur celui-ci.

Norme hermitienne

Soit $(E, (\,|\,))$ un espace préhilbertien complexe On note $\|x\| := \sqrt{(\,x \mid x\,)} = (\,x \mid x\,)^{1/2}$. L'analogue de la proposition 1 de la page 143 est vraie (dans (ii), il faut évidemment supposer que le scalaire α est complexe), $\|.\|$ est une norme sur E, nous dirons que $\|x\|$ est la *norme hermitienne* du vecteur $x \in E$. On peut donc lui associer une distance d sur E. On vérifie aussi l'égalité de la médiane.

3.2 Orthogonalité, distance d'un point à un sous-espace de dimension finie, inégalité de Bessel

L'orthogonalité et les systèmes de vecteurs deux à deux orthogonaux sont définis comme dans le cas réel (*cf.* page 144). Un système orthogonal de vecteurs *non nuls* d'un espace préhilbertien complexe est *libre*.

Proposition 83. Soit E un espace préhilbertien complexe admettant une base orthonormée $(e_h)_{h\in I}$. Si $x, y \in E$, on note $x := \sum_{h\in I} x_h e_h$, $y := \sum_{h\in I} y_h e_h$. On a :

- pour tout $h \in I$, $x_h = (\, e_h \mid x\,)$, d'où $x = \sum_{h\in I} (\, e_h \mid x\,)\, e_h$;
- $(\, x \mid y\,) = \sum_{h\in I} \overline{x_h}\, y_h$, $\|x\|^2 = (\, x \mid x\,) = \sum_{i\in I} |x_h|^2$.

Comme dans le cas réel, on a un théorème de Pythagore (*cf.* la proposition 6 de la page 145), mais on *prendra garde* à la différence de formulation. Le lecteur adaptera la démonstration du cas réel.

Proposition 84 (Théorème de Pythagore).
Soient E un espace préhilbertien complexe. Si les deux vecteurs $x, y \in E$ sont orthogonaux, on a $\|x+y\|^2 = \|x\|^2 + \|y\|^2$. Plus généralement, si $x_1, \ldots, x_k$ sont des vecteurs deux à deux orthogonaux de E, on a $\|x_1 + \cdots + x_k\|^2 = \|x_1\|^2 + \cdots + \|x_k\|^2$.

Tout ce qui concerne les sous-espaces orthogonaux, les sommes directes orthogonales, les supplémentaires orthogonaux, est formulé et prouvé comme dans le cas réel.
Si E est un espace *hermitien*, on prouve *l'important* résultat suivant.

Théorème 85 (Toute forme linéaire sur un espace hermitien est un produit scalaire). Soit E un espace vectoriel hermitien :
(i) Pour toute forme linéaire f sur E il *existe* un *unique* vecteur $a \in E$ tel que $f(x) = (\, a \mid x\,)$, pour tout $x \in E$.
(ii) L'application $a \mapsto a^\flat$ de E dans E^* qui associe au vecteur $a \in E$, la forme linéaire $a^\flat : x \mapsto (\, a \mid x\,)$ est un *isomorphisme semi-linéaire* d'espaces vectoriels complexes.

On notera que ce résultat n'est pas vrai, en général, pour un espace préhilbertien de dimension infinie.

Théorème 86. Soit E un espace vectoriel *hermitien*. Pour tout sous-espace vectoriel F de E, $E = F \overset{\perp}{\oplus} F^\perp$, $\dim F + \dim F^\perp = \dim E$, $(F^\perp)^\perp = F$. Il existe une base orthonormée de E formée de la réunion d'une base orthonormée de F et d'une base orthonormée de $F^\perp$. En particulier E admet une base orthonormée.

Projecteurs orthogonaux, symétries orthogonales

On peut recopier tout ce qui concerne les projecteurs orthogonaux et les symétries orthogonales d'un espace préhilbertien réel. On remplace *espace préhilbertien réel* par *espace préhilbertien complexe* et *euclidien* par *hermitien*.

Calcul du projecteur orthogonal sur un sous-espace

Le calcul de la projection orthogonale sur un sous-espace vectoriel de dimension finie est le même que dans le cas réel :

Proposition 87. Soient E un espace hermitien, $F \subset E$ un sous-espace vectoriel de dimension $k \geqslant 1$ et p le projecteur orthogonal de E sur F. Soit $(\varepsilon_1, \ldots, \varepsilon_k)$ une base *orthonormée* de F. Pour tout $x \in E$, on a : $p(x) = \sum_{h=1}^{k} (\varepsilon_h \mid x)\, \varepsilon_h$.

Distance d'un point à un sous-espace de dimension finie

La distance d'un point à un sous-espace de dimension finie est définie et étudiée comme dans le cas réel. On a l'analogue du théorème 14 de la page 152 et de l'inégalité de Bessel qui s'en déduit.

Théorème 88. Soient E un espace préhilbertien complexe et $F \subset E$ un sous-espace vectoriel de dimension *finie* $k \geqslant 1$. Le sous-espace F admet un *supplémentaire orthogonal*. Pour tout $a \in E$, $d(a, F) = \|a - p(a)\|$, p étant le projecteur orthogonal sur F. De plus, pour tout $b \in F$ tel que $b \neq p(a)$, on a $d(a, F) < d(a, b)$.

Corollaire 89 (Inégalité de Bessel).
Soient E un espace préhilbertien complexe, $F \subset E$ un sous-espace vectoriel de dimension *finie* $k \geqslant 1$ et $(\varepsilon_1, \ldots, \varepsilon_k)$ une base *orthonormée* de F. Soit p le projecteur orthogonal sur F. Pour tout $x \in E$, on a $\|x\|^2 \geqslant \|p(x)\|^2 = \sum_{h=1}^{k} (\varepsilon_h \mid x)^2$.

Matrices de Gram

On définit les matrices de Gram comme dans le cas réel. leurs propriétés sont les mêmes, *mutatis mutandis* (on change $\mathbb{R}$ en $\mathbb{C}$ et S_n^{++} en H_n^{++}).

Orthogonalisation de Gram-Schmidt

Le procédé d'orthogonalisation de Gram-Schmidt s'étend au cas préhilbertien complexe et l'on peut énoncer la variante du théorème 18 de la page 155. Les applications de ce procédé sont similaires, on a en particulier une variante de l'inégalité d'Hadamard pour une matrice carrée à coefficients complexes (en utilisant la norme hermitienne usuelle des vecteurs colonnes).

3.3 Adjonction, groupe unitaire, endomorphismes hermitiens

On définit l'adjoint d'un endomorphisme d'un espace préhilbertien complexe comme dans le cas réel.

Théorème et définition 90. Soit E un espace vectoriel préhilbertien complexe. Soit u un endomorphisme de E, s'il existe un endomorphisme v de E tel que, pour tout $(x, y) \in E^2$, $(\,u(x) \mid y\,) = \big(x \mid v(y)\big)$.
Cet endomorphisme est unique ; on l'appelle *l'adjoint* de u ; on le note u^*.
Si v est l'adjoint de u, v admet u pour adjoint : $v^* = (u^*)^* = u$.

Notons que, si $\alpha \in \mathbb{C}$, on a : $(\alpha u)^* = \overline{\alpha} u^*$, en particulier l'adjoint de $\alpha\ \mathrm{Id}_E$ est $\overline{\alpha}\ \mathrm{Id}_E$. Ainsi l'ensemble des endomorphismes de E ayant un adjoint est une sous-algèbre de la $\mathbb{C}$-algèbre $\mathcal{L}(E)$, mais le passage à l'adjoint est *semi-linéaire*.

Proposition 91. Soit E un espace préhilbertien complexe. On a $(\mathrm{Id}_E)^* = \mathrm{Id}_E$, si $u, v \in \mathcal{L}(E)$ ont des adjoints u^* et v^*, $(uv)^* = v^* u^*$, si u et son inverse u^{-1} admettent des adjoints, alors $(u^*)^{-1} = (u^{-1})^*$.

Proposition 92. Soit u un endomorphisme de l'espace préhilbertien E admettant un adjoint u^*. Alors $\operatorname{Ker} u^* = (\operatorname{Im} u)^\perp$ et $\operatorname{Im} u^* \subset (\operatorname{Ker} u)^\perp$. Si $F \subset E$ est un sous-espace vectoriel *stable* par u, alors $F^\perp \subset E$ est un sous-espace vectoriel stable par u^*.

Adjoint d'un endomorphisme d'un espace hermitien

On a, comme dans le cas euclidien, le résultat suivant.

Proposition 93. Tout endomorphisme u d'un espace hermitien E admet un adjoint.

Proposition 94. Soient $\mathcal{B}$ une base orthonormée de l'espace hermitien E et u un endomorphisme de E.
Si $\mathrm{Mat}_{\mathcal{B}}(u) := M$, alors $\mathrm{Mat}_{\mathcal{B}}(u^*) = \overline{{}^t M}$. On a $\operatorname{rang} u = \operatorname{rang} u^*$ et $\overline{\det u} = \det u^*$.

Proposition 95. Soit u un endomorphisme d'un espace hermitien E.
On a $\operatorname{Ker} u^* = (\operatorname{Im} u)^\perp$ et $\operatorname{Im} u^* = (\operatorname{Ker} u)^\perp$.
Un sous-espace $F \subset E$ est u-stable si, et seulement si, $F^\perp$ est u^*-stable.

Groupe unitaire

Théorème et définition 96. Soit E un espace vectoriel préhilbertien complexe. Pour tout automorphisme $u \in \mathcal{GL}(E)$ les assertions suivantes sont équivalentes :
(i) Pour tout $x \in E$, $\|u(x)\| = \|x\|$ (on dit que u conserve la norme).
(ii) Pour tout $(x, y) \in E^2$, $(\,u(x) \mid u(y)\,) = (\,x \mid y\,)$ (on dit que u conserve le produit scalaire).
(iii) L'automorphisme u a un adjoint u^* et $u^* = u^{-1}$.
Si u vérifie l'une quelconque de ces assertions, on dit que u est un automorphisme *unitaire* de E. On note $\mathcal{U}(E)$ l'ensemble des automorphismes unitaires de E.

Les automorphismes unitaires sont des isométries. La proposition 29 de la page 166 reste vraie en remplaçant $\mathcal{O}(E)$ par $\mathcal{U}(E)$.

Proposition 97. L'ensemble $\mathcal{U}(E)$ des automorphismes unitaires de l'espace préhilbertien complexe E est un *groupe* (un sous-groupe de $\mathcal{GL}(E)$), on l'appelle *le groupe unitaire* de E.

Groupe unitaire d'un espace hermitien

Un endomorphisme d'un espace *hermitien* est unitaire si, et seulement s'il conserve la norme hermitienne.

Proposition 98. Soit u un *endomorphisme* de l'espace hermitien E, les assertions suivantes sont équivalentes :
(i) $u \in \mathcal{U}(E)$;
(ii) l'image par u de toute base orthonormée de E est une base orthonormée ;
(iii) il existe une base orthonormée de E dont l'image par u est une base orthonormée.

Théorème et définition 99. Soit E un espace hermitien de dimension strictement positive. L'application qui associe à tout élément de $\mathcal{U}(E)$ son déterminant est un morphisme de $\mathcal{U}(E)$ dans le sous groupe multiplicatif $\mathbb{U}$ des nombres complexes de module 1. Le noyau de ce morphisme est un sous-groupe de $\mathcal{U}(E)$ dit *groupe spécial unitaire* de E. Il est noté $\mathcal{SU}(E)$.

Démonstration. Soit $u \in \mathcal{U}(E)$, on a $uu^* = \mathrm{Id}_E$ et $\det u = \overline{\det u^*}$, d'où $|\det u|^2 = 1$ et $\det u \in \mathbb{U}$ puisque $\det u \in \mathbb{C}$. ∎

Proposition 100. Si $\lambda \in \mathbb{C}$ est une valeur propre de $u \in \mathcal{U}(E)$, alors $\lambda \in \mathbb{U}$.

Matrices unitaires

Nous nous intéressons ici au groupe unitaire $\mathcal{U}(\mathbb{C}^n)$, l'espace $\mathbb{C}^n$ étant muni de son produit scalaire canonique. C'est à un isomorphisme (non canonique) près le cas général : si E est un espace hermitien de dimension n, le choix d'une base orthonormée de E donne un isomorphisme de groupes entre $\mathcal{U}(E)$ et $\mathcal{U}(\mathbb{C}^n)$. Cette étude sera approfondie dans le module I.6.

Théorème et définition 101. Soit $\mathcal{B}$ la base canonique de $\mathbb{C}^n$, l'application qui associe à tout automorphisme unitaire de $\mathbb{C}^n$ sa matrice dans la base $\mathcal{B}$ est un homomorphisme injectif de groupes de $\mathcal{U}(\mathbb{C}^n)$ dans $GL_n(\mathbb{C})$. Son image est appelée *groupe des matrices unitaires* de $M_n(\mathbb{C})$, et notée $U_n(\mathbb{C})$. Le groupe $U_n(\mathbb{C})$ est isomorphe au groupe $\mathcal{U}(\mathbb{C}^n)$. L'image de $\mathcal{SU}(\mathbb{C}^n)$ est notée SU_n, c'est un sous-groupe du groupe spécial linéaire $SL_n(\mathbb{C})$.

La démonstration est laissée au lecteur, de même que celle de la proposition suivante.

Proposition 102. Soit $P \in M_n(\mathbb{C})$, les conditions suivantes sont équivalentes :
(i) la matrice P est unitaire ;
(ii) $PP^* = P\overline{{}^tP} = I_n$;
(iii) $P^*P = \overline{{}^tP}P = I_n$;
(iv) les vecteurs colonnes de P forment une base orthonormée de $\mathbb{C}^n$;
(v) les vecteurs lignes de P forment une base orthonormée de $\mathbb{C}^n$.

Endomorphismes hermitiens et antihermitiens

Théorème et définition 103. Soit E un espace préhilbertien complexe, pour tout endomorphisme u de E, les deux assertions suivantes sont équivalentes :
(i) pour tout $(x, y) \in E^2$, $\big(\, u(x) \,\big|\, y \,\big) = \big(\, x \,\big|\, u(y) \,\big)$ (resp. $\big(\, u(x) \,\big|\, y \,\big) = -\big(\, x \,\big|\, u(y) \,\big)$) ;
(ii) u admet un adjoint u^* et $u^* = u$ (resp. $u^* = -u$).
Si u vérifie l'une de ces deux propriétés, on dit que u est un endomorphisme *hermitien* ou *autoadjoint* (resp. *antihermitien*). L'ensemble des endomorphismes hermitiens de E (respectivement antihermitiens) est noté $\mathbf{H}(E)$ (respectivement $\mathbf{AS}(E)$), ce sont des $\mathbb{R}$-espaces vectoriels supplémentaires du $\mathbb{R}$-espace vectoriel des endomorphismes de E ayant un adjoint, l'application $u \mapsto iu$ est un isomorphisme de $\mathbf{H}(E)$ sur $\mathbf{AS}(E)$.

Projecteurs orthogonaux. Un projecteur d'un espace préhilbertien complexe est *orthogonal* si, et seulement s'il est hermitien. Les projecteurs orthogonaux sont extrêmement importants dans un certain nombre d'applications, surtout en dimension infinie (analyse, mécanique quantique), leur étude approfondie sort du cadre de cet ouvrage et nous ne l'aborderons pas.

Définition 19. Soit E un espace préhilbertien complexe, un endomorphisme $u \in \mathbf{H}(E)$ est dit *positif* (resp. *défini positif*) si, et seulement si, pour tout $x \in E$ (resp. $x \in E \setminus \{0\}$), $\big(\, u(x) \;\big|\; x \,\big) \geqslant 0$ (resp. $\big(\, u(x) \;\big|\; x \,\big) > 0$).
On note $\mathbf{H}^{+}(E)$ l'ensemble des endomorphismes hermitiens positifs, $\mathbf{H}^{++}(E)$ celui des endomorphismes hermitiens définis positifs.

On montre sans difficulté les résultats suivants.

Proposition 104. Soit E un espace préhilbertien complexe, pour tout $u \in \mathbf{H}(E)$, l'application de E^2 dans $\mathbb{C}$, qui associe à (x, y) le nombre $\big(\, u(x) \;\big|\; y \,\big)$ est une forme sesquilinéaire hermitienne sur E. Elle est positive (resp. définie positive) si, et seulement si, u l'est.

Proposition 105. Soit E un espace préhilbertien complexe, pour tout $u \in \mathcal{L}(E)$ ayant un adjoint u^*, les endomorphismes u^*u et uu^* sont hermitiens *positifs*, u^*u est défini positif si, et seulement si, u est *injectif*, de même uu^* est défini positif si, et seulement si, u^* est *injectif*.

Endomorphismes hermitiens et antihermitiens d'un espace hermitien

Proposition 106. Soit $\mathcal{B} := (e_1, \ldots, e_n)$ une base *orthonormée* de l'espace vectoriel hermitien E de dimension $n > 0$.
(i) L'application qui à un endomorphisme $u \in \mathbf{H}(E)$ $\big($respectivement $u \in \mathbf{AH}(E)\big)$ associe sa matrice M dans la base $\mathcal{B}$ est un *isomorphisme* d'espaces vectoriels réels de $\mathbf{H}(E)$ $\big($respectivement $\mathbf{AH}(E)\big)$ sur $H_n(\mathbb{C})$ (respectivement $AH_n(\mathbb{C})$).
(ii) Un endomorphisme hermitien $u \in \mathbf{H}(E)$ est *positif* (respectivement *défini positif*) si, et seulement si, $M \in H_n^{+}(\mathbb{C})$ (resp. $M \in H_n^{++}(\mathbb{C})$).

Proposition 107. Soit E un espace vectoriel hermitien. Pour tout $u \in \mathcal{L}(E)$, les endomorphismes hermitiens positifs u^*u et uu^* ont même rang : celui de u.

Proposition 108. Soit E un espace vectoriel hermitien. L'application qui à tout *endomorphisme* hermitien $u \in \mathbf{H}(E)$ associe la *forme sesquilinéaire* hermitienne ψ définie par $\psi(x, y) = \big(\, u(x) \;\big|\; y \,\big)$ est un *isomorphisme* d'espaces vectoriels *réels* de $\mathbf{H}(E)$ sur l'espace vectoriel réel des formes sesquilinéaires hermitiennes sur E. Si E est de dimension strictement positive, les *matrices* de l'endomorphisme u et de la forme ψ correspondante dans une *base orthonormée* sont les *mêmes*. Cette matrice commune reste *semblable* à elle même quand on change de base orthonormée.
De plus l'endomorphisme u est *positif* (resp. *défini positif*) si, et seulement si, la forme sesquilinéaire hermitienne ψ est *positive* (resp. *définie positive*).

Le lecteur retiendra en particulier que *les matrices d'une forme hermitienne donnée sur un espace vectoriel hermitien dans les bases orthonormées de cet espace sont toutes semblables*. En particulier, elles ont toutes le *même spectre*.

3.4 Réduction d'un endomorphisme normal, applications

3.4.1 Réduction d'un endomorphisme normal

Définition 20. Soit E un espace préhilbertien complexe. Un endomorphisme u de E est dit *normal* s'il admet un adjoint u^* et s'il commute avec celui-ci : $uu^* = u^*u$.

Exemples. Les endomorphismes *hermitiens, antihermitiens, unitaires* d'un espace préhilbertien complexe sont des endomorphismes normaux.

Proposition 109. Soit u un endomorphisme normal d'un espace préhilbertien complexe E :
(i) pour tout $x \in E$, $\|u(x)\| = \|u^*(x)\|$;
(ii) pour tout $(\lambda, x) \in \mathbb{C} \times E$, $\|u(x) - \lambda x\| = \|u^*(x) - \bar{\lambda}x\|$;
(iii) si $\lambda \in \mathbb{C}$ est une valeur propre de u, $E_\lambda(u) = E_{\bar{\lambda}}(u^*)$;
(iv) si λ et μ sont deux valeurs propres *distinctes* de u, les sous-espaces propres $E_\lambda(u)$ et $E_\mu(u)$ sont *orthogonaux*.

Démonstration.
(i) et (ii) Pour tout $x \in E$:

$$\begin{aligned}\|u(x)\|^2 = \big(u(x) \mid u(x)\big) &= \big(x \mid u^*u(x)\big)\\ &= \big(x \mid uu^*(x)\big) = \big(u^*(x) \mid u^*(x)\big) = \|u^*(x)\|^2.\end{aligned}$$

De (i) on déduit (ii) en posant $v = u - \lambda \operatorname{Id}$ ($v^* = u^* - \bar{\lambda}\operatorname{Id}$).
(iii) Soient λ une valeur propre de u et x un vecteur propre associé : $u(x) - \lambda x = 0$, donc $\|u(x) - \lambda x\| = \|u^*(x) - \bar{\lambda}x\| = 0$ et $\operatorname{Ker}\big(u - \lambda \operatorname{Id}\big) \subset \operatorname{Ker}\big(u^* - \bar{\lambda}\operatorname{Id}\big)$.
De $u = (u^*)^*$, on déduit l'inclusion opposée $\operatorname{Ker}\big(u^* - \bar{\lambda}\operatorname{Id}\big) \subset \operatorname{Ker}\big(u - \lambda \operatorname{Id}\big)$.
(iv) Soient $x \in E_\lambda(u)$ et $y \in E_\mu(u)$.
On a $\big(x \mid u(y)\big) = \mu(x \mid y)$ et $\big(u^*(x) \mid y\big) = (\overline{\lambda}x \mid y) = \lambda(x \mid y)$.
De $\big(x \mid u(y)\big) = \big(u^*(x) \mid y\big)$, on déduit alors : $(\mu - \lambda)(x \mid y) = 0$, puis $(x \mid y) = 0$, puisque $\lambda \neq \mu$. ∎

Corollaire 110. Soit E un espace vectoriel hermitien.
(i) Les valeurs propres de tout endomorphisme *hermitien* de E sont *réelles*.
(ii) Les valeurs propres de tout endomorphisme *antihermitien* de E sont *imaginaires pures*.
(iii) Les valeurs propres de tout automorphisme *unitaire* de E sont *de module* 1.

Théorème 111 (Diagonalisation des endomorphismes normaux).
Soit E un espace vectoriel hermitien de dimension $n > 0$. Tout endomorphisme *normal* de E est *diagonalisable* et ses sous-espaces propres sont deux à deux *orthogonaux* :

$$E = \overset{\perp}{\bigoplus_{\lambda \in \operatorname{sp} u}} E_\lambda(u).$$

Démonstration. D'après la proposition précédente les sous-espaces propres sont en somme directe orthogonale : $F = \overset{\perp}{\bigoplus_{\lambda \in \operatorname{sp} u}} E_\lambda(u) = \overset{\perp}{\bigoplus_{\lambda \in \operatorname{sp} u}} E_{\overline{\lambda}}(u^*)$. Ainsi, F est u^* stable, donc $F^\perp$ est u-stable (*cf.* proposition 92 de la page 199). Si l'on avait $F \neq E$, $u_{|F}$ admettrait au moins une droite propre, ce qui est impossible, donc $E = F$. ∎

Un endomorphisme hermitien étant normal, on déduit du théorème 111 et du corollaire 110 de la page précédente l'important résultat suivant.

Théorème 112 (Théorème spectral pour les endomorphismes hermitiens d'un espace hermitien). Soit E un espace vectoriel hermitien de dimension $n > 0$. Tout endomorphisme *hermitien* de E est *diagonalisable, à spectre réel*, et ses sous-espaces propres sont deux à deux *orthogonaux* :

$$E = \overset{\perp}{\bigoplus_{\lambda \in \operatorname{sp} u}} E_\lambda(u).$$

Interprétation matricielle

Définition 21.
Une matrice $A \in M_n(\mathbb{C})$ est dite *normale* si, et seulement si, $AA^* = A^*A$.

Soit $\mathcal{B}$ une base orthonormée de l'espace vectoriel hermitien E, un endomorphisme de E est normal si, et seulement si, la matrice $\operatorname{Mat}_{\mathcal{B}}(u)$ est normale.

Théorème 113.
Soit $A \in M_n(\mathbb{C})$ une matrice normale. Il existe un couple (D, P) formé d'une matrice *diagonale* D et d'une matrice *unitaire* P tel que $A = PDP^{-1} = PDP^*$.

Si la matrice A est hermitienne (respectivement antihermitienne,) les éléments de la matrice diagonale D sont réels (respectivement imaginaires purs), si elle est unitaire, ils sont de module 1.

> ***Remarque importante.*** On peut appliquer ces résultats aux matrices réelles symétriques, antisymétriques, orthogonales (en tant que matrices complexes elles sont respectivement hermitiennes, antihermitiennes, unitaires).

Exercice 17.
Soit $A \in A_n(\mathbb{R})$ une matrice *antisymétrique* réelle.
Montrer qu'il existe une matrice orthogonale $P \in O_n(\mathbb{R})$ telle que PAP^{-1} soit une matrice diagonale par blocs de la forme $B = \mathrm{Diag}(0, \dots, 0, B_1, \dots, B_p)$, les matrices $B_k \in M_2(\mathbb{R})$ ($k \in [\![1, p]\!]$) étant de la forme $\begin{pmatrix} 0 & \alpha_k \\ -\alpha_k & 0 \end{pmatrix}$, où $\alpha_k \neq 0$.

Solution. Soit $A \in A_n(\mathbb{R})$, considérée comme matrice de $M_n(\mathbb{C})$ elle est *antihermitienne*. Si $\lambda = i\alpha$ ($\alpha \in \mathbb{R}^*$) est une valeur propre de A, $\bar{\lambda} = -i\alpha$ est aussi une valeur propre de A de même multiplicité (car le polynôme caractéristique de A est *à coefficients réels*). La conjugaison dans $\mathbb{C}^n$ induit un isomorphisme entre $E_\lambda(A)$ et $E_{\bar{\lambda}}(A)$. Si $(V_1, \dots, V_k)$ est une base orthonormée de $E_\lambda(A)$, $(\bar{V}_1, \dots, \bar{V}_k)$ est une base orthonormée de $E_{\bar{\lambda}}(A)$. La réunion de ces deux bases est une base du $\mathbb{C}$-espace vectoriel $E_\lambda(A) \overset{\perp}{\oplus} E_{\bar{\lambda}}(A)$ ($\lambda \neq \bar{\lambda}$). Pour tout $h \in [\![1, k]\!]$ définissons $W_h = \frac{1}{\sqrt{2}}(V_h + \overline{V}_h)$ et $W'_h = \frac{1}{i\sqrt{2}}(V_h - \overline{V}_h)$; on a : $A(W_h) = -\alpha W'_h$ et $A(W'_h) = \alpha W_h$. La famille $(W_1, W'_1, \dots, W_k, W'_k)$ est une famille *orthonormée* de $\mathbb{R}^n$, elle engendre un sous-espace vectoriel F_α de dimension $2k$ de $\mathbb{R}^n$.
Si les valeurs propres non nulles de A sont $\pm i\alpha_1, \dots, \pm i\alpha_p$, les réels non nuls $\alpha_1, \dots, \alpha_p$ étant deux à deux distincts, on a : $\mathbb{R}^n = \mathrm{Ker}\, A \overset{\perp}{\oplus} F_{\alpha_1} \overset{\perp}{\oplus} \dots \overset{\perp}{\oplus} F_{\alpha_p}$, le résultat s'en déduit.

Remarque. Il existe un résultat similaire pour les matrices *orthogonales* (*cf.* le module I.6).

3.4.2 Applications de la réduction d'un endomorphisme hermitien

Les démonstrations des résultats suivants sont analogues à celles des cas réels parallèles.

Signature d'une forme hermitienne

Proposition 114. Soit E un espace vectoriel complexe de dimension $n > 0$.
Soient Φ une forme hermitienne sur E, de signature (p, q), $\mathcal{B}$ une base de E et $A = \mathrm{Mat}(\Phi; \mathcal{B})$, alors p est égal au nombre de racines strictement positives du polynôme caractéristique de A et q est égal au nombre de racines strictement négatives de ce polynôme.

Problèmes d'extremum

Proposition 115. Soient E un espace vectoriel hermitien de dimension strictement positive et u un endomorphisme hermitien de E, on a :

$$\sup_{\|x\|=1} \big(u(x) \mid x\big) = \sup_{x \in E\setminus\{0\}} \frac{\big(u(x) \mid x\big)}{\|x\|^2} = \sup\big(\mathrm{sp}(u)\big),$$

$$\inf_{\|x\|=1} \big(u(x) \mid x\big) = \inf_{x \in E\setminus\{0\}} \frac{\big(u(x) \mid x\big)}{\|x\|^2} = \inf\big(\mathrm{sp}(u)\big).$$

De plus $u \in \mathbf{H}^+(E)$ (resp. $u \in \mathbf{H}^{++}(E)$) si, et seulement si, toutes les valeurs propres de u sont positives (resp. strictement positives).

Corollaire 116. Soit $A \in H_n(\mathbb{C})$:
(i) On a $\sup\limits_{\substack{x\in\mathbb{C}^n\\ x\neq 0}} \dfrac{X^*AX}{X^*X} = \sup\big(\mathrm{sp}(A)\big)$ et $\inf\limits_{\substack{x\in\mathbb{C}^n\\ x\neq 0}} \dfrac{X^*AX}{X^*X} = \inf\big(\mathrm{sp}(A)\big)$
(ii) On a $A \in H_n^+(\mathbb{C})$ (resp. $A \in H_n^{++}(\mathbb{C})$) si, et seulement si, toutes les valeurs propres de A sont positives (resp. strictement positives).

Corollaire 117. Soient Φ une forme hermitienne sur l'espace vectoriel hermitien E, $\mathcal{B}$ une base orthonormée de E et $A := \mathrm{Mat}(\Phi;\mathcal{B})$.
On a $\sup\limits_{\|x\|=1} \Phi(x) = \sup\big(\mathrm{sp}(A)\big)$ et $\inf\limits_{\|x\|=1} \Phi(x) = \inf\big(\mathrm{sp}(A)\big)$.

Quotient de Rayleigh. Soit $A \in H_n(\mathbb{C})$, on appelle *quotient de Rayleigh* de A l'application $R_A : \mathbb{C}^n \setminus \{0\} \to \mathbb{C}$ définie par $R_A(x) = \dfrac{X^*AX}{X^*X}$.

Diagonalisation simultanée, racine carrée

La diagonalisation simultanée de deux formes hermitiennes et la racine carrée d'un endomorphisme hermitien positif se traitent comme dans le cas réel.

Exercice type corrigé

I.4.0 Cet exercice est une généralisation de l'exercice 4 de la page 153.

Calculer $\delta = \inf\limits_{(a_1,a_2,\dots,a_n)\in\mathbb{R}^n} \displaystyle\int_0^1 (1 + a_1t + a_2t^2 + \cdots + a_nt^n)^2 \,\mathrm{d}t$

Solution. On munit $\mathbb{R}[X]$ du produit scalaire $(\,P \mid Q\,) = \displaystyle\int_0^1 P(t)Q(t)\,\mathrm{d}t$. Soit :

$$P(t) := a_1t + a_2t^2 + \cdots + a_nt^n,$$

$\displaystyle\int_0^1 (1 + P(t))^2 \,\mathrm{d}t = \|1 + P(t)\|^2 = \|1 - (-P(t))\|^2$, $\sqrt{\delta}$ est la distance de 1 au sous-espace vectoriel $E_n := \mathrm{Vect}(t, t^2, \dots, t^n)$ de $\mathbb{R}[X]$.
Soit p le projecteur orthogonal sur E_n : $p(1) = \sum\limits_{k=1}^{n} \alpha_k t^k$, $(\alpha_1, \dots \alpha_n)$ étant solution du système linéaire non homogène (S) de n équations :

$$\sum_{k=1}^{n} \alpha_k (\,t^k \mid t^h\,) = (\,1 \mid t^h\,), \quad \forall h \in [\![1, n]\!],$$

soit $\displaystyle\sum_{k=1}^{n} \frac{\alpha_k}{k+h+1} = \frac{1}{h+1}$.

Considérons la fraction rationnelle $F(t) := \sum_{k=1}^{n} \frac{\alpha_k}{k+t} - \frac{1}{t}$, $(\alpha_1, \dots \alpha_n)$ est solution de (S) si, et seulement si, $F(2) = \cdots = F(n+1) = 0$. On a $F(t) = \dfrac{P_n(t)}{t(t+1)\dots(t+n)}$ avec $\deg P_n \leqslant n$, d'où $P_n(t) = C_n(t-2)\dots(t-n-1)$. Le coefficient de $\frac{1}{t}$ dans la décomposition en éléments simples de F (ou résidu en 0) est -1, donc :

$$-1 = \frac{P_n(0)}{n!} = \frac{(-1)^n C_n (n+1)!}{n!} \quad \text{et} \quad C_n = \frac{(-1)^{n+1}}{n+1}\cdot$$

Ainsi $\delta = d^2(1, E_n) = (\,1 \mid 1 - p(1)\,) = 1 - \sum_{k=1}^{n} \frac{\alpha_k}{k+1} = -F(1) = \dfrac{1}{(n+1)^2}$ et $d(1, E_n) = \dfrac{1}{n+1}\cdot$

Pour $n = 2$, on retrouve $\sqrt{\delta} = \dfrac{1}{3}\cdot$

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

Exercices sur la section 1

I.4.1 $*$ Soient p et q deux projecteurs orthogonaux d'un espace vectoriel euclidien de dimension $n > 0$. Montrer que les assertions suivantes sont équivalentes :
(i) $pq = qp$;
(ii) il existe une base de E formée de vecteurs propres communs à p et q ;
(iii) pq est un projecteur.

I.4.2 Déterminer une base orthonormée de l'hyperplan de $\mathbb{R}^n$ d'équation $\sum_{i=1}^{n} x_i = 0$.

I.4.3 Soit E un espace vectoriel euclidien, montrer que toute *application* qui conserve le produit scalaire appartient à $\mathcal{O}(E)$.

I.4.4 $*$ Soit E un espace vectoriel euclidien de dimension finie $n > 0$. Soient H un hyperplan de E, s la symétrie orthogonale d'axe H et $g \in \mathcal{O}(E)$, on dira que H est l'hyperplan de s.
a) Montrer que $g \circ s \circ g^{-1}$ est une symétrie orthogonale hyperplane. Quel est son hyperplan ?
b) On rappelle que deux hyperplans sont perpendiculaires si l'un contient la droite orthogonale à l'autre. Soient k hyperplans $H_1, \dots, H_k$, deux à deux perpendiculaires et $v_1, \dots, v_k$ des vecteurs non-nuls tels que, pour tout $i \in [\![1, k]\!]$, v_i est orthogonal à H_i. Montrer que le système $(v_1, \dots, v_k)$ est orthogonal et que $k \leqslant n$.
c) Démontrer que deux symétries orthogonales hyperplanes commutent si, et seulement si, elles sont égales ou si l'hyperplan de l'une est perpendiculaire à l'hyperplan de l'autre.
d) On suppose que $n = 3$. Décrire tous les sous-groupes *commutatifs finis* de $\mathcal{O}(E)$ engendrés par des symétries orthogonales hyperplanes (un tel sous-groupe contient au plus 3 symétries orthogonales hyperplanes distinctes).

I.4.5 Soit E un espace vectoriel euclidien de dimension $n > 0$. Étant donné un entier p strictement supérieur à 1, une famille $(v) = (v_1, \dots, v_p)$ de vecteurs de E est dite *obtuse* si pour tout couple $(i,j) \in [\![1,p]\!]^2$ tel que $i \neq j$, on a $(\, v_i \mid v_j \,) < 0$ (*i.e.* les deux vecteurs forment un angle obtus).

a) Soit $x = \sum_{i=1}^{p} x_i v_i$, montrer que $\| \sum_{i=1}^{p} |x_i| v_i \| \leqslant \|x\|$.

b) On suppose que la famille (v) est obtuse et qu'il existe $a \in E$ tel que, pour tout $i \in [\![1,p]\!]$, on ait $(\, a \mid v_i \,) > 0$. Montrer que (v) est une famille libre.

c) On suppose que la famille (v) est obtuse et liée. Soit $\sum_{i=1}^{p} \lambda_i v_i = 0$ une relation de dépendance. Montrer que, pour tout $i \in [\![1,p]\!]$: $\lambda_i \neq 0$. En déduire que le rang de la famille (v) est $p-1$ et que tous les coefficients de la relation sont de même signe. Quel est le cardinal maximal d'un système obtus de E ?

d) On suppose que l'on a un système (v), formé de $n+1$ vecteurs unitaires, et un réel $\alpha \in]0,1[$, tels que, pour tous i,j tels que $i \neq j$, on ait $(\, v_i \mid v_j \,) = -\alpha$. Déterminer α (en fonction de n) et une relation de dépendance liant $v_1, \dots, v_{n+1}$.

I.4.6 Soit E un espace vectoriel euclidien de dimension $n > 0$.
Soient $a \in \mathbf{S}^{++}(E)$ et $b \in \mathbf{S}^{+}(E)$. Montrer qu'il existe un unique endomorphisme c de E tel que $ac + ca = b$, et vérifier que $c \in \mathbf{S}^{+}(E)$.

I.4.7 Soit $A := \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$. Déterminer $D \in D_n(\mathbb{R})$ et $P \in O_n(\mathbb{R})$ telles que $A = PDP^{-1}$. Même question pour $A := \begin{pmatrix} 2 & -2 & 1 \\ -2 & 4 & -4 \\ 1 & -4 & -2 \end{pmatrix}$.

I.4.8 Soit $n \in \mathbb{N}^*$, montrer que l'application $f : \mathbb{R}^n \to \mathbb{R}$ définie par :

$$f(a_0, \dots, a_n) := \int_0^1 \left(\ln t + \sum_{k=0}^{n} a_k t^k \right)^2 \mathrm{d}t$$

admet un minimum et le calculer (on considérera $F(x) = \dfrac{1}{(x+1)^2} + \sum_k \dfrac{a_k}{x+k+1}$).

I.4.9 Soit $E = \mathcal{C}^2([0,1], \mathbb{R})$. Pour tous $f, g \in E$, on pose :

$$(\, f \mid g \,) := \int_0^1 \big(f(t)g(t) + f'(t)g'(t) \big) \, \mathrm{d}t.$$

a) Montrer que $(f,g) \mapsto (\, f \mid g \,)$ est un produit scalaire sur E.

b) Soient $V := \{ f \in E \mid \ f(0) = f(1) = 0 \}$ et $W := \{ f \in E \mid \ f'' = f \}$.
Montrer que V et W sont deux sous-espaces supplémentaires orthogonaux de E.

c) Soit $F := \{ f \in E \mid \ f(0) = \alpha, \ f(1) = \beta \}$. Calculer $\inf_{f \in F} \int_0^1 \big(f^2(t) + f'^2(t) \big) \, \mathrm{d}t$

I.4.10 $*$ Soit E un espace vectoriel euclidien, soit $A := \{u \in \mathcal{L}(E) |\ u = uu^*u\}$.
a) Montrer que $u \in A$ si, et seulement si, u^*u est un projecteur orthogonal.
b) Montrer que $u \in A$ si, et seulement si, $\big\|u(x)\big\| = \|x\|$, quel que soit $x \in (\operatorname{Ker} u)^\perp$.
c) On suppose que $u \in A$. Montrer que $(\operatorname{Ker} u)^\perp = \{x \in E \mid \ \|x\| = \big\|u(x)\big\|\}$.

I.4.11 Soit $E = \mathbb{R}_n[X]$ muni du produit scalaire $(\,P \mid Q\,) := \displaystyle\int_0^1 P(t)Q(t)\ \mathrm{d}t$. Pour $P \in E$, on définit $u(P)(x) := \displaystyle\int_0^1 (x+t)^n P(t)\ \mathrm{d}t$.
a) Vérifier que u définit un endomorphisme de E, symétrique et bijectif.
b) Soit $(P_0, \ldots, P_n)$ une base orthonormée de vecteurs propres de u $\big(u(P_i) = \lambda_i P_i\big)$. Montrer que $(x+y)^n = \displaystyle\sum_{k=0}^{n} \lambda_k P_k(x) P_k(y)$ et $\displaystyle\int_0^1 (x+t)^{2n}\ \mathrm{d}t = \sum_{k=0}^{n} \lambda_k^2 P_k^2(x)$
c) Calculer $\operatorname{Tr} u$ et $\operatorname{Tr} u^2$.

I.4.12 $*$ Soit $E = \mathbb{R}[X]$, on le munit du produit scalaire $(\,P \mid Q\,) := \displaystyle\int_0^1 P(t)Q(t)\ \mathrm{d}t$.
1) On considère la forme linéaire h définie sur E par $h : P \mapsto P(2)$. Montrer qu'il n'existe pas de polynôme $Q_0 \in E$ tel que, pour tout $P \in E$: $h(P) = (\,Q_0 \mid P\,)$ (la forme linéaire h n'est pas un produit scalaire).
2.a) Montrer que, pour tout $P \in E$, l'intégrale $\displaystyle\int_0^1 \frac{P(t)}{\sqrt{t}}\ \mathrm{d}t$ est convergente. On peut donc définir une forme linéaire sur E par $f(P) = \displaystyle\int_0^1 \frac{P(t)}{\sqrt{t}}\ \mathrm{d}t$. On suppose qu'il existe un polynôme $Q_1 = \displaystyle\sum_{k=0}^{p} \alpha_k t^k \in E$ tel que, pour tout $P \in E$: $f(P) = (\,Q_1 \mid P\,)$, calculer $f(t^n)$ et $(\,Q_1 \mid t^n\,)$, pour tout $n \in \mathbb{N}$, en déduire que les fractions rationnelles $\dfrac{1}{x + \frac{1}{2}}$ et $\displaystyle\sum_{k=0}^{p} \frac{\alpha_k}{x+k+1}$ coïncident. Montrer que l'on aboutit à une contradiction et conclure.
2.b) Soit w l'endomorphisme de E qui associe à un polynôme P le polynôme constant $f(P) \in E$. Montrer que w n'admet pas d'adjoint.
On pourra supposer qu'il en existe un, considérer $w^*(1) \in E$ et utiliser la question 2.a).

I.4.13 $*$ Soit $A \in S_n(\mathbb{R})$, soit $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ la suite ordonnée décroissante des valeurs propres de A. Montrer que, pour tout $k \in [\![1, n]\!]$: $\displaystyle\sum_{j=1}^{k} a_{jj} \leqslant \sum_{j=1}^{k} \lambda_j$.

I.4.14 Soit $A \in S_n^{++}(\mathbb{R})$.
a) Montrer que pour tous $x, y \in \mathbb{R}^n$: $({}^tXAX)({}^tYA^{-1}Y) \geqslant ({}^tXY)^2$.
b) Soient $i \in [\![1, n]\!]$ et A_i la matrice extraite de A obtenue en supprimant la $i^{\text{ème}}$ ligne et la $i^{\text{ème}}$ colonne de A, vérifier que $\det A_i > 0$ et montrer que

$$\frac{\det A}{\det A_i} = \inf\{{}^tXAX \mid \ x \in \mathbb{R}^n \text{ tel que } x_i = 1\}$$

(on pourra utiliser l'inégalité de Cauchy-Schwarz).

I.4.15 $**$ Soit $A = (a_{ij}) \in S_n^+(\mathbb{R})$, on suppose que $a_{ij} \neq 0$, pour tous $i, j \in [\![1, n]\!]$.
Soit $B := (b_{ij} = \frac{1}{a_{ij}})$. Montrer que $B \in S_n^+(\mathbb{R})$ si, et seulement si, $\operatorname{rang} A = 1$.
On pourra étudier le cas $n = 2$.

I.4.16 Sans utiliser les résultats sur les endomorphismes normaux d'un espace vectoriel hermitien montrer que les valeurs propres d'une matrice antisymétrique réelle considérée comme matrice à coefficients complexes sont imaginaires pures.

I.4.17 Soient E un espace vectoriel euclidien et $a, b \in E$ deux vecteurs linéairement indépendants. On définit $u : x \mapsto u(x) := (\, a \mid x \,)\, b$.

(i) Déterminer u^*.

(ii) Soit $f := u + u^*$; quels sont les éléments propres de f ?

(iii) Déterminer les bornes supérieures et inférieures de $\{(\, x \mid f(x) \,) \mid \ x \in E, \|x\| = 1\}$.

I.4.18 Soit $S \in S_n(\mathbb{R})$. Montrer qu'il existe $A \in A_n(\mathbb{R})$ telle que $S = A^2$ si, et seulement si, $\operatorname{sp} S \subset \mathbb{R}^-$, les valeurs propres de S strictement négatives étant de multiplicité paire.

I.4.19 Soit $M := (i + j - 1)_{i,j \in [\![1,n]\!]} \in S_n(\mathbb{R})$.
Déterminer les valeurs propres et les sous-espaces propres de M.

I.4.20 Réduire simultanément les deux formes quadratiques sur $\mathbb{R}^4$:

$$\Phi(x) := x_1^2 + 3x_2^2 + x_3^2 - x_4^2 - 2x_1x_2 - 4x_2x_3 + 2x_3x_4$$
$$\Psi(x) := x_1^2 + 2x_2^2 + 2x_3^2 + 2x_4^2 - 2x_1x_2 - 2x_2x_3 - 2x_3x_4.$$

On montrera d'abord que Ψ est définie positive.

I.4.21 $*$ a) Question préliminaire.
Soient $\lambda_1, \ldots, \lambda_n$ des nombres réels positifs, montrer l'inégalité $1 + (\prod_{i=1}^{n} \lambda_i)^{1/n} \leqslant \prod_{i=1}^{n} (1 + \lambda_i)^{1/n}$.
On pourra utiliser la fonction $t \mapsto \ln(1 + e^t)$.
b) Soient $A, B \in S_n^+(\mathbb{R})$. Montrer que $(\det A)^{1/n} + (\det B)^{1/n} \leqslant (\det(A + B))^{1/n}$.
On pourra utiliser, quand A est inversible, la matrice $A^{-1}B$.

Exercices sur les sections 2 et 3

I.4.22 Soit $A := \begin{pmatrix} 0 & -2 & -1 \\ 2 & 0 & 2 \\ 1 & -2 & 0 \end{pmatrix}$.
Déterminer $D \in D_n(\mathbb{C})$ et $P \in \mathcal{U}_n(\mathbb{C})$ telles que $A = PDP^{-1}$

I.4.23 $*$ Soit E un espace vectoriel normé sur $\mathbb{R}$. On suppose que la norme de E vérifie, pour tous $x, y \in E$, *l'égalité de la médiane* $\|x+y\|^2 + \|x-y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right)$.
On se propose de montrer qu'il existe une unique forme bilinéaire symétrique $(x,y) \mapsto \varphi(x,y)$ sur E telle que, pour tout $x \in E$: $\varphi(x,x) = \|x\|^2$.
a) On définit $\varphi(x,y) := \|\frac{x+y}{2}\|^2 - \|\frac{x-y}{2}\|^2$. Montrer :

$$\forall x, y, x_1, x_2 \in E\,,\ \varphi(x,y) = \varphi(y,x) \quad \text{et} \quad \varphi(x_1+x_2, y) = \varphi(x_1,y) + \varphi(x_2,y).$$

b) Montrer que, pour tous $\alpha \in \mathbb{R}$ et $x, y \in E$: $\varphi(\alpha x, y) = \alpha\varphi(x,y)$ (on supposera d'abord que $\alpha \in \mathbb{N}$, puis $\alpha \in \mathbb{Q}$; on terminera par un argument de continuité). Conclure.
c) On suppose maintenant que E est un espace vectoriel normé sur $\mathbb{C}$, montrer un résultat analogue au précédent.

I.4.24 Minimax d'une forme hermitienne (théorème de Courant-Fischer).
Soient $A := (a_{hk}) \in H_n(\mathbb{C})$ et $(\varepsilon_1, \ldots, \varepsilon_n)$ une base orthonormée de $\mathbb{C}^n$ formée de vecteurs propres de A ($A(\varepsilon_h) = \lambda_h \varepsilon_h$, pour tout $h \in [\![1,n]\!]$). On suppose que : $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$.
a) Soient $k \in \mathbb{N}^*$ et $\mathcal{V}_k$ l'ensemble des sous–espaces vectoriels de dimension k de $\mathbb{C}^n$; montrer que, pour tout $V \in \mathcal{V}_k$: $V \cap \mathrm{Vect}(\varepsilon_k, \varepsilon_{k+1}, \ldots, \varepsilon_n) \neq \{0\}$, en déduire : $\inf\limits_{V \in \mathcal{V}_k} \left(\sup\limits_{\substack{x \in V \\ x \neq 0}} \frac{X^* A X}{X^* X} \right) = \lambda_k$.
b) Soit $p \in [\![1,n]\!]$; on note $\mathcal{F}_p$ l'ensemble des sous–espaces vectoriels de $\mathbb{C}^n$, de codimension *strictement inférieure* à p. Montrer que pour tout $F \in \mathcal{F}_p$: $F \cap \mathrm{Vect}(\varepsilon_1, \ldots, \varepsilon_p) \neq \{0\}$.
En déduire : $\sup\limits_{F \in \mathcal{F}_p} \left(\inf\limits_{\substack{x \in F \\ x \neq 0}} \frac{X^* A X}{X^* X} \right) = \lambda_p$.
c) Application.
Soit Φ la forme hermitienne sur $\mathbb{C}^n$, dont la matrice associée dans la base canonique de $\mathbb{C}^n$ est A. Soient $r \in [\![0, n-1]\!]$, G un sous–espace vectoriel de dimension $n-r$ de $\mathbb{C}^n$, et A_r la matrice associée à la forme $\Phi_r = \Phi_{|G}$ dans une base orthonormée de G. Soient $\mu_1 \leqslant \cdots \leqslant \mu_{n-r}$ les valeurs propres de A_r comptées, si nécessaire, avec leurs multiplicités. Montrer que, pour tout $p \in [\![1, n-r]\!]$: $\lambda_p \leqslant \mu_p \leqslant \lambda_{p+r}$.
En déduire que, pour tout $q \in [\![1,n]\!]$: $\lambda_1 + \cdots + \lambda_q \leqslant a_{11} + \cdots + a_{qq}$.

I.4.25 a) Soit A une matrice antisymétrique réelle. Montrer que la relation $(I+B)(I+A) = 2I$ définit une matrice B unique. Vérifier que B est orthogonale et n'admet pas la valeur propre -1.
b) Soit B une matrice orthogonale n'admettant pas la valeur propre -1. Montrer que la relation $(I+B)(I+A) = 2I$ définit une matrice A unique et vérifier que A est antisymétrique.

I.4.26 a) Soient $A \in H_n(\mathbb{C})$ une matrice hermitienne, non nulle, et $\lambda \in \mathbb{C}$, non réel. Montrer que la matrice $B(\lambda) := (A + \lambda I)(A - \lambda I)^{-1}$ est unitaire si, et seulement si, $\lambda \in \mathbb{R}^* i$. Dans ce cas 1 n'est pas valeur propre de $B(\lambda)$.
b) Inversement montre que si B est unitaire et n'admet pas 1 pour valeur propre et si $\lambda \in i\mathbb{R}^*$, il existe une matrice A hermitienne unique telle que $B = (A + \lambda I)(A - \lambda I)^{-1}$

I.4.27 Soit u un endomorphisme d'un espace hermitien.
Montrer que u est normal si, et seulement s'il existe un polynôme $P \in \mathbb{C}[X]$ tel que $u^* = P(u)$.

Réduction des matrices

I.5

Dans tout ce module, K désigne un corps commutatif, E un K-espace vectoriel non nul, u un endomorphisme de E, n un entier strictement positif et A un élément de $M_n(K)$. Sauf mention explicite du contraire, E est supposé de dimension finie égale à n.

Nous abordons une étude, plus approfondie que dans [L1], des endomorphismes de E et des matrices carrées à coefficients dans K. Cette étude est motivée entre autres par ses applications (systèmes différentiels linéaires à coefficients constants, calcul de l'exponentielle d'une matrice carrée, suites récurrentes, ...).

Réduire u, c'est trouver une base $\mathcal{B}$ de E dans laquelle la matrice $\mathrm{Mat}_{\mathcal{B}}(u)$ de u soit aussi simple que possible (par exemple diagonale, triangulaire, ...), notamment afin de simplifier les calculs. *Réduire* A c'est trouver une matrice *semblable* à A et aussi simple que possible. Rappelons que deux matrices $A, B \in M_n(K)$ sont dites semblables s'il existe $P \in GL_n(K)$ telle que $B = PAP^{-1}$. La réduction des endomorphismes de E et la réduction des matrices carrées d'ordre n sont deux problèmes essentiellement équivalents.

Ce module fait appel à la partie de [L1] concernant l'arithmétique des polynômes. Les notions de K-algèbre et de polynôme minimal d'un élément d'une K-algèbre (*cf.* le module I.1) seront importantes, à cause des algèbres $K[X]$, $\mathcal{L}(E)$ et $M_n(K)$. Enfin la notion de sous-espace stable par un endomorphisme (*cf.* la section 2.1) va s'avérer cruciale.

1 Rappels et compléments

1.1 Sommes directes de sous-espaces vectoriels

Étant donnés deux sous-espaces vectoriels V, W d'un espace vectoriel E, $V + W$ est le sous-espace vectoriel de E engendré par $V \cup W$; il est formé des sommes $v + w$, où (v, w) décrit $V \times W$. D'après [L1], la somme $V + W$ est *directe* si, et seulement si, $V \cap W = \{0\}$, *i.e.* si, et seulement si, tout élément de $V + W$ s'écrit *de façon unique* $v + w$, où $(v, w) \in V \times W$, ou encore si, et seulement si, l'application linéaire $(v, w) \mapsto v + w$ de $V \times W$ dans E est injective. Cette notion de somme directe s'étend au cas de plus de deux sous-espaces vectoriels :

Proposition et définition 1. Soient $V_1, \dots, V_p$ des sous-espaces vectoriels d'un espace vectoriel E et V leur somme. Les conditions suivantes sont équivalentes :

1. Tout $x \in V$ s'écrit de façon *unique* $x = \sum_{i=1}^{p} v_i$, où $(v_1, \dots, v_p) \in V_1 \times \cdots \times V_p$.
2. L'application $f : (v_1, \dots, v_p) \mapsto v_1 + \cdots + v_p$ de $V_1 \times \cdots \times V_p$ dans E est injective.
3. Pour tout $i \in [\![1, p]\!]$, l'intersection $V_i \cap \sum_{j \neq i} V_j$ est égale à $\{0\}$.

Lorsque ces conditions sont vérifiées, on dit que la somme des V_i est *directe*, et l'on écrit $V = V_1 \oplus V_2 \oplus \cdots \oplus V_p$ ou $V = \bigoplus_{i=1}^{p} V_i$.

Démonstration. L'image de l'application linéaire f est V, d'où l'équivalence des conditions 1 et 2.

Prouvons que 2 $\Rightarrow$ 3. Soient $i \in [\![1, p]\!]$ et $a \in V_i \cap \sum_{j \neq i} V_j$; a peut s'écrire $a = \sum_{j=1,\, j \neq i}^{p} x_j$, où $x_j \in V_j$, pour tout $j \neq i$. Posons $x_i := -a$. Alors $x := (x_1, \dots, x_p) \in \operatorname{Ker} f$, donc $x = 0$ et en particulier $a = 0$.

Prouvons que 3 $\Rightarrow$ 2. Soit $x := (x_1, \dots, x_p) \in \operatorname{Ker} f$, de sorte que $\sum_{i=1}^{p} x_i = 0$.

Soit $i \in [\![1, p]\!]$. Puisque $x_i = -\sum_{j \neq i} x_j$, x_i appartient à $V_i \cap \sum_{j \neq i} V_j$, donc $x_i = 0$. C'est vrai pour tout i, donc $x = 0$. ∎

Si la condition 3 est satisfaite, on a $V_i \cap V_j = \{0\}$ dès que $i \neq j$, mais la réciproque est fausse. Soient en effet D, D', D'' trois droites vectorielles de $\mathbb{R}^2$ distinctes. On a $D \cap D' = D' \cap D'' = D'' \cap D = \{0\}$, mais $D \cap (D' + D'') = D \cap \mathbb{R}^2 = D \neq \{0\}$.

Exemple. Si $(e_1, \dots, e_n)$ est une base d'un espace vectoriel E, l'espace E est somme directe des droites Ke_i, $i = 1, \dots, n$ (il est instructif de le vérifier).

Supposons maintenant que l'espace vectoriel E soit somme directe de ses sous-espaces vectoriels $V_1, \dots, V_p$:

$$E = V_1 \oplus V_2 \oplus \cdots \oplus V_p. \tag{1}$$

Cette « égalité » signifie deux choses : d'une part la somme des V_i est directe, et d'autre part cette somme est égale à E. Tout $x \in E$ s'écrivant de façon *unique* $x = \sum_{i=1}^{p} v_i$, où $(v_1, \dots, v_p) \in V_1 \times V_2 \cdots \times V_p$, nous pouvons poser $\pi_i(x) := v_i$ pour $i = 1, \dots, p$. On définit ainsi des endomorphismes $\pi_1, \dots, \pi_r$ de E, appelés *projecteurs de E associés*

à la décomposition en somme directe (1). Les égalités suivantes sont immédiates :

$$\begin{cases} \pi_1 + \pi_2 + \cdots + \pi_p = \mathrm{Id}_E, \\ \pi_i \circ \pi_j = 0 \ \text{ dès que } i \neq j, \\ \pi_i^2 = \pi_i \ \text{ pour tout } i. \end{cases} \tag{2}$$

Parmi ces égalités, les dernières résultent d'ailleurs des précédentes. Par exemple :

$$\pi_1 = \pi_1 \circ \mathrm{Id}_E = \pi_1 \circ (\pi_1 + \pi_2 + \cdots + \pi_p) = \pi_1^2 + \pi_1 \circ \pi_2 + \cdots + \pi_1 \circ \pi_p = \pi_1^2.$$

Notons aussi pour tout i les égalités suivantes, qui résultent des définitions :

$$\mathrm{Im}(\pi_i) \ = \ V_i \quad \text{et} \quad \mathrm{Ker}(\pi_i) \ = \ \sum_{j \neq i} V_j. \tag{3}$$

Exercice 1.
Soient inversement $\pi_1, \ldots, \pi_p \in \mathcal{L}(E)$. On suppose que les π_i vérifient les égalités (2). Pour tout i, posons $V_i := \mathrm{Im}(\pi_i)$. Montrer que E est somme directe de $V_1, \ldots, V_p$ et que les π_i sont les projecteurs associés à cette décomposition de E en somme directe.

Solution. Soit $x \in E$. On a : $x = \mathrm{Id}_E(x) = \pi_1(x) + \pi_2(x) + \cdots + \pi_p(x) \in V_1 + V_2 + \cdots + V_p$, ce qui montre que $V_1 + V_2 + \cdots + V_p = E$. Montrons que la somme des V_i est directe. Soient $i \in [\![1, p]\!]$ et $a \in V_i \cap \sum_{j \neq i} V_j$. Pour tout $j \neq i$, l'égalité $\pi_i \circ \pi_j = 0$ implique $V_j = \mathrm{Im}(\pi_j) \subset \mathrm{Ker}(\pi_i)$. Ainsi $a \in \sum_{j \neq i} V_j \subset \mathrm{Ker}(\pi_i)$. Or $a \in V_i = \mathrm{Im}(\pi_i)$, donc il existe $b \in E$ tel que $a = \pi_i(b)$, d'où $0 = \pi_i(a) = \pi_i^2(b) = \pi_i(b) = a$.

Soit $x \in E$. Pour tout i, $\pi_i(x) \in V_i$. Puisque $x = \mathrm{Id}_E(x) = \sum_{i=1}^{p} \pi_i(x)$, les π_i sont bien les projecteurs associés à la décomposition de E comme somme directe des V_i.

Proposition 2. Partons d'une décomposition de E en somme directe, comme en (1). Pour tout $i \in [\![1, p]\!]$, soit $\mathcal{B}_i$ une base de V_i. Les $\mathcal{B}_i$ sont deux à deux disjointes, et leur réunion $\mathcal{B}$ est une base de E, dite *adaptée* à la décomposition (1). Réciproquement, soient $\mathcal{B} := (e_1, \ldots, e_n)$ une base de E et $(I_1, \ldots, I_m)$ une partition de $[\![1, n]\!]$. Alors E est somme directe des sous-espaces $V_i := \mathrm{Vect}\big(e_h \mid h \in I_i\big)$, $1 \leqslant i \leqslant m$.

Démonstration. Le cas $p := 2$ a été traité dans [L1]. On raisonne alors par récurrence sur p. Pour la première assertion, on observe que $E = (V_1 \oplus V_2 \oplus \cdots \oplus V_{p-1}) \oplus V_p$ (le faire en exercice). ∎

> ***Remarque.*** Soit $(V_i)_{i \in I}$ une famille de sous-espaces vectoriels de l'espace vectoriel E, l'ensemble d'indices I étant quelconque (fini ou infini). La somme $V := \sum_{i \in I} V_i$ est le sous-espace vectoriel de E engendré par la réunion des V_i, sous-espace formé des sommes $\sum_{i \in I} x_i$, où $(x_i)_{i \in I}$ est une famille à support fini et $x_i \in V_i$ pour tout i.

Nous dirons que la somme des V_i est *directe* si une telle somme $\sum_{i\in I} x_i$ n'est nulle que si chaque x_i est nul. Cela équivaut évidemment à dire que, pour toute partie *finie* J de I, la somme $\sum_{i\in J} V_i$ est directe, ou encore, vu la proposition 1 de la page 214, que $V_i \cap \sum_{j\in I,\, j\neq i} V_j = \{0\}$ pour tout i. On écrit alors : $V = \bigoplus_{i\in I} V_i$.

Exercice 2.

On suppose que $V_i \neq \{0\}$ pour tout i. Montrer que la somme des V_i est directe si, et seulement si, toute famille $(x_i)_{i\in I}$ telle que $x_i \in V_i$ et $x_i \neq 0$ pour tout i est libre.

Solution. La remarque permet de supposer que I est fini, et même que $I := [\![1, p]\!]$, avec $p \in \mathbb{N}^*$. Supposons que la somme des V_i soit directe, et soit $(x_i)_{i\in I}$ comme dans l'énoncé. Soit $(\lambda_i)_{i\in I}$ une famille de scalaires telle que $\sum_{i=1}^{p} \lambda_i x_i = 0$. Pour tout i, $\lambda_i x_i \in V_i$. Vu l'égalité $\sum_{i=1}^{p} \lambda_i x_i = 0$, la proposition 1 de la page 214 montre que $\lambda_i x_i = 0$, c'est-à-dire $\lambda_i = 0$, et cela pour tout i. La famille $(x_i)_{i\in I}$ est donc libre.
Inversement, supposons que la condition de l'énoncé soit satisfaite et considérons une famille $(x_1, \ldots, x_p) \in V_1 \times \cdots \times V_p$ telle que $\sum_{i=1}^{p} x_i = 0$. Il s'agit de montrer que tous les x_i sont nuls (*cf.* l'assertion 2 de la proposition 1 de la page 214). Soit $J := \{i \in [\![1, p]\!] \mid x_i \neq 0\}$. Si $i \in J$, posons $y_i := x_i$ et $\lambda_i := 1$. Si $i \in [\![1, p]\!] \setminus J$, posons $\lambda_i := 0$, et soit y_i un vecteur non nul arbitraire de V_i. Alors $\sum_{i=1}^{p} \lambda_i y_i = 0$. Par hypothèse, la famille $(y_1, \ldots, y_p)$ est libre, donc $\lambda_i = 0$ pour tout i. Autrement dit J est vide, *i.e.* les x_i sont tous nuls.

Dans le cas de la dimension finie, on dispose d'une importante condition nécessaire et suffisante pour qu'une somme de sous-espaces vectoriels soit directe :

Proposition 3. Soient E un espace vectoriel de dimension finie et $V_1, \ldots, V_p$ des sous-espaces vectoriels de E. On a toujours :

$$\dim(V_1 + V_2 + \cdots + V_p) \leqslant \dim(V_1) + \dim(V_2) + \cdots + \dim(V_p).$$

Cette inégalité est une égalité si, et seulement si, la somme des V_i est directe.

Démonstration. Posons $F := V_1 \times V_2 \times \cdots \times V_p$ et $V := V_1 + V_2 + \cdots + V_p$. Soit f l'application linéaire $(x_1, x_2, \ldots, x_p) \mapsto x_1 + x_2 + \cdots + x_p$ de F dans E. L'image de f étant V, le théorème du rang donne $\dim(V) = \dim(F) - \dim(\operatorname{Ker} f)$, soit :

$$\dim(V_1 + V_2 + \cdots + V_p) = \dim(V_1) + \dim(V_2) + \cdots + \dim(V_p) - \dim(\operatorname{Ker} f).$$

D'où l'inégalité annoncée. Il y a égalité si, et seulement si, f est injective, *i.e.* si, et seulement si, la somme des V_i est directe, vu la proposition 1 de la page 214.

1.2 Calculs matriciels par blocs

Soient $m, n \in \mathbb{N}^*$. Découpons $[\![1, m]\!]$ en intervalles $I_1, \ldots, I_p$ *consécutifs* : pour tous $h, k \in [\![1, p]\!]$ tels que $h < k$ et tout couple $(i, j) \in I_h \times I_k$, on a $i < j$. Découpons de même $[\![1, n]\!]$ en intervalles $J_1, \ldots, J_q$ consécutifs. Si $h \in [\![1, p]\!]$, posons $m_h := \operatorname{card} I_h$; si $k \in [\![1, q]\!]$, posons $n_k := \operatorname{card} J_k$. Soit ensuite $A = (a_{i,j}) \in M_{m,n}(K)$. Pour tout $(h, k) \in [\![1, p]\!] \times [\![1, q]\!]$, notons $A_{h,k} \in M_{m_h,n_k}(K)$ la matrice ayant pour coefficients les $a_{i,j}$, $(i, j) \in I_h \times J_k$. Il est évidemment équivalent de connaître A ou de connaître toutes les matrices $A_{h,k}$, appelées *blocs*. On convient d'écrire $A = \begin{pmatrix} A_{1,1} & A_{1,2} & \ldots & A_{1,q} \\ A_{2,1} & A_{2,2} & \ldots & A_{2,q} \\ \ldots & \ldots & \ldots & \ldots \\ A_{p,1} & A_{p,2} & \ldots & A_{p,q} \end{pmatrix}$ (« écriture de A par blocs »). Prenons par exemple $m := n := 5$, $p := q := 2$, $I_1 := J_1 := \{1, 2\}$ et $I_2 := J_2 := \{3, 4, 5\}$. Si $A := \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & a_{1,5} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,5} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & a_{4,5} \\ a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & a_{5,5} \end{pmatrix} \in M_5(K)$, on écrit $A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$, où $A_{1,1} := \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$, $A_{1,2} := \begin{pmatrix} a_{1,3} & a_{1,4} & a_{1,5} \\ a_{2,3} & a_{2,4} & a_{2,5} \end{pmatrix}$, $A_{2,1} := \begin{pmatrix} a_{3,1} & a_{3,2} \\ a_{4,1} & a_{4,2} \\ a_{5,1} & a_{5,2} \end{pmatrix}$, $A_{2,2} := \begin{pmatrix} a_{3,3} & a_{3,4} & a_{3,5} \\ a_{4,3} & a_{4,4} & a_{4,5} \\ a_{5,3} & a_{5,4} & a_{5,5} \end{pmatrix}$. Dans le cas général, si $u \in [\![1, m_h]\!]$ et $v \in [\![1, n_k]\!]$, le coefficient d'indices (u, v) de $A_{h,k}$ est le coefficient de A ayant pour indices $(m_1 + \cdots + m_{h-1} + u,\ n_1 + \cdots + n_{k-1} + v)$.

Les opérations matricielles usuelles, addition et multiplication, peuvent être effectuées « par blocs », sous certaines conditions. Pour l'addition, si :

$$A := \begin{pmatrix} A_{1,1} & A_{1,2} & \ldots & A_{1,q} \\ A_{2,1} & A_{2,2} & \ldots & A_{2,q} \\ \ldots & \ldots & \ldots & \ldots \\ A_{p,1} & A_{p,2} & \ldots & A_{p,q} \end{pmatrix} \quad \text{et} \quad B := \begin{pmatrix} B_{1,1} & B_{1,2} & \ldots & B_{1,q} \\ B_{2,1} & B_{2,2} & \ldots & B_{2,q} \\ \ldots & \ldots & \ldots & \ldots \\ B_{p,1} & B_{p,2} & \ldots & B_{p,q} \end{pmatrix}$$

sont écrites par blocs de mêmes « formats », on a :

$$A + B = \begin{pmatrix} A_{1,1} + B_{1,1} & A_{1,2} + B_{1,2} & \ldots & A_{1,q} + B_{1,q} \\ A_{2,1} + B_{2,1} & A_{2,2} + B_{2,2} & \ldots & A_{2,q} + B_{2,q} \\ \ldots & \ldots & \ldots & \ldots \\ A_{p,1} + B_{p,1} & A_{p,2} + B_{p,2} & \ldots & A_{p,q} + B_{p,q} \end{pmatrix},$$

La preuve est évidente. Pour le produit « par blocs », soit $r \in \mathbb{N}^*$. Découpons $[\![1, r]\!]$ en intervalles $H_1, \ldots, H_s$ *consécutifs*. On écrit une matrice $A \in M_{m,n}(K)$ par blocs, selon le découpage $(I_1, \ldots, I_p)$ pour les lignes et le découpage $(J_1, \ldots, J_q)$ pour les colonnes. De même, on écrit une matrice $B \in M_{n,r}(K)$ par blocs, selon le découpage $(J_1, \ldots, J_q)$ pour les lignes et le découpage $(H_1, \ldots, H_s)$ pour les colonnes. Pour tout l, on pose $t_l := \operatorname{card}(H_l)$. Écrivons la matrice $C := AB \in M_{m,r}$ par blocs, selon

le découpage $(I_1, \dots, I_p)$ pour les lignes et le découpage $(H_1, \dots, H_s)$ pour les colonnes. Avec des notations évidentes on a, pour tout $(h,l) \in [\![1,p]\!] \times [\![1,r]\!]$, la formule « naturelle » :

$$C_{h,l} = A_{h,1}B_{1,l} + A_{h,2}B_{2,l} + \cdots + A_{h,q}B_{q,l}. \tag{4}$$

La preuve de cette formule nécessite de bien manier les indices.

Posons $\alpha := m_1 + \cdots + m_{h-1}$ et $\gamma := t_1 + \cdots + t_{l-1}$. Soit $(u,w) \in [\![1,m_h]\!] \times [\![1,t_l]\!]$; notons ξ (resp. η) le coefficient d'indices (u,w) du premier (resp. second) membre de l'égalité (4). Puisque $(J_1, \dots, J_q)$ est une partition de $[\![1,n]\!]$, on a, avec des notations évidentes :

$$\xi = c_{\alpha+u,\gamma+w} = \sum_{j=1}^{n} a_{\alpha+u,j}b_{j,\gamma+w} = \sum_{k=1}^{q}\left(\sum_{j\in J_k} a_{\alpha+u,j}b_{j,\gamma+w}\right)$$

$$= \sum_{k=1}^{q}\left(\sum_{v=1}^{n_k}(A_{h,k})_{u,v}(B_{k,l})_{v,w}\right) = \sum_{k=1}^{q}(A_{h,k}B_{k,l})_{u,w} = \eta.$$

Dans ce module, nous rencontrerons surtout des matrices carrées $A \in M_n(K)$, écrites par blocs en prenant pour les lignes et les colonnes un même découpage $(I_1, \dots, I_p)$ de $[\![1,n]\!]$: $A = \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,p} \\ A_{2,1} & A_{2,2} & \dots & A_{2,p} \\ \dots & \dots & \dots & \dots \\ A_{p,1} & A_{p,2} & \dots & A_{p,p} \end{pmatrix}$. Noter que les matrices $A_{1,1}, \dots, A_{p,p}$ sont alors carrées. Nous dirons de plus que A est *triangulaire par blocs* si $A_{h,k} = 0$ dès que $h,k \in [\![1,p]\!]$ et $h > k$. Nous dirons que A est *diagonale par blocs* si $A_{h,k} = 0$ dès que $h,k \in [\![1,p]\!]$ sont distincts. Si tel est le cas, nous écrirons $A := \mathrm{Diag}(A_{1,1}, \dots, A_{p,p})$. La somme et le produit de matrices diagonales par blocs (de mêmes formats) se calculent simplement : $\mathrm{Diag}(A_1, \dots, A_p) + \mathrm{Diag}(B_1, \dots, B_p) = \mathrm{Diag}(A_1 + B_1, \dots, A_p + B_p)$ et $\mathrm{Diag}(A_1, \dots, A_p)\,\mathrm{Diag}(B_1, \dots, \dots, B_p) = \mathrm{Diag}(A_1B_1, \dots, A_pB_p)$.

La proposition suivante fournit un procédé commode pour réduire un endomorphisme (*cf.* la section 2.1 pour la notion de sous-espace stable par un endomorphisme) :

Proposition 4. Soient E un espace vectoriel de dimension finie $n \geqslant 1$, u un endomorphisme de E et $\mathcal{B}$ une base de E. Pour que $A := \mathrm{Mat}_{\mathcal{B}}(u)$ soit diagonale par blocs, il faut, et il suffit, que E soit somme directe de sous-espaces non nuls $V_1, \dots, V_m$ stables par u tels que $\mathcal{B}$ soit obtenue en juxtaposant des bases de chacun des V_i.

Démonstration. Notons $a_{h,k}$ les coefficients de A, et posons $\mathcal{B} := (e_1, \dots, e_n)$. Supposons que $E = V_1 \oplus V_2 \oplus \cdots \oplus V_m$, chaque V_i étant un sous-espace non nul de E stable par u, et que $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_m$ où, pour tout i, $\mathcal{B}_i$ est une base de V_i. Pour tout i, soit $I_i := \{h \in [\![1,n]\!] \mid e_h \in \mathcal{B}_i\}$. Alors $(I_1, \dots, I_m)$ est une partition de $[\![1,n]\!]$, formée d'intervalles consécutifs, d'où une écriture de A par blocs : $A := \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,m} \\ A_{2,1} & A_{2,2} & \dots & A_{2,m} \\ \dots & \dots & \dots & \dots \\ A_{m,1} & A_{m,2} & \dots & A_{m,m} \end{pmatrix}$. Il s'agit de montrer que, si $i,j \in [\![1,m]\!]$

sont distincts, $A_{i,j} = 0$. Chaque coefficient de $A_{i,j}$ est de la forme $a_{h,k}$, où $h \in I_i$ et $k \in I_j$. Mais $a_{h,k}$ est la $h^{ème}$ coordonnée de $u(e_k)$ dans la base $\mathcal{B}$. De plus $e_k \in V_j$, donc $u(e_k) \in V_j$; comme $h \notin I_j$, $a_{h,k} = 0$. Noter que, si $i \in [\![1, m]\!]$, $A_{i,i}$ est la matrice dans la base $\mathcal{B}_i$ de la restriction $u_i \in \mathcal{L}(V_i)$ de u à V_i.
Supposons inversement que A soit diagonale par blocs : $A := \mathrm{Diag}(A_1, \dots, A_m)$, et soit $(I_1, \dots, I_m)$ la partition de $[\![1, n]\!]$ correspondante. Pour tout $i \in [\![1, m]\!]$, posons $V_i := \mathrm{Vect}\big(e_h \mid h \in I_i\big)$. D'après la proposition 2 de la page 215, E est somme directe de $V_1, \dots, V_m$. Soit $j \in [\![1, m]\!]$. Pour tout $k \in I_j$, $u(e_k)$ est combinaison linéaire des e_h pour $h \in I_j$, donc $u(e_k) \in V_j$. Il en résulte que V_j est stable par u. ∎

Pour une matrice carrée M définie par blocs, par exemple $M := \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ (A et D sont des matrices carrées), il n'y a pas de moyen simple de calculer $\det M$ en fonction des blocs, ici A, B, C, D. Voici cependant un cas particulier simple et important :

Théorème 5. Soit A une matrice triangulaire par blocs :

$$A := \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,m} \\ 0 & A_{2,2} & \dots & A_{2,m} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_{m,m} \end{pmatrix}.$$

En particulier, $A_{1,1}, \dots, A_{m,m}$ sont des matrices carrées. Alors :

$$\det(A) = \det(A_{1,1}) \det(A_{2,2}) \cdots \det(A_{m,m}).$$

Démonstration. En raisonnant par récurrence sur m, on se ramène au cas $m = 2$ puisque, si $m \geqslant 3$ et si l'on pose $A' := \begin{pmatrix} A_{2,2} & A_{2,3} & \dots & A_{2,m} \\ 0 & A_{3,3} & \dots & A_{3,m} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_{m,m} \end{pmatrix}$ et $A'' := (A_{1,2}\, A_{1,3} \dots\, A_{1,m})$, on a $A = \begin{pmatrix} A_{1,1} & A'' \\ 0 & A' \end{pmatrix}$. Supposons donc que $A := \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$, où $B \in M_p(K)$, $C \in M_{p,q}(K)$ et $D \in M_q(K)$, avec $p, q \in \mathbb{N}^*$ et $n := p + q$. On vérifie l'égalité :

$$\begin{pmatrix} B & C \\ 0 & D \end{pmatrix} = \begin{pmatrix} I_p & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I_p & C \\ 0 & I_q \end{pmatrix} \begin{pmatrix} B & 0 \\ 0 & I_q \end{pmatrix},$$

en notant, par exemple, que le produit des deux premières matrices du second membre s'obtient en multipliant (à gauche) par D la deuxième ligne de $\begin{pmatrix} I_p & C \\ 0 & I_q \end{pmatrix}$. La matrice $\begin{pmatrix} I_p & C \\ 0 & I_q \end{pmatrix}$ est triangulaire supérieure et ses termes diagonaux valent 1, donc $\det \begin{pmatrix} I_p & C \\ 0 & I_q \end{pmatrix} = 1$. En développant par rapport à la dernière ligne, on voit que $\det \begin{pmatrix} B & 0 \\ 0 & I_q \end{pmatrix}$ vaut $\det B$ si $q := 1$ et $\det \begin{pmatrix} B & 0 \\ 0 & I_{q-1} \end{pmatrix}$ si $q \geqslant 2$. Par récurrence sur q, on obtient donc $\det \begin{pmatrix} B & 0 \\ 0 & I_q \end{pmatrix} = \det B$. On montre de même que $\det \begin{pmatrix} I_p & 0 \\ 0 & D \end{pmatrix} = \det D$, d'où la conclusion. ∎

1.3 Polynômes d'endomorphismes

Deux polynômes jouent un rôle essentiel dans la réduction de l'endomorphisme u (ou de la matrice A), ce sont le polynôme caractéristique et le polynôme minimal. Ces deux polynômes sont d'ailleurs reliés l'un à l'autre, comme nous le verrons.

Commençons par quelques rappels sur les polynômes. Un polynôme $P \in K[X]$ non constant, *i.e.* de degré $n \geqslant 1$, est dit *scindé*, ou *scindé sur* K, s'il est produit de polynômes de degré 1. Si c'est le cas, on peut écrire $P := a(X-\alpha_1)(X-\alpha_2)\cdots(X-\alpha_n)$, où $a \in K^*$ est le *coefficient dominant* de P, *i.e.* le coefficient de X^n dans P, et $\alpha_1, \ldots, \alpha_n \in K$. Dans cette écriture, il faut se rappeler que les α_i ne sont pas nécessairement distincts. On dit que $(\alpha_1, \ldots, \alpha_n)$ est un *système de racines* de P. Il y a une écriture plus intrinsèque de P. Notons $\{\lambda_1, \ldots, \lambda_p\}$ l'ensemble des racines de P. Alors P s'écrit ainsi :

$$P \;:=\; a(X-\lambda_1)^{m_1}(X-\lambda_2)^{m_2}\cdots(X-\lambda_p)^{m_p}. \tag{5}$$

Ici, chaque m_j est un entier strictement positif, appelé *multiplicité de* λ_j dans P, et l'on a $\sum_{j=1}^{p} m_j = n$. On dit que la racine λ_j de P est simple si $m_j = 1$ et multiple si $m_j > 1$ (double si $m_j = 2$, etc). Pour que λ_j soit racine simple de P, il faut, et il suffit, que λ_j ne soit pas racine de P'. Lorsque tous les m_j valent 1, on dit que P est *à racines simples*. Cela revient à dire que P et P' sont premiers entre eux.

Le corps K est dit *algébriquement clos* si tout polynôme non constant de $K[X]$ est scindé, *i.e.* si tout polynôme non constant de $K[X]$ possède au moins une racine (*cf.* [L1]). Ainsi $\mathbb{C}$ est algébriquement clos (théorème de D'Alembert-Gauß).

Rappelons que le polynôme caractéristique de A est le déterminant de la matrice $XI_n - A$, il est noté χ_A. Le polynôme caractéristique de u est, par définition, le polynôme caractéristique de la matrice de u dans n'importe quelle base de E, il est noté χ_u.

Définition 1.
On dit que u est *scindé* si χ_u est scindé. De même A est dite *scindée* si χ_A est scindé.

Dans ce module, certains résultats essentiels, portant sur u (resp. sur A), seront démontrés sous l'hypothèse que u est scindé (resp. que A est scindée). Ces hypothèses ne sont pas trop limitatives puisque, dans la pratique, le cas le plus important est celui du corps $\mathbb{C}$.

Soit maintenant R une K-algèbre (en ce qui concerne cette notion d'algèbre, on se reportera au module I.1). Étant donné un élément x de R, il existe un unique morphisme d'algèbres $\varphi : K[X] \to R$ tel que $\varphi(X) := x$. Si $P := a_n X^n + \cdots + a_1 X + a_0 \in K[X]$, on a explicitement $\varphi(P) = a_n x^n + \cdots + a_1 x + a_0 1_A$, ce qui fait que $\varphi(P)$ est plutôt noté $P(x)$. L'image de φ est la sous-algèbre $K[x]$ de R engendrée par x, ses éléments sont appelés (abusivement) *polynômes en* x *à coefficients dans* K. Le noyau de φ est un idéal $K[X]$ Si cet idéal est non nul, *ce qui est automatique si* R *est de dimension finie*, il possède un unique générateur unitaire, appelé *polynôme minimal de* x et noté μ_x. Chacune des deux propriétés ci-dessous caractérise μ_x :

1. C'est le polynôme unitaire P de plus bas degré annulant x, *i.e.* tel que $P(x) = 0$.
2. Il est unitaire, et les polynômes annulant x sont exactement les multiples de μ_x.

Dans ce qui précède, prenons $R := \mathcal{L}(E)$ et $x := u$. Le fait que $P \mapsto P(u)$ soit un morphisme d'anneaux montre que $(PQ)(u) = P(u) \circ Q(u)$ pour tous $P, Q \in K[X]$. Comme $QP = PQ$, il en résulte que $P(u) \circ Q(u) = Q(u) \circ P(u)$. L'élément neutre multiplicatif de $\mathcal{L}(E)$ est Id_E. Pour tout polynôme $P := a_m X^m + \cdots + a_1 X + a_0$, on a donc $P(u) = a_m u^m + \cdots + a_1 u + a_0 \mathrm{Id}_E$. Puisque $\mathcal{L}(E)$ est de dimension finie, le polynôme minimal de u est défini, il est noté μ_u. Noter que $\mu_u \neq 1$ parce que $E \neq \{0\}$ et que, par exemple, le polynôme minimal d'une homothétie $\lambda \mathrm{Id}_E$ est $X - \lambda$.

Prenons ensuite $R := M_n(K)$ et $x := A$. Ici, pour tous $P, Q \in K[X]$, on a simplement $(PQ)(A) = P(A)Q(A)$ et, pour tout polynôme $P := a_m X^m + \cdots + a_1 X + a_0$, on a $P(A) = a_m A^m + \cdots + a_1 A + a_0 I_n$, puisque l'élément neutre multiplicatif de $M_n(K)$ est la matrice identité I_n. Puisque $M_n(K)$ est de dimension finie, le polynôme minimal de A est défini, il est noté μ_A.

La sous-algèbre $K[A]$ de $M_n(K)$ engendrée par A est formée des matrices $P(A)$, où P décrit $K[X]$. Le morphisme $\varphi : P \mapsto P(A)$ de $K[X]$ dans $M_n(K)$ a pour image $K[A]$, et son noyau est l'idéal (μ_A). Posons $p := \deg \mu_A$. Alors $K[X] = K_{p-1}[X] \oplus (\mu_A)$, donc φ induit un isomorphisme de $K_{p-1}[X]$ sur $K[A]$. Comme $K_{p-1}[X]$ est de dimension p, il vient : $\dim K[A] = \deg \mu_A$.

Le résultat suivant est utile pour passer du langage des endomorphismes au langage matriciel ou inversement :

Proposition 6. Supposons que $A := \mathrm{Mat}_{\mathcal{B}}(u)$, où $\mathcal{B}$ est une base de E.
Alors $\chi_u = \chi_A$ et $\mu_u = \mu_A$.

Démonstration. L'égalité $\chi_u = \chi_A$ vient de la définition de χ_u. Ensuite, $v \mapsto \mathrm{Mat}_{\mathcal{B}}(v)$ est un isomorphisme d'algèbres de $\mathcal{L}(E)$ sur $M_n(K)$. Pour tout $P \in K[X]$, $P(A)$ est donc la matrice de $P(u)$ dans la base $\mathcal{B}$. En particulier $P(A) = 0$ si, et seulement si, $P(u) = 0$. L'égalité $\mu_A = \mu_u$ se déduit alors de l'une des caractérisations 1 ou 2. ∎

Voici un résultat très important faisant intervenir des polynômes en l'endomorphisme u :

Théorème 7 (Lemme des noyaux). Soient $P_1, \ldots, P_r \in K[X]$ des polynômes premiers entre eux *deux à deux*. Posons $P := P_1 P_2 \cdots P_r$. Alors :

1. Le noyau $N := \mathrm{Ker}\big(P(u)\big)$ est somme directe des $N_i := \mathrm{Ker}\big(P_i(u)\big)$, pour $i = 1, \ldots, r$.
2. Les projecteurs $\pi_1, \ldots, \pi_r \in \mathcal{L}(N)$ associés à la décomposition $N = \bigoplus_{i=1}^{r} N_i$ sont des polynômes en l'endomorphisme induit $u_N \in \mathcal{L}(N)$.

L'espace vectoriel E est ici de dimension quelconque, finie ou infinie.

Démonstration. L'endomorphisme induit u_N est l'application $x \mapsto u(x)$ de N dans N, *cf.* la section 2.1. Pour tout i, notons Q_i le produit des P_j pour $j \neq i$, de sorte que $P = P_i Q_i$. Puisque $P(u) = (Q_i P_i)(u) = Q_i(u) \circ P_i(u)$, on a $N_i \subset N$. Quitte à remplacer E par N et u par $u_N \in \mathcal{L}(N)$, nous pouvons alors supposer que $E = N$, *i.e.* $P(u) = 0$. Par récurrence sur r, on se ramène aisément au cas $r = 2$ (le cas $r = 1$ est clair). Puisque P_1, P_2 sont premiers entre eux, il existe $U_1, U_2 \in K[X]$ tels que $U_1P_1 + U_2P_2 = 1$ (théorème de Bézout). Posons alors $\pi_1 := (U_2P_2)(u)$ et $\pi_2 := (U_1P_1)(u)$, de sorte que $\pi_1 + \pi_2 = \mathrm{Id}_E$. D'abord $\pi_1 \circ \pi_2 = 0$, *i.e.* $(U_2P_2)(u) \circ (U_1P_1)(u) = 0$. En effet, on a déjà :

$$P_2(u) \circ (U_1P_1)(u) = U_1(u) \circ (P_1P_2)(u) = U_1(u) \circ P(u) = 0.$$

D'après l'exercice 1 de la page 215, E est somme directe de $V_1 := \mathrm{Im}(\pi_1) = \mathrm{Ker}(\pi_2)$ et $V_2 := \mathrm{Im}(\pi_2) = \mathrm{Ker}(\pi_1)$. De l'égalité $P_2(u) \circ \pi_2 = 0$, on déduit ensuite l'inclusion $V_2 = \mathrm{Im}(\pi_2) \subset \mathrm{Ker}\big(P_2(u)\big) = N_2$.

Or $V_2 = \mathrm{Ker}\big(U_2(u) \circ P_2(u)\big) \supset \mathrm{Ker}\big(P_2(u)\big) = N_2$, d'où $V_2 = N_2$ et de même $V_1 = N_1$. L'assertion 1 est alors une conséquence de l'exercice précité. L'assertion 2 aussi, car π_1 et π_2 sont par construction des polynômes en u. ∎

Exemple. Soient V, W deux sous-espaces vectoriels supplémentaires de E. Supposons que u soit la symétrie par rapport à V et de direction W (ici K est de caractéristique distincte de 2). Puisque $u^2 = \mathrm{Id}_E$, le polynôme $P := X^2 - 1$ annule u. Le théorème dit que E est somme directe de $V = \mathrm{Ker}(u - \mathrm{Id}_E)$ et de $W = \mathrm{Ker}(u + \mathrm{Id}_E)$, ce que nous savions déjà.

1.4 Trace d'une matrice carrée, d'un endomorphisme

Définition 2. On appelle *trace* d'une matrice carrée A la somme de ses termes diagonaux, notée $\mathrm{Tr}\, A$: $\mathrm{Tr}\, A := \sum_{i=1}^{n} a_{i,i}$ si $A := (a_{i,j}) \in M_n(K)$.

Une matrice carrée et sa transposée ont évidemment la même trace. Dans [L1], nous avons vu que le coefficient de X^{n-1} dans le polynôme caractéristique χ_A d'une matrice $A \in M_n(K)$ est l'opposé de la somme des coefficients diagonaux de A. Ainsi :

$$\chi_A = X^n - \mathrm{Tr}(A)X^{n-1} + \cdots + (-1)^n \det(A). \tag{6}$$

Par exemple, si $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(K)$, $\mathrm{Tr}(A) = a + d$, $\det(A) = ad - bc$, et donc :

$$\chi_A = X^2 - \mathrm{Tr}(A)X + \det(A) = X^2 - (a + d)X + (ad - bc).$$

Proposition 8. Soit n un entier strictement positif.

1. L'application $\mathrm{Tr} : M_n(K) \to K$ est une *forme linéaire* sur $M_n(K)$. Pour toutes matrices $A, B \in M_n(K)$ et tout $\lambda \in K$, on a donc :
$$\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B) \quad \text{et} \quad \mathrm{Tr}(\lambda A) = \lambda\, \mathrm{Tr}(A).$$
2. Pour toutes matrices $A, B \in M_n(K)$, on a $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.
3. Deux matrices carrées *semblables* ont la même trace.

Démonstration. Soient $A = (a_{i,j})$ et $B = (b_{i,j})$. On a : $A + B = (a_{i,j} + b_{i,j})$ et $\lambda A = (\lambda a_{i,j})$, d'où l'assertion 1. Écrivons ensuite $AB = (c_{i,j})$. Alors :

$$\mathrm{Tr}(AB) = \sum_{i=1}^{n} c_{i,i} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{i,j} b_{j,i} \right) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} a_{j,i} b_{i,j} \right).$$

La première égalité est la définition de la trace, la deuxième vient de la définition de AB et la troisième d'un changement d'indices muets (échange de i et j). Échangeons A et B :

$$\mathrm{Tr}(BA) \;=\; \sum_{i=1}^{n} \left(\sum_{j=1}^{n} b_{i,j} a_{j,i} \right).$$

D'où $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, vu la commutativité de la multiplication de K.

Si A et B sont semblables, il existe $P \in GL_n(K)$ telle que $B := P^{-1}AP$, et 2 donne :

$$\mathrm{Tr}(B) = \mathrm{Tr}\big(P^{-1}AP\big) = \mathrm{Tr}\big((P^{-1}A)P\big) = \mathrm{Tr}\big(P(P^{-1}A)\big) = \mathrm{Tr}(A). \qquad \blacksquare$$

> ***Remarque.*** Dans la définition de la trace d'une matrice carrée, comme dans la proposition précédente, on peut supposer que K est seulement un anneau commutatif. Rappelons qu'alors $P \in M_n(K)$ est inversible, *i.e.* appartient à $GL_n(K)$ si, et seulement si, $\det A$ est inversible dans K.

Théorème et définition 9. Soient E un K-espace vectoriel de dimension finie $n \geqslant 1$ et u un endomorphisme de E. Toutes les matrices représentant u (dans une certaine base) ont la même trace. Cette trace commune est appelée *trace* de u et notée $\mathrm{Tr}(u)$.

Démonstration. Soient $\mathcal{B}, \mathcal{B}'$ deux bases de E.

Les matrices $\mathrm{Mat}_{\mathcal{B}}(u), \mathrm{Mat}_{\mathcal{B}'}(u) \in M_n(K)$ sont semblables, elles ont donc même trace, vu la proposition précédente. $\blacksquare$

La proposition suivante résulte aussitôt de ce théorème :

Proposition 10. Soit E un K-espace vectoriel de dimension finie $n \geqslant 1$.

1. L'application $\mathrm{Tr} : \mathcal{L}(E) \to K$ est une *forme linéaire* sur $\mathcal{L}(E)$: pour tous $u, v \in \mathcal{L}(E)$ et $\lambda \in K$, on a :
$$\mathrm{Tr}(u + v) = \mathrm{Tr}(u) + \mathrm{Tr}(v) \qquad \text{et} \qquad \mathrm{Tr}(\lambda u) = \lambda \,\mathrm{Tr}(u).$$
2. Pour tous $u, v \in \mathcal{L}(E)$, on a $\mathrm{Tr}(u \circ v) = \mathrm{Tr}(v \circ u)$.
3. Deux endomorphismes *semblables* ont la même trace.
4. Pour tout $u \in \mathcal{L}(E)$, on a :
$$\chi_u \;=\; X^n - \mathrm{Tr}(u) X^{n-1} + \cdots + (-1)^n \det(u). \tag{7}$$

2 Diagonalisabilité — Trigonalisabilité

2.1 Sous-espaces propres, sous-espaces stables

Rappelons qu'un sous-espace vectoriel V de E est dit *stable* par u si $u(V) \subset V$. Lorsque c'est le cas, l'application $x \mapsto u(x)$ de V dans V (obtenue par restriction et corestriction de u) est un endomorphisme de V, nous le noterons u_V. On dit que u_V est l'endomorphisme de V *induit* par u. Voici un résultat simple mais utile.

Proposition 11. Soient u, v deux endomorphismes d'un espace vectoriel E. On suppose que u et v commutent ($u \circ v = v \circ u$). Alors v laisse stables le noyau et l'image de u, ainsi que chaque espace propre de u.

Démonstration. Si $x \in \operatorname{Ker} u$, on a $u\big(v(x)\big) = v\big(u(x)\big) = v(0) = 0$, donc $v(x) \in \operatorname{Ker} u$. D'où $v(\operatorname{Ker} u) \subset \operatorname{Ker} u$. Si $b \in \operatorname{Im} u$, soit $a \in E$ tel que $b = u(a)$. Alors $v(b)$ est égal à $v\big(u(a)\big) = u\big(v(a)\big)$, donc $v(b) \in \operatorname{Im} u$. D'où $v(\operatorname{Im} u) \subset \operatorname{Im} u$. Soit λ une valeur propre de u. Puisque v commute avec u, il commute avec $u - \lambda \operatorname{Id}_E$. Ce qui précède montre que v laisse stable le noyau de $u - \lambda \operatorname{Id}_E$, à savoir l'espace propre de u associé à λ. ∎

Les sous-espaces propres d'un endomorphisme sont des sous-espaces stables particuliers, ils sont extrêmement importants. Quelques rappels à ce sujet peuvent être utiles (*cf.* [L1]). Pour l'instant, la dimension de E est quelconque. Un scalaire $\lambda \in K$ est appelé *valeur propre* de u si $u - \lambda \operatorname{Id}_E$ n'est pas injectif. Si E est de dimension finie, cela revient à dire que $u - \lambda \operatorname{Id}_E$ n'est pas inversible, *i.e.* que $\det(u - \lambda \operatorname{Id}_E) = 0$; les valeurs propres de u sont ainsi les racines de χ_u. La définition suivante vaut en dimension quelconque :

Définition 3. On appelle *spectre* de u l'ensemble des $\lambda \in K$ tels que $u - \lambda \operatorname{Id}_E$ ne soit pas inversible. Ce spectre est noté $\operatorname{sp} u$. Le spectre de $A \in M_K$, noté $\operatorname{sp} A$, est par définition le spectre de l'endomorphisme de K^n canoniquement associé à A.

En dimension finie, nous savons qu'un endomorphisme est injectif si, et seulement s'il est inversible. D'où le résultat suivant :

Proposition 12. En dimension finie, le spectre de u est l'ensemble des valeurs propres de u, ou encore l'ensemble des racines du polynôme χ_u. Ainsi $\operatorname{sp} u$ est *fini*.

Définition 4. On appelle *multiplicité* d'une valeur propre λ de u (resp. A) la multiplicité de λ comme racine de χ_u (resp. χ_A), *i.e.* le plus grand entier naturel m tel que $(X - \lambda)^m$ divise χ_u (resp. χ_A).
On parle ainsi de valeur propre simple, double, etc.

Soit $\lambda \in K$. Le noyau de $u - \lambda \operatorname{Id}_E$ est un sous-espace vectoriel de E, nous le noterons $E_\lambda(u)$. Ainsi $E_\lambda(u) \neq \{0\}$ si, et seulement si, λ est une valeur propre de u. Supposons que ce soit le cas. On appelle *vecteur propre associé à* λ tout vecteur *non nul* de $E_\lambda(u)$. Par exemple, on a toujours $E_0(u) = \operatorname{Ker} u$, et 0 est valeur propre de u si, et seulement

si, u n'est pas injectif. Pour une matrice carrée $A \in M_n(K)$, les éléments propres (valeurs et vecteurs propres) ou le spectre de A sont ceux de l'endomorphisme de K^n canoniquement associé à A. Revoir [L1] pour les liens entre éléments propres d'endomorphismes et éléments propres des matrices carrées.

Remarque. Soient $u \in \mathcal{L}(E)$ et $x \in E_\lambda(u)$. De l'égalité $u(x) = \lambda x$, on tire $u^2(x) = u(\lambda x) = \lambda^2 x$ puis, par une récurrence immédiate, $u^k(x) = \lambda^k x$ pour tout $k \in \mathbb{N}$. On en déduit que $P(u)(x) = P(\lambda)x$ pour tout polynôme P (par linéarité en P). Ainsi $E_\lambda(u) \subset E_{P(\lambda)}\big(P(u)\big)$. Cette inclusion est en général stricte (*cf.* le cas où $P = \mu_u$ et $u \neq \lambda \operatorname{Id}$).
Si u est inversible et $x \neq 0$, on a $\lambda \neq 0$ et $u^{-1}(x) = \lambda^{-1}u^{-1}(\lambda x) = \lambda^{-1}x$, donc $x \in E_{1/\lambda}(u^{-1})$. Ainsi $E_\lambda(u) \subset E_{1/\lambda}(u^{-1})$.
En remplaçant (u, λ) par (u^{-1}, λ^{-1}), on voit qu'il y a égalité : $E_\lambda(u) = E_{1/\lambda}(u^{-1})$.

Théorème 13. La dimension de E est ici quelconque. Lorsque λ décrit K, la somme des $E_\lambda(u)$ est directe.

Démonstration. Vu la remarque de la page 215, il suffit de montrer que, si $p \in \mathbb{N}^*$ et si $\lambda_1, \dots, \lambda_p \in K$ sont distincts, la somme $E_{\lambda_1}(u) + \cdots + E_{\lambda_p}(u)$ est directe. Pour tout i, posons $P_i := X - \lambda_i$, de sorte que $P_i(u) = u - \lambda_i \operatorname{Id}_E$ et $E_{\lambda_i}(u) = \operatorname{Ker}\big(P_i(u)\big)$. Puisque les P_i sont premiers entre eux deux à deux, il suffit d'appliquer le lemme des noyaux. ∎

Corollaire 14. Toute famille de vecteurs propres de u associés à des valeurs propres deux à deux distinctes est libre.

Démonstration. C'est une conséquence immédiate du théorème précédent et de l'exercice de la page 216 (rappelons qu'un vecteur propre n'est jamais nul). ∎

Exercice 3.
Pour tout réel a strictement positif, notons f_a la fonction $x \mapsto \sin(ax)$. Montrer que la famille $(f_a)_{a>0}$ est libre.
Solution. L'application $u : f \mapsto f''$ est un endomorphisme de l'espace vectoriel des fonctions indéfiniment dérivables de $\mathbb{R}$ dans $\mathbb{R}$. Pour tout $a > 0$, la dérivée seconde de $\sin(ax)$ est $-a^2 \sin(ax)$, donc f_a est un vecteur propre de u, associé à la valeur propre $-a^2$. Lorsque a varie, ces valeurs propres sont distinctes, et l'on conclut grâce au corollaire.

Les valeurs propres de A sont exactement les racines du polynôme caractéristique χ_A (de même pour u). Le polynôme minimal possède la même propriété :

Proposition 15. Les valeurs propres de u (resp. de A) sont exactement les racines de son polynôme minimal μ_u (resp. μ_A).

Démonstration. Traitons le cas d'un endomorphisme (le cas d'une matrice s'en déduit aisément). Soit $\lambda \in K$. Supposons que λ soit une valeur propre de u : il existe un vec-

teur $x \neq 0$ tel que $u(x) = \lambda x$. Pour tout $k \in \mathbb{N}$, on a $u^k(x) = \lambda^k x$ (récurrence sur k), et l'on en déduit par linéarité que $Q(u)(x) = Q(\lambda)x$ pour tout $Q \in K[X]$. Comme $\mu_u(u) = 0$, on a en particulier $\mu_u(\lambda)x = 0$, d'où $\mu_u(\lambda) = 0$ puisque $x \neq 0$.

Supposons inversement que $\mu_u(\lambda) = 0$. Alors $X - \lambda$ divise μ_u, *i.e.* il existe $R \in K[X]$ tel que $\mu_u = (X-\lambda)R$. Puisque $\deg(R) < \deg(\mu_u)$, on a $R(u) \neq 0$ (définition du polynôme minimal). Mais $0 = \mu_u(u) = (u - \lambda \operatorname{Id}_E) \circ R(u)$, donc $u - \lambda \operatorname{Id}_E$ n'est pas inversible, *i.e.* n'est pas injectif. Ainsi λ est une valeur propre de u. ∎

Donnons quelques résultats qui seront utiles pour la réduction.

Proposition 16. Soient V un sous-espace vectoriel non nul de E stable par u et $u_V \in \mathcal{L}(V)$ l'endomorphisme induit.

1. Le polynôme minimal de u_V divise le polynôme minimal de u.
2. Le polynôme caractéristique de u_V divise le polynôme caractéristique de u.

Démonstration. Les assertions 1 et 2 se ressemblent, mais leurs preuves sont différentes.

1. Soit $v := u_V$. Pour tout $Q \in K[X]$, $Q(v) \in \mathcal{L}(V)$ est évidemment induit par $Q(u)$. En particulier $\mu_u(v)$ est induit par $\mu_u(u) = 0$. Ainsi μ_u annule v, donc μ_v divise μ_u.

2. Posons $n := \dim E$ et $p := \dim V$. Soit $(e_1, \dots, e_p)$ une base de V, complétons-la en une base $\mathcal{B} := (e_1, \dots, e_n)$ de E. Notons $A := (a_{i,j})$ la matrice de u dans la base $\mathcal{B}$. Pour tout $j \in [\![1, p]\!]$, on a $e_j \in V$, donc $u(e_j) = \sum_{i=1}^{n} a_{i,j} e_i \in V$, *i.e.* $a_{i,j} = 0$ pour tout $i \in [\![p+1, n]\!]$. Cela montre que A s'écrit par blocs : $A := \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$, où $B \in M_p(K)$, $C \in M_{p,n-p}(K)$ et $D \in M_{n-p}(K)$. En fait, B est la matrice de v dans la base $(e_1, \dots, e_p)$ de V. Puisque $XI_n - A = \begin{pmatrix} XI_p - B & -C \\ 0 & XI_{n-p} - D \end{pmatrix}$, le calcul d'un déterminant par blocs montre que $\chi_u = \chi_A = \det(XI_p - B)\det(XI_{n-p} - D) = \chi_B \chi_D = \chi_v \chi_D$, et par suite χ_v divise χ_u. ∎

Comme application de ce résultat, voici un théorème dont la preuve, très simple, met en évidence l'intérêt des sous-espaces stables :

Théorème 17. Pour toute valeur propre λ de u, la dimension de l'espace propre $E_\lambda(u)$ est au plus égale à la multiplicité de λ.

Démonstration. Posons $V := E_\lambda(u)$, $p := \dim V$, et soit $v := u_V \in \mathcal{L}(V)$ l'endomorphisme induit par u. En fait v est l'homothétie de rapport λ, de sorte que $\chi_v = (X - \lambda)^p$. D'après la proposition précédente, $(X - \lambda)^p$ divise χ_u, d'où la conclusion. ∎

Proposition 18. Supposons que E soit somme directe de sous-espaces vectoriels $V_1, \dots, V_p$ de E *stables par* u. Pour tout i, notons u_i l'endomorphisme de V_i induit par u. Alors :

1. Le polynôme minimal μ_u est le ppcm des μ_{u_i}, $i = 1, \dots, p$.
2. Le polynôme caractéristique χ_u est le produit des χ_{u_i}, $i = 1, \dots, p$.

Démonstration.

1. Si $Q \in K[X]$ et $i \in [\![1, p]\!]$, $Q(u_i) \in \mathcal{L}(V_i)$ est induit par $Q(u)$. Il en résulte que $Q(u)_{V_i} = 0$ si, et seulement si, Q est multiple de μ_{u_i}. Soit alors $Q \in K[X]$. Puisque E est somme des V_i, $Q(u) = 0$ si, et seulement si, chaque $Q(u)_{V_i}$ est nul, *i.e.* si, et seulement si, Q est multiple de chacun des polynômes μ_{u_i}, c'est-à-dire multiple du ppcm M de ces polynômes. Ainsi M est un générateur unitaire de l'idéal de $K[X]$ formé des polynômes annulant u, donc $M = \mu_u$, par définition de μ_u.
2. Pour tout i, soient $\mathcal{B}_i$ une base de V_i ; posons $n_i := \dim(V_i)$, $A_i := \mathrm{Mat}_{\mathcal{B}_i}(u_i)$. La réunion $\mathcal{B}$ des $\mathcal{B}_i$ est une base de E, et $A := \mathrm{Mat}_{\mathcal{B}}(u)$ est diagonale par blocs : $A = \mathrm{Diag}(A_1, \dots, A_p)$. D'où $XI_n - A = \mathrm{Diag}(XI_{n_1} - A_1, \dots, XI_{n_p} - A_p)$, l'égalité étant vérifiée dans $M_n\big(K[X]\big)$.
   Il suffit alors d'appliquer le théorème 5 de la page 219. ∎

Voici la traduction matricielle de ce résultat :

Proposition 19. Supposons que $A := \mathrm{Diag}(A_1, \dots, A_p)$ soit diagonale par blocs.

1. Le polynôme minimal μ_A est le ppcm des μ_{A_i}, $i = 1, \dots, p$.
2. Le polynôme caractéristique χ_A est le produit des χ_{A_i}, $i = 1, \dots, p$.

Démonstration. Compte tenu de la proposition 4 de la page 218, il suffit d'appliquer la proposition précédente à l'endomorphisme de K^n canoniquement associé à A. ∎

Donnons un autre exemple de sous-espace stable. Cet exemple nous servira dans la preuve du théorème de Cayley-Hamilton (section 2.2), et il sera important dans [MPA1], tant du point de vue théorique que du point de vue du calcul numérique.
Soit x un vecteur de E fixé. Notons $E_{u,x}$ le sous-espace vectoriel de E engendré par les $u^k(x)$, $k \in \mathbb{N}$. Noter que $x = u^1(x) \in E_{u,x}$. Ce sous-espace est stable par u car, pour tout k, $u\big(u^k(x)\big) = u^{k+1}(x) \in E_{u,x}$. C'est même, au sens de l'inclusion, le *plus petit* sous-espace de E stable par u et contenant x. En effet, soit V un tel sous-espace. Puisque $(y \in V) \Rightarrow \big(u(y) \in V\big)$, on voit aussitôt par récurrence que, pour tout $k \in \mathbb{N}$, $(y \in V) \Rightarrow \big(u^k(y) \in V\big)$. En particulier $u^k(x) \in E_{u,x}$ pour tout k, d'où $E_{u,x} \subset V$.
L'application f de $K[X]$ dans E définie par $f(Q) := Q(u)(x)$ est linéaire. Puisque les X^k engendrent $K[X]$, l'image de f est engendrée par les $u^k(x)$, d'où :

$$E_{u,x} := \mathrm{Vect}\big(u^k(x) \mid k \in \mathbb{N}\big) = \big\{Q(u)(x) \mid Q \in K[X]\big\} = \mathrm{Im}\, f. \tag{8}$$

Lemme 20. Le noyau de f est un idéal non nul de $K[X]$. Soient P le générateur unitaire de $\operatorname{Ker} f$ et $p := \deg P$. Alors $\mathcal{C} := \big(x, u(x), \dots, u^{p-1}(x)\big)$ est une base de $E_{u,x}$. Par ailleurs, P divise μ_u.

Démonstration. Nous savons que $K[X] = K_{p-1}[X] \oplus (P)$, où $K_{p-1}[X]$ est formé des polynômes de degré au plus égal à $p-1$. Ainsi $K_{p-1}[X]$ est un supplémentaire de $\operatorname{Ker} f = (P)$. Il en résulte que l'image par f de la base $(1, X, \dots, X^{p-1})$ de $K_{p-1}[X]$ est une base de $\operatorname{Im} f = E_{u,x}$, d'où la première assertion. La deuxième est claire : $\mu_u(u) = 0$, *a fortiori* $0 = \mu_u(u)(x) = f(\mu_u)$, donc $\mu_u \in \operatorname{Ker} f = (P)$. ∎

Soit v l'endomorphisme de $E_{u,x}$ induit par u (rappelons que $E_{u,x}$ est stable par u). Quelle est la matrice de v dans la base $\mathcal{C}$? Pour le savoir, écrivons P sous la forme :

$$P := X^p + a_{p-1}X^{p-1} + \cdots + a_1 X + a_0. \tag{9}$$

Alors $0 = f(P) = u^p(x) + \sum_{k=0}^{p-1} a_k u^k(x)$, d'où $v\big(u^{p-1}(x)\big) = u^p(x) = -\sum_{k=0}^{p-1} a_k u^k(x)$. Les images par v de $x, u(x), \dots, u^{p-2}(x)$ étant $u(x), u^2(x), \dots, u^{p-1}(x)$ respectivement, la matrice $\operatorname{Mat}_{\mathcal{C}}(v)$ est égale à la matrice suivante :

$$C_P := \begin{pmatrix} 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & \ddots & \vdots & -a_2 \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 0 & 0 & \dots & 1 & -a_{p-1} \end{pmatrix} \in M_p(K). \tag{10}$$

La matrice C_P ainsi définie ne dépend en fait que du polynôme P, *i.e.* de $a_0, \dots, a_{p-1}$, ce qui justifie la définition suivante :

Définition 5. Soit $P \in K[X]$ un polynôme unitaire de degré $p \geqslant 1$, écrit comme en (9). La matrice C_P définie en (10) est appelée *matrice compagnon* du polynôme P.

Si par exemple $P := X^2 + 1$, on a $C_P := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Lorsque $K := \mathbb{R}$, il s'agit de la matrice de rotation d'angle $\pi/2$. Dans le cas général, donnons une propriété importante de cette matrice compagnon :

Proposition 21. Soient $P \in K[X]$ un polynôme unitaire de degré $p \geqslant 1$ et C_P sa matrice compagnon. Le polynôme caractéristique et le polynôme minimal de C_P sont tous deux égaux à P.

Démonstration. Pour le polynôme minimal, raisonnons en termes d'endomorphismes. Soient u l'endomorphisme de K^p canoniquement associé à C_P et $\mathcal{B} := (e_1, \dots, e_p)$ la

base canonique de K^p. Ainsi :

$$u(e_1) = e_2, u(e_2) = e_3, \dots, u(e_{p-1}) = e_p, u(e_p) = -(a_0e_1 + a_1e_2 + \cdots + a_{p-1}e_p).$$

D'où $u^k(e_1) = e_{k+1}$ pour $k = 0, \dots, p-1$, et $u^p(e_1) = -(a_0e_1 + a_1e_2 + \cdots + a_{p-1}e_p)$. Il vient ensuite $P(u)(e_1) = \big[u^p + (a_{p-1}u^{p-1} + \cdots + a_1u + a_0\,\mathrm{Id})\big](e_1)$, c'est-à-dire $P(u)(e_1) = -(a_0e_1 + a_1e_2 + \cdots + a_{p-1}e_p) + (a_{p-1}e_p + \cdots + a_1e_2 + a_0e_1) = 0$. Ainsi $P(u)(e_1) = 0$. Pour tout $k \in [\![1, p]\!]$, $P(u)(e_k)$ vaut alors :

$$P(u)\big(u^{k-1}(e_1)\big) = \big(P(u) \circ u^{k-1}\big)(e_1) = \big(u^{k-1} \circ P(u)\big)(e_1) = u^{k-1}\big(P(u)(e_1)\big) = 0.$$

L'endomorphisme $P(u)$ est donc nul, car son noyau contient tous les e_k.
Soit maintenant $Q = b_{p-1}X^{p-1} + \cdots + b_1X + b_0 \in K_{p-1}[X]$ un polynôme annulant u. En particulier $Q(u)(e_1) = 0$, soit $b_{p-1}e_p + \cdots + b_1e_2 + b_0e_1 = 0$. Les b_i sont donc tous nuls, *i.e.* $Q = 0$. En conclusion, u est annulé par P, mais n'est annulé par aucun polynôme de degré strictement inférieur à p, à part 0. Ainsi $\mu_u = P$, d'où $\mu_{C_P} = P$.

Le polynôme caractéristique de C_P vaut :

$$\begin{vmatrix} X & 0 & \dots & \dots & 0 & a_0 \\ -1 & X & \dots & \dots & 0 & a_1 \\ 0 & -1 & \ddots & & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -1 & X & a_{p-2} \\ 0 & 0 & \dots & 0 & -1 & X + a_{p-1} \end{vmatrix}.$$

Notons L_i la ligne contenant a_i. L'opération $L_0 \leftarrow L_0 + XL_1 + \cdots + X^{p-1}L_{p-1}$ transforme ce déterminant en le suivant :

$$\begin{vmatrix} 0 & 0 & \dots & \dots & 0 & P \\ -1 & X & \dots & \dots & 0 & a_1 \\ 0 & -1 & \ddots & & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -1 & X & a_{p-2} \\ 0 & 0 & \dots & 0 & -1 & X + a_{p-1} \end{vmatrix}.$$

En développant par rapport à la première ligne, on obtient : $(-1)^{p+1} \times P \times (-1)^{p-1}$, c'est-à-dire P.

Ainsi les valeurs propres de C_P ne sont autres que les racines de P. Certaines méthodes d'analyse numérique utilisent cette remarque et permettent de calculer des valeurs approchées des racines de P à l'aide du calcul matriciel.

2.2 Théorème de Cayley-Hamilton

Un théorème essentiel relie le polynôme caractéristique et le polynôme minimal :

Théorème 22 (Cayley-Hamilton). Le polynôme caractéristique χ_u de u annule u, c'est-à-dire $\chi_u(u) = 0$. En d'autres termes, le polynôme minimal μ_u de u divise le polynôme caractéristique χ_u. De même $\chi_A(A) = 0$, autrement dit μ_A divise χ_A.

Démonstration. Soit $x \in E$. Nous allons montrer que $\chi_u(u) \in \mathcal{L}(E)$ est nul en x. Soit V le sous-espace de E engendré par les $u^k(x)$, $k \in \mathbb{N}$. Nous savons qu'il est stable par u et contient x (*cf.* la section 2.1). Soit $v := u_V$ l'endomorphisme de V induit par u. D'après la proposition 21 de la page 228, $\mu_v = \chi_v$. Comme $\mu_v(v) = 0$, on en déduit que $\chi_v(v) = 0$. Mais $\chi_v(v)$ est l'endomorphisme de V induit par $\chi_v(u)$, et par suite la restriction de $\chi_v(u)$ à V est nulle, en particulier $\chi_v(u)(x) = 0$. Or, la proposition 16 de la page 226 montre que χ_v divise χ_u, écrivons $\chi_u = Q\chi_v$. Alors $\chi_u(u) = Q(u) \circ \chi_v(u)$, d'où $\chi_u(u)(x) = Q(u)\big(\chi_v(u)(x)\big) = Q(u)(0) = 0$. Ainsi $\chi_u(u)(u)(x) = 0$ pour tout $x \in E$, donc $\chi_u(u) = 0$. Par définition du polynôme minimal μ_u, cela signifie que μ_u divise χ_u.

Soit maintenant $A \in M_n(K)$. Notons $\mathcal{B}$ la base canonique de K^n et u l'endomorphisme de K^n canoniquement associé à A. Alors $\chi_A = \chi_u$ et, par ailleurs, $\chi_A(A)$ est la matrice dans la base $\mathcal{B}$ de $\chi_A(u) = \chi_u(u) = 0$. D'où $\chi_A(A) = 0$. ∎

Si $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, on a donc : $A^2 - (a+d)A + (ad - bc)I_2 = 0$. Prenons par exemple $A := \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}$. Alors $A^2 - 2A + I_2 = 0$, soit $A^2 = 2A - I_2$. Multiplions par A : $A^3 = 2A^2 - A = 2(2A - I_2) - A = 3A - 2I_2$. On en déduit que, pour tout $k \in \mathbb{N}$, A^k s'écrit $A^k = a_k A + b_k I_2$, où $(a_k), (b_k)$ sont les suites définies par récurrence ainsi :

$$a_0 := 0, a_1 := 1, b_0 := 1, b_1 := 0 \quad \text{et} \quad a_{k+2} := 2a_{k+1} - a_k,\ b_{k+2} := 2b_{k+1} - b_k \quad \text{pour tout } k.$$

Le lecteur pourra vérifier que $a_k = k$ et $b_k = 1 - k$ pour tout k.

Le théorème de Cayley-Hamilton peut faciliter la détermination du polynôme minimal.

Exemple. Soit $A := \begin{pmatrix} -1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}$. On a $\chi_A = (X+1)(X-1)^2$. Nous savons que μ_A divise χ_A et a les mêmes racines que χ_A, à savoir 1 et -1. Il n'y a donc que deux possibilités : $\mu_A = X^2 - 1$ ou $\mu_A = (X+1)(X-1)^2$. Le calcul donne $A^2 = \begin{pmatrix} 1 & 0 & 6 \\ 0 & 1 & 6 \\ 0 & 0 & 1 \end{pmatrix}$.

Ainsi $A^2 \neq I_2$, *i.e.* $X^2 - 1$ n'annule pas A. Il en résulte que $\mu_A = (X+1)(X-1)^2$.

2.3 Endomorphismes et matrices diagonalisables

Les calculs portant sur des matrices diagonales sont simples. Si $A := \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$, $B := \mathrm{Diag}(\mu_1, \ldots, \mu_n)$ et $k \in \mathbb{N}$, on a en effet : $A + B = \mathrm{Diag}(\lambda_1 + \mu_1, \ldots, \lambda_n + \mu_n)$, $AB = \mathrm{Diag}(\lambda_1\mu_1, \ldots, \lambda_n\mu_n)$ et $A^k = \mathrm{Diag}(\lambda_1^k, \ldots, \lambda_n^k)$. De même, lorsque $K := \mathbb{C}$, le calcul de l'exponentielle d'une matrice diagonale est très simple (*cf.* le module II.2).
Les matrices ou endomorphismes « diagonalisables » conduisent encore à des calculs simples. Voici les définitions correspondantes :

Définition 6. On dit que u est *diagonalisable* s'il existe une base de E dans laquelle la matrice de u soit *diagonale*. La matrice $A \in M_n(K)$ est dite *diagonalisable* si l'endomorphisme de K^n canoniquement associé est diagonalisable.

Une homothétie $\lambda \mathrm{Id}_E$ est diagonalisable, car sa matrice dans n'importe quelle base est λI_n. Dans le module I.4, nous avons vu des exemples non triviaux de matrices ou d'endomorphismes diagonalisables. Ainsi toute matrice symétrique *réelle* est diagonalisable. De même tout endomorphisme *normal* (par exemple hermitien) d'un espace hermitien est diagonalisable. À propos de ces résultats, insistons sur les faits suivants :

- Leurs démonstrations font intervenir la nature particulière du corps de base ($\mathbb{R}$ ou $\mathbb{C}$), et elles sont très différentes de celles que nous rencontrerons dans cette section.
- Le théorème spectral hermitien (par exemple) ne se contente pas de dire qu'un endomorphisme hermitien u est diagonalisable, il affirme que les sous-espaces propres de u sont orthogonaux, ce qui est capital.

Donnons une traduction simple des définitions ci-avant :

Proposition 23.

1. Pour que A soit diagonalisable, il faut, et il suffit, qu'elle soit semblable à une matrice diagonale.
2. Soit $\mathcal{B}$ une base de E. Pour que u soit diagonalisable, il faut, et il suffit, que $A := \mathrm{Mat}_{\mathcal{B}}(u)$ soit diagonalisable.

Démonstration. Pour l'assertion 1, prenons pour u l'endomorphisme de K^n canoniquement associé à A : $A = \mathrm{Mat}_{\mathcal{B}}(u)$, $\mathcal{B}$ étant la base canonique de K^n. Pour que A soit diagonalisable, il faut, et il suffit, qu'il existe une base $\mathcal{C}$ de E telle que $\mathrm{Mat}_{\mathcal{C}}(u)$ soit diagonale. Posant $P := \mathrm{Pass}(\mathcal{B}, \mathcal{C})$, on a $\mathrm{Mat}_{\mathcal{C}}(u) = P^{-1}AP$. Comme $\mathcal{C} \mapsto P$ est une bijection de l'ensemble des bases de K^n sur $GL_n(K)$, A est diagonalisable si, et seulement s'il existe $P \in GL_n(K)$ telle que $P^{-1}AP$ soit diagonale. D'où l'équivalence annoncée.
Prouvons l'assertion 2. Nous savons que A est diagonalisable si, et seulement s'il existe $P \in GL_n(K)$ telle que $P^{-1}AP$ soit diagonale, *i.e.* si, et seulement s'il existe une base $\mathcal{C}$ de E telle que $\mathrm{Mat}_{\mathcal{C}}(u)$ soit diagonale, ce qui équivaut à dire que u est diagonalisable. ∎

Voici ensuite d'importantes caractérisations de la diagonalisabilité.

Théorème 24. Les conditions suivantes sont équivalentes :

1. L'endomorphisme u est diagonalisable.
2. Il existe une base de E formée de vecteurs propres de u.
3. L'espace E est somme (directe) des espaces propres $E_\lambda(u)$ pour $\lambda \in \mathrm{sp}(u)$.

Démonstration. Supposons u diagonalisable et soit $\mathcal{B} := (e_1, \dots, e_n)$ une base de E telle que $\mathrm{Mat}_{\mathcal{B}}(u) := \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$ soit diagonale. Si $j \in [\![1, n]\!]$, $u(e_j) = \lambda_j e_j$, donc e_j est vecteur propre de u. D'où $1 \Longrightarrow 2$.
La somme V des sous-espaces propres de u est directe (théorème 13 de la page 225). Si une base de E est formée de vecteurs propres de u, chacun de ses éléments appartient à V, donc V contient $\mathcal{B}$ et par suite $V = E$. Ainsi $2 \Longrightarrow 3$.
Enfin, si $V = E$, soit $\mathcal{B}$ une base de E obtenue par juxtaposition d'une base de chacun des sous-espaces propres de u. Il est clair que $\mathrm{Mat}_{\mathcal{B}}(u)$ est diagonale, donc $3 \Longrightarrow 1$. ∎

Exemple. La matrice $A := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ n'est pas diagonalisable. En effet, $\chi_A = X^2$. Si A était diagonalisable, elle serait semblable à une matrice diagonale $D := \mathrm{Diag}(\lambda, \mu)$, et l'on aurait $X^2 = \chi_A = \chi_D = (X-\lambda)(X-\mu)$, d'où $\lambda = \mu = 0$ et donc $D = 0$. C'est absurde : la seule matrice semblable à la matrice nulle 0_2 est 0_2 et $A \neq 0_2$. De façon directe, on vérifie que les vecteurs propres de A sont les λe_1, $\lambda \in K^*$, où (e_1, e_2) est la base canonique de K^2. Il n'y a donc pas de base de K^2 formée de vecteurs propres de l'endomorphisme u canoniquement associé à A, et ainsi u et A ne sont pas diagonalisables.

Supposons que u soit diagonalisable. Le *diagonaliser*, c'est trouver une base $\mathcal{B}$ de E formée de vecteurs propres de u, *i.e.* telle que $\mathrm{Mat}_{\mathcal{B}}(u)$ soit diagonale. Ce problème se ramène à la détermination des espaces propres de u, *i.e.* à la résolution de systèmes linéaires. Diagonaliser une matrice $A \in M_n(K)$ c'est trouver $P \in GL_n(K)$ telle que $D = P^{-1}AP$ soit diagonale. Cela peut servir à calculer A^k, $k \in \mathbb{N}$, *cf.* l'exercice ci-après. En effet, d'une part le calcul de D^k est immédiat, et d'autre part $A = PDP^{-1}$, d'où $A^k = PD^kP^{-1}$ puisque $M \mapsto PMP^{-1}$ est un automorphisme de l'algèbre $M_n(K)$.

Exercice 4.

Calculer les puissances de la matrice $A := \begin{pmatrix} 2 & -1 & 1 \\ -4 & 2 & -4 \\ -2 & 1 & -1 \end{pmatrix} \in M_3(\mathbb{C})$.

Solution. Nous ne savons pas *a priori* que A est diagonalisable. Notons $\mathcal{B} := (e_1, e_2, e_3)$ la base canonique de $\mathbb{C}^3$ et $u \in \mathcal{L}(\mathbb{C}^3)$ l'endomorphisme canoniquement associé à A. Le polynôme caractéristique de A s'écrit $\chi_A = \begin{vmatrix} X-2 & 1 & -1 \\ 4 & X-2 & 4 \\ 2 & -1 & X+1 \end{vmatrix}$. Développons :

$$\begin{aligned}\chi_A &= (X-2)\begin{vmatrix} X-2 & 4 \\ -1 & X+1 \end{vmatrix} - \begin{vmatrix} 4 & 4 \\ 2 & X+1 \end{vmatrix} - \begin{vmatrix} 4 & X-2 \\ 2 & -1 \end{vmatrix} \\ &= (X-2)(X^2-X+2) - (4X-4) - (-2X) \\ &= X^3 - 3X^2 + 2X = X(X-1)(X-2).\end{aligned}$$

Ainsi A possède trois valeurs propres : $0, 1, 2$. Déterminons les espaces propres associés. Le système correspondant à la valeur propre 0 s'écrit :

$$\begin{pmatrix} -2 & 1 & -1 \\ 4 & -2 & 4 \\ 2 & -1 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \text{soit} \quad \begin{cases} -2x + y - z = 0 \\ 4x - 2y + 4z = 0 \\ 2x - y + z = 0\,. \end{cases}$$

On trouve la solution $(1, 2, 0)$: $v_1 := e_1 + 2e_2$ est un vecteur propre de u associé à 0. Le système correspondant à la valeur propre 1 s'écrit :

$$\begin{pmatrix} -1 & 1 & -1 \\ 4 & -1 & 4 \\ 2 & -1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \text{soit} \quad \begin{cases} -x + y - z = 0 \\ 4x - y + 4z = 0 \\ 2x - y + 2z = 0\,. \end{cases}$$

On trouve la solution $(-1, 0, 1)$: $v_2 := e_3 - e_1$ est un vecteur propre de u associé à 1. Le système correspondant à la valeur propre 2 s'écrit :

$$\begin{pmatrix} 0 & 1 & -1 \\ 4 & 0 & 4 \\ 2 & -1 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \text{soit} \quad \begin{cases} y - z = 0 \\ 4x + 4z = 0 \\ 2x - y + 3z = 0\,. \end{cases}$$

On trouve la solution $(1, -1, -1)$: $v_3 := e_1 - e_2 - e_3$ est un vecteur propre de u associé à 2. D'après le corollaire 14 de la page 225, $\mathcal{C} := (v_1, v_2, v_3)$ est une base de $\mathbb{C}^3$, donc A est diagonalisable et $\mathrm{Mat}_{\mathcal{C}}(u) = \mathrm{Diag}(0, 1, 2) := D$. La matrice de passage de $\mathcal{B}$ à $\mathcal{C}$ est $P := \begin{pmatrix} 1 & -1 & 1 \\ 2 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$. En exprimant les e_i à partir des v_j, on trouve : $P^{-1} = \begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 3 \\ 2 & -1 & 2 \end{pmatrix}$. Si $k \in \mathbb{N}$, $A^k = PD^kP^{-1}$ et $D^k = \mathrm{Diag}(0, 1, 2^k)$, d'où :

$$\begin{aligned}A^k &= \begin{pmatrix} 1 & -1 & 1 \\ 2 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2^k \end{pmatrix}\begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 3 \\ 2 & -1 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -1 & 2^k \\ 0 & 0 & -2^k \\ 0 & 1 & -2^k \end{pmatrix}\begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 3 \\ 2 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 2^{k+1} - 2 & 1 - 2^k & 2^{k+1} - 3 \\ -2^{k+1} & 2^k & -2^{k+1} \\ 2 - 2^{k+1} & 2^k - 1 & 3 - 2^{k+1} \end{pmatrix}.\end{aligned}$$

Supposons que u soit diagonalisable et $\operatorname{sp} u := \{\lambda_1, \dots, \lambda_r\}$. D'après le théorème 24 de la page 232, on en déduit une décomposition de E en somme directe :

$$E \;=\; E_{\lambda_1}(u) \oplus E_{\lambda_2}(u) \oplus \cdots \oplus E_{\lambda_r}(u). \tag{11}$$

Cette décomposition est appelée *décomposition spectrale*, et les projecteurs $\pi_1, \dots, \pi_r$ correspondants sont appelés *projecteurs spectraux* de u. Rappelons que les π_i vérifient les égalités suivantes : $\pi_1 + \pi_2 + \cdots + \pi_r = \mathrm{Id}_E$, $\pi_i^2 = \pi_i$ pour tout i, et $\pi_i \circ \pi_j = 0$ dès que $i \neq j$. Si $(a_1, \dots, a_r)$ et $(b_1, \dots, b_r)$ appartiennent à K^r, on en déduit l'égalité :

$$(a_1\pi_1 + \cdots + a_r\pi_r) \circ (b_1\pi_1 + \cdots + b_r\pi_r) \;=\; a_1b_1\pi_1 + \cdots + a_rb_r\pi_r.$$

Par ailleurs, on a :

$$u \;=\; \lambda_1\pi_1 + \lambda_2\pi_2 + \cdots + \lambda_r\pi_r. \tag{12}$$

En effet, pour tout $i \in [\![1, r]\!]$, les restrictions à $E_{\lambda_i}(u)$ des deux membres de cette égalité sont égales à $x \mapsto \lambda_i x$. Pour tout $k \in \mathbb{N}$, on en déduit (par récurrence) la formule :

$$u^k \;=\; \lambda_1^k\pi_1 + \lambda_2^k\pi_2 + \cdots + \lambda_r^k\pi_r. \tag{13}$$

Ces formules permettent de calculer les puissances de u, *à condition* de connaître les π_i. Par combinaisons linéaires, il en résulte en tous cas ceci :

$$Q(u) \;=\; Q(\lambda_1)\pi_1 + Q(\lambda_2)\pi_2 + \cdots + Q(\lambda_r)\pi_r \quad \text{pour tout} \quad Q \in K[X]. \tag{14}$$

Soit $i \in [\![1, r]\!]$. Pour calculer π_i (comme polynôme en u), il suffit de trouver un polynôme Q_i tel que $Q_i(\lambda_i) = 1$ et $Q_i(\lambda_j) = 0$ pour tout $j \neq i$, puisqu'alors on aura $\pi_i = Q_i(u)$. Le polynôme d'interpolation de Lagrange répond à la question. L'exercice suivant illustre cette méthode.

Exercice 5.

Calculer par cette méthode les puissances de la matrice A de l'exercice 4 de la page 232.

Solution. Notons u l'endomorphisme de $\mathbb{C}^3$ canoniquement associé à A et $\mathcal{B}$ la base canonique de $\mathbb{C}^3$. Nous savons que $\chi_u = X(X-1)(X-2)$. Ainsi $\operatorname{sp} u = \{\lambda_1, \lambda_2, \lambda_3\}$, où $\lambda_1 := 0$, $\lambda_2 := 1$, $\lambda_3 := 2$. En outre, le corollaire 14 de la page 225 montre que u est diagonalisable. Pour calculer les puissances de A, il faut calculer les *matrices spectrales* $\Pi_i := \mathrm{Mat}_{\mathcal{B}}(\pi_j)$ pour $j = 1, 2, 3$. Les polynômes d'interpolation de Lagrange Q_2, Q_3 sont donnés par les formules classiques (Q_1 ne sert pas) :

$$Q_2 := \frac{X(X-2)}{(1-0)(1-2)} = -X(X-2)\,, \;\; Q_3 := \frac{X(X-1)}{(2-0)(2-1)} = X(X-1)/2.$$

Le calcul donne alors Π_2 et Π_3 :

$$\Pi_2 = -A(A-2I_3) = \begin{pmatrix} -2 & 1 & -3 \\ 0 & 0 & 0 \\ 2 & -1 & 3 \end{pmatrix}, \;\; \Pi_3 = \frac{1}{2}A(A-I_3) = \begin{pmatrix} 2 & -1 & 2 \\ -2 & 1 & -2 \\ -2 & 1 & -2 \end{pmatrix}.$$

Pour tout entier $k \geqslant 1$, on a donc :

$$A^k \;=\; \begin{pmatrix} -2 & 1 & -3 \\ 0 & 0 & 0 \\ 2 & -1 & 3 \end{pmatrix} + 2^k \begin{pmatrix} 2 & -1 & 2 \\ -2 & 1 & -2 \\ -2 & 1 & -2 \end{pmatrix}.$$

Nous retrouvons donc les formules obtenues dans la solution de l'exercice 4 de la page 232, mais de façon beaucoup plus rapide !

La décomposition spectrale ci-dessus concernait un endomorphisme diagonalisable. Pour un endomorphisme scindé, une décomposition plus générale sera étudiée dans la section 3.1. Donnons maintenant un critère de diagonalisabilité très puissant. Dans la section suivante, nous donnerons un autre critère, faisant intervenir le polynôme caractéristique.

Théorème 25. Les conditions suivantes sont équivalentes :

1. L'endomorphisme u est diagonalisable.
2. Le polynôme minimal μ_u de u est *scindé* et à racines simples sur K.
3. Il existe $P \in K[X]$, scindé et à racines simples, tel que $P(u) = 0$ et $P \neq 0$.

Démonstration. Supposons u diagonalisable. Si $\lambda \in \mathrm{sp}(u)$, l'endomorphisme de $E_\lambda(u)$ induit par u est l'homothétie de rapport λ, de polynôme minimal $X - \lambda$. Puisque E est somme directe des $E_\lambda(u)$, $\lambda \in \mathrm{sp}(u)$, la proposition 18 de la page 227 montre que μ_u est le ppcm des $X - \lambda, \lambda \in \mathrm{sp}(u)$, *i.e.* leur produit. Ainsi μ_u est scindé à racines simples.
L'implication $2 \Longrightarrow 3$ est évidente, car $\mu_u(u) = 0$. Supposons que $P \in K[X] \setminus \{0\}$ soit scindé à racines simples et annule u. D'après le lemme des noyaux, E est somme directe des $V_\lambda := \mathrm{Ker}(u - \lambda\,\mathrm{Id}_E)$, λ décrivant l'ensemble R des racines de P. Si $\lambda \in R$, ou bien $\lambda \in \mathrm{sp}(u)$, auquel cas $V_\lambda = E_\lambda(u)$, ou bien $\lambda \notin \mathrm{sp}(u)$, auquel cas $V_\lambda = \{0\}$. Il en résulte que E est somme des $E_\lambda(u)$ lorsque λ décrit $R \cap \mathrm{sp}(u)$, *a fortiori* E est somme des $E_\lambda(u)$, $\lambda \in \mathrm{sp}(u)$, donc u est diagonalisable. ∎

Corollaire 26. Supposons que u soit diagonalisable. Pour tout sous-espace V de E stable par u, l'endomorphisme induit $u_V \in \mathcal{L}(V)$ est diagonalisable.

Démonstration. D'après la proposition 16 de la page 226, μ_{u_V} divise μ_u. Le théorème précédent montre que μ_u est scindé et à racines simples, donc μ_{u_V} a la même propriété. Toujours d'après le théorème, u_V est donc diagonalisable. ∎

Corollaire 27. Soit $A \in M_n(K)$. Les conditions suivantes sont équivalentes :

1. La matrice A est diagonalisable.
2. Le polynôme minimal μ_A de A est *scindé* et à racines simples.
3. Il existe un polynôme non nul, scindé et à racines simples, qui annule A.

Démonstration. Soit $u \in \mathcal{L}(K^n)$ l'endomorphisme canoniquement associé à A. Puisque $\chi_u = \chi_A$ et $\mu_A = \mu_u$, la conclusion vient de la proposition 23 de la page 231 et du théorème précédent appliqué à u. ∎

Exercice 6.

Montrer que toute matrice $A \in GL_n(\mathbb{C})$ d'ordre (multiplicatif) fini est diagonalisable.

Solution. Soit $p \in \mathbb{N}^*$ l'ordre de A, de sorte que $A^p = I_n$. Le polynôme $P := X^p - 1$ annule donc A. Ce polynôme est scindé (comme tous les polynômes sur $\mathbb{C}$), et il est à racines simples : ses racines sont les p racines $p^{èmes}$ de l'unité dans $\mathbb{C}$. Il suffit alors d'appliquer le corollaire 27 de la page précédente.

Soit $A \in M_n(K)$. Les valeurs propres de A sont les racines du polynôme χ_A, qui est de degré n, donc A possède *au plus* n valeurs propres. Étudions le cas limite :

Proposition 28. Si u possède n valeurs propres distinctes, il est diagonalisable. De même toute matrice $A \in M_n(K)$ ayant n valeurs propres distinctes est diagonalisable.

Démonstration. Supposons que u possède n valeurs propres distinctes $\lambda_1, \ldots, \lambda_n$, avec $\dim E = n$. Pour tout i, soit e_i un vecteur propre de u associé à λ_i. Les e_i sont linéairement indépendants (corollaire 14 de la page 225), donc forment une base de E. Le théorème 24 de la page 232 montre alors que u est diagonalisable. L'énoncé matriciel s'en déduit. ∎

La réciproque est fausse : si $n \geqslant 2$, la matrice nulle de $M_n(K)$ est diagonalisable mais 0 est sa seule valeur propre. La proposition ne donne donc qu'une condition *suffisante* pour qu'une matrice soit diagonalisable. Cette condition est cependant utile : nous verrons plus loin que la « plupart » des matrices de $M_n(\mathbb{C})$ possèdent n valeurs propres distinctes.

Exercice 7.

Montrer que la matrice $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \in M_3(\mathbb{Q})$ est diagonalisable.

Solution. Calculons χ_A :

$$\chi_A = \begin{vmatrix} X-1 & -1 & 0 \\ 0 & X-2 & -1 \\ 0 & 0 & X-3 \end{vmatrix} = (X-1)(X-2)(X-3),$$

car le déterminant d'une matrice triangulaire est le produit des termes diagonaux de cette matrice. Ainsi χ_A a trois racines, à savoir $1, 2, 3$, donc A a trois valeurs propres distinctes. On applique alors la proposition. Noter que nous n'avons pas diagonalisé A.

Exercice 8.

Soit $A := \begin{pmatrix} 1 & 7 & -7 \\ 4 & 22 & -23 \\ 4 & 14 & -15 \end{pmatrix}$. Combien y a-t-il de matrices $M \in M_3(\mathbb{C})$ vérifiant l'équation $M^3 = A$? Expliciter l'une de ces matrices.

Solution. Le calcul donne : $\chi_A = X^3 - 8X^2 - X + 8 = (X-1)(X+1)(X-8)$. Ainsi A a trois valeurs propres, à savoir $1, -1$ et 8. Elle est donc diagonalisable, plus précisément elle

est semblable à $D := \mathrm{Diag}(1,-1,8)$. Il existe donc $P \in GL_3(\mathbb{C})$ telle que $A = PDP^{-1}$. Posons $S := \{M \in M_3(\mathbb{C}) \mid M^3 = A\}$ et $T := \{N \in M_3(\mathbb{C}) \mid N^3 = D\}$. Puisque $N \mapsto PNP^{-1}$ est un automorphisme de l'algèbre $M_3(\mathbb{C})$ appliquant D sur A, sa restriction à T est une bijection de T sur S. Il nous suffit donc de résoudre l'équation $N^3 = D$. Posons $j := \exp(2\pi i/3)$; les racines cubiques de l'unité sont $1, j, j^2$. Si $a, b, c \in \{0,1,2\}$ la matrice $N := \mathrm{Diag}(j^a, -j^b, 2j^c)$ est évidemment solution de l'équation $N^3 = D$. On obtient ainsi $3^3 = 27$ solutions de cette équation. Montrons que ce sont les seules.

Soit donc $N \in T$. Notons u (resp. v) l'endomorphisme de $\mathbb{C}^3$ canoniquement associé à D (resp. N), de sorte que $v^3 = u$. Notant $\mathcal{B} := (e_1, e_2, e_3)$ la base canonique de $\mathbb{C}^3$, u possède trois droites propres : $\mathbb{C}e_1, \mathbb{C}e_2, \mathbb{C}e_3$. Le point crucial est que v (que l'on ne connaît pas) commute avec $u = v^3$. D'après la proposition 11 de la page 224, v laisse stable chacune des droites propres de u, ce qui signifie que la matrice N est diagonale. Ainsi $N := \mathrm{Diag}(x, y, z)$, avec $x^3 = 1, y^3 = -1, z^3 = 8$. Il en résulte que N est l'une des 27 matrices ci-dessus. Finalement $\mathrm{card}\,(T) = 27$, et donc $\mathrm{card}\,(S) = 27$.

Pour expliciter un élément de S, il suffit de diagonaliser A, *i.e.* de trouver $P \in GL_3(\mathbb{C})$ telle que $A = PDP^{-1}$. La matrice $M := P\,\mathrm{Diag}(1,-1,2)P^{-1}$ répondra à la question. Soit f l'endomorphisme de $\mathbb{C}^3$ canoniquement associé à A. Le calcul montre que les vecteurs $\varepsilon_1 := e_1 + 2e_2 + 2e_3$, $\varepsilon_2 := e_2 + e_3$, $\varepsilon_3 := e_1 + 3e_2 + 2e_3$ sont des vecteurs propres de f, associés aux valeurs propres $1, -1, 8$ respectivement. Ils forment donc une base $\mathcal{C}$ de $\mathbb{C}^3$. Il suffit de prendre pour P la matrice de passage de $\mathcal{B}$ à $\mathcal{C}$, c'est-à-dire $P := \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 3 \\ 2 & 1 & 2 \end{pmatrix}$. On vérifie que $P^{-1} = \begin{pmatrix} 1 & -1 & 1 \\ -2 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}$. D'où une solution de l'équation $M^3 = A$, à savoir :

$$\begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 3 \\ 2 & 1 & 2 \end{pmatrix} \mathrm{Diag}(1,-1,2) \begin{pmatrix} 1 & -1 & 1 \\ -2 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 2 \\ 2 & -1 & 6 \\ 2 & -1 & 4 \end{pmatrix} \begin{pmatrix} 1 & -1 & 1 \\ -2 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix},$$

soit $M := \begin{pmatrix} 1 & 1 & -1 \\ 4 & 4 & -5 \\ 4 & 2 & -3 \end{pmatrix}$.

L'exercice suivant fait apparaître l'intérêt des sous-espaces stables.

Exercice 9.

Soit $(u_i)_{i \in I}$ une famille d'endomorphismes de E. On suppose que les u_i commutent deux à deux et que chacun d'eux est diagonalisable. Montrer que ces endomorphismes sont *simultanément diagonalisables*, c'est-à-dire qu'il existe une base de E dans laquelle la matrice de chaque u_i est diagonale. On pourra utiliser la proposition 11 de la page 224.

Solution. Supposons que $I := [\![1, p]\!]$, où $p \in \mathbb{N}^*$. Raisonnons par récurrence sur p. Le cas $p = 1$ est clair.
Supposons $p \geqslant 2$, la conclusion étant vraie à l'ordre $p-1$. Soient $u_1, \ldots, u_p$ des endomorphismes diagonalisables de E commutant deux à deux. Notons $V_1, \ldots, V_r$ les espaces propres de u_p. Puisque u_p est diagonalisable, $E = V_1 \oplus V_2 \oplus \ldots \oplus V_r$.
Soient $k \in [\![1, r]\!]$ et $i \in [\![1, p-1]\!]$. Puisque u_i commute avec u_p, il laisse stable V_k (proposition 11 de la page 224). D'après le corollaire 26 de la page 235, l'endomorphisme $u_{i,k}$ induit par u_i sur V_k est un endomorphisme diagonalisable de V_k. Si k est fixé, $u_{1,k}, \ldots, u_{p-1,k}$ sont des endomorphismes diagonalisables de V_k commutant deux à deux. D'après l'hypothèse de récurrence, il existe une base $\mathcal{B}_k$ de V_k telle que, pour tout $i \in [\![1, p-1]\!]$, la matrice $A_{i,k}$ de $u_{i,k}$ dans cette base soit diagonale. Soit $\mathcal{B}$ la base de E obtenue par juxtaposition des $\mathcal{B}_k$. Pour tout $i \in [\![1, p-1]\!]$, la matrice de u_i dans la base $\mathcal{B}$ est la matrice diagonale par blocs $\mathrm{Diag}(A_{i,1}, \ldots, A_{i,r})$, elle est diagonale puisque les $A_{i,k}$ sont diagonales. Pour tout $k \in [\![1, r]\!]$, posons $d_k := \dim(V_k)$ et soit λ_k le scalaire tel que $V_k = \mathrm{Ker}(u_p - \lambda_k \mathrm{Id}_E)$. La matrice de u_p dans la base $\mathcal{B}$ est la matrice diagonale par blocs $\mathrm{Diag}(\lambda_1 I_{d_1}, \ldots, \lambda_r I_{d_r})$, elle est aussi diagonale. D'où la conclusion lorsque $I = [\![1, p]\!]$, et donc lorsque I est fini.
Supposons I infini. Le sous-espace vectoriel T de $\mathcal{L}(E)$ engendré par les u_i est de dimension finie, parce que $\mathcal{L}(E)$ l'est. Il existe donc une partie finie J de I tel que les $u_j, j \in J$ engendrent T. D'après le cas précédent, il existe une base $\mathcal{B}$ de E dans laquelle la matrice de chaque $u_j, j \in J$ soit diagonale. Soit $i \in I \setminus J$. Puisque u_i est combinaison linéaire des $u_j, j \in J$, sa matrice dans la base $\mathcal{B}$ est encore diagonale, comme combinaison linéaire de matrices diagonales. La base $\mathcal{B}$ répond donc à la question.

2.4 Endomorphismes et matrices trigonalisables

Le déterminant d'une matrice triangulaire est le produit de ses termes diagonaux, mais le calcul des puissances d'une matrice triangulaire n'est pas toujours immédiat. Les matrices triangulaires, et les endomorphismes ou matrices trigonalisables qui sont définis ci-dessous, jouent cependant un rôle important, notamment théorique.

Définition 7. L'endomorphisme u est dit *trigonalisable* s'il existe une base de E dans laquelle la matrice de u soit triangulaire supérieure. Une matrice $A \in M_n(K)$ est dite *trigonalisable* si elle est semblable à une matrice triangulaire supérieure.

Toute matrice diagonale étant triangulaire, il est évident que « diagonalisable » implique « trigonalisable ». L'exemple de la page 232 montre que la réciproque est fausse. Voici le principal critère de trigonalisabilité :

Théorème 29. L'endomorphisme u est trigonalisable si, et seulement s'il est scindé. De même la matrice A est trigonalisable si, et seulement si, elle est scindée.

Démonstration. Les deux assertions du théorème sont équivalentes, nous laissons le lecteur le vérifier. Supposons que $A \in M_n(K)$ soit trigonalisable. Il existe $P \in GL_n(K)$

telle que $B := P^{-1}AP$ soit triangulaire supérieure : $B := (a_{i,j})$, avec $a_{i,j} = 0$ dès que $i > j$. Puisque A et B sont semblables, $\chi_A = \chi_B$. Par ailleurs (déterminant d'une matrice triangulaire), $\chi_B = \prod_{i=1}^{n}(X - a_{i,i})$, donc χ_A est scindé, *i.e.* A est scindée.

La réciproque est moins évidente. Montrons par récurrence sur l'entier $m \geqslant 1$ que tout endomorphisme *scindé* v d'un espace vectoriel F de dimension m est trigonalisable. Le cas $m = 1$ est clair. Supposons le résultat vrai à l'ordre $m-1$ ($m \geqslant 2$). Soient F un espace vectoriel de dimension m et v un endomorphisme scindé de F. En particulier, χ_v possède au moins une racine λ_m, qui est une valeur propre de v. Soit e_m un vecteur propre de v associé à λ_m. Complétons la famille libre (e_m) en une base $\mathcal{B} := (e_1, \ldots, e_m)$ de F. Puisque $v(e_m) = \lambda_m e_m$, la matrice $A := \mathrm{Mat}_{\mathcal{B}}(v)$ s'écrit par blocs : $A := \begin{pmatrix} M & U \\ 0 & \lambda_m \end{pmatrix}$, où $M \in M_{m-1}(K)$ et $U \in M_{m-1,1}(K)$. Le calcul d'un déterminant par blocs montre que $\chi_A = (X - \lambda_m)\chi_M$. Le polynôme χ_M est donc scindé, puisqu'il divise χ_A. Vu l'hypothèse de récurrence, il existe $Q \in GL_{m-1}(K)$ telle que $T := Q^{-1}MQ$ soit triangulaire supérieure. La matrice $P := \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \in M_m(K)$ est inversible, son inverse étant $\begin{pmatrix} Q^{-1} & 0 \\ 0 & 1 \end{pmatrix}$. On en déduit que $P^{-1}AP$ est triangulaire, et donc que v est trigonalisable. En effet, $P^{-1}AP$ vaut :

$$P^{-1}\begin{pmatrix} M & U \\ 0 & \lambda_m \end{pmatrix}\begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} Q^{-1} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} MQ & U \\ 0 & \lambda_m \end{pmatrix}$$
$$= \begin{pmatrix} Q^{-1}MQ & Q^{-1}U \\ 0 & \lambda_m \end{pmatrix},$$

et la dernière matrice est triangulaire supérieure parce que $Q^{-1}MQ$ l'est. ∎

Corollaire 30. Sur un corps algébriquement clos (par exemple $\mathbb{C}$), toute matrice carrée (et tout endomorphisme en dimension finie) est trigonalisable.

Démonstration. Cela vient du fait que *tout* polynôme non constant à coefficients dans un corps algébriquement clos est scindé. ∎

Exemple. Soit $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{R})$. À quelle condition A est-elle trigonalisable, ou diagonalisable, sur $\mathbb{C}$ ou sur $\mathbb{R}$?

Le polynôme $\chi_A = X^2 - (a+d)X + (ad - bc)$ a pour discriminant :

$$\Delta := (a+d)^2 - 4(ad - bc) = (a-d)^2 + 4bc.$$

Sur $\mathbb{C}$, A est toujours trigonalisable. Si $\Delta < 0$, χ_A a deux racines complexes conjuguées non réelles, donc distinctes, et par suite A est diagonalisable sur $\mathbb{C}$. Puisque χ_A n'a pas de racine réelle, A n'est pas trigonalisable sur $\mathbb{R}$. Si $\Delta > 0$, χ_A a deux racines réelles

distinctes, donc A est diagonalisable sur $\mathbb{R}$, *a fortiori* sur $\mathbb{C}$. Supposons enfin que $\Delta = 0$. Dans ce cas, χ_A a une racine double $\lambda \in \mathbb{R}$; ainsi A est scindée, *i.e.* trigonalisable, sur $\mathbb{R}$. Il reste à étudier, dans ce cas, la diagonalisabilité de A. Puisque $\chi_A = (X-\lambda)^2$, la seule matrice diagonale qui peut être semblable à A est λI_2. Ainsi A est diagonalisable si $A = \lambda I_2$ et non diagonalisable si $A \neq \lambda I_2$ (que ce soit sur $\mathbb{R}$ ou sur $\mathbb{C}$).

Remarque. Tout endomorphisme annulé par un polynôme *scindé* est trigonalisable. Pour le voir, adaptons la preuve du théorème précédent. Reprenant les notations de cette preuve, supposons que $v \in \mathcal{L}(F)$ soit annulé par $P := \prod_{j=1}^{p}(X-\alpha_j)$, de sorte que $(v-\alpha_1 \operatorname{Id}_F)\circ(v-\alpha_2 \operatorname{Id}_F)\circ\cdots(v-\alpha_p \operatorname{Id}_F) = 0$. Puisque $F \neq \{0\}$, l'endomorphisme nul n'est pas inversible, donc l'un des $v - \alpha_j \operatorname{Id}_F$ n'est pas inversible. Nous pouvons ainsi supposer, par exemple, que α_p est une valeur propre de v. Comme dans ladite preuve, on en déduit une écriture par blocs $A = \begin{pmatrix} M & U \\ 0 & \alpha_p \end{pmatrix}$. L'important est d'observer alors que $P(A) = \begin{pmatrix} P(M) & * \\ 0 & P(\alpha_p) \end{pmatrix}$, d'où $P(M) = 0$. Ainsi M est annulée par un polynôme scindé, elle est donc trigonalisable, d'après l'hypothèse de récurrence. On conclut comme précédemment.

Supposons que u soit *scindé*. Le polynôme caractéristique χ_u s'écrit donc :

$$\chi_u := (X-\lambda_1)^{n_1}(X-\lambda_2)^{n_2}\cdots(X-\lambda_r)^{n_r}, \tag{15}$$

où $r \in \mathbb{N}^*$ et $\lambda_1, \ldots, \lambda_r \in K$ sont distincts, les n_i étant des entiers strictement positifs. Ainsi $\operatorname{sp} u = \{\lambda_1, \ldots, \lambda_r\}$ et les n_i sont les multiplicités des valeurs propres λ_i. Puisque $\deg(\chi_u) = n$, la somme des n_i vaut n. Vu l'égalité (15), le coefficient constant de χ_u est le produit des $(-\lambda_i)^{n_i}$, et le coefficient de X^{n-1} dans χ_u est $-\sum_{i=1}^{r} n_i\lambda_i$ (développer le second membre de (15)). La comparaison avec l'égalité (7) de la page 223 donne ceci :

Proposition 31. Avec les hypothèses et notations ci-dessus, on a les égalités :

$$n = \sum_{i=1}^{r} n_i\,, \qquad \det(u) = \prod_{i=1}^{r} \lambda_i^{n_i}\,, \quad \text{et} \quad \operatorname{Tr}(u) = \sum_{i=1}^{r} n_i\lambda_i. \tag{16}$$

Ainsi le déterminant (resp. la trace) de u est le produit (resp. la somme) des valeurs propres de u, comptées avec leurs multiplicités.

Le polynôme minimal μ_u divise χ_u et a les mêmes racines que χ_u, d'où :

$$\mu_u = (X-\lambda_1)^{k_1}(X-\lambda_2)^{k_2}\cdots(X-\lambda_r)^{k_r}, \tag{17}$$

avec $1 \leqslant k_i \leqslant n_i$ pour $i = 1, \ldots, r$. D'après le théorème 25 de la page 235, u est diagonalisable si, et seulement si, tous les k_i valent 1.

Si u est diagonalisable, il est trigonalisable et donc scindé. Cependant un endomorphisme scindé n'est pas nécessairement diagonalisable, comme le montrent l'exemple de

la page 232 ou le théorème 25 de la page 235. Comme application du théorème précédent, voici un important critère de diagonalisabilité. La preuve en est simple (à ce stade) mais instructive.

Théorème 32. L'endomorphisme u est diagonalisable si, et seulement s'il est scindé et si la multiplicité de chaque valeur propre est égale à la dimension de l'espace propre associé.

Démonstration. Si u est diagonalisable, il est *a fortiori* trigonalisable et donc scindé. Nous pouvons donc supposer que χ_u est scindé et factorisé comme en (15). Pour tout i, le théorème 17 de la page 226 donne : $d_i := \dim\big(E_{\lambda_i}(u)\big) \leqslant n_i$. La somme V des $E_{\lambda_i}(u)$ est directe, donc $\dim(V) = \sum_{i=1}^{r} d_i$. Par ailleurs, $n = \sum_{i=1}^{r} n_i$. D'après le théorème 24 de la page 232, u est diagonalisable si, et seulement si, $V = E$, ce qui équivaut à $\sum_{i=1}^{r}(n_i - d_i) = 0$ (*cf.* la proposition 3 de la page 216). Comme $d_i \leqslant n_i$ pour tout i, cette égalité est vraie si, et seulement si, $d_i = n_i$ pour tout i. ∎

Dans l'exemple de la page 232, il y a une seule valeur propre, à savoir 0, et la dimension de l'espace propre associé est 1. On retrouve ainsi le fait que A ne soit pas diagonalisable.
Comme autre application du théorème 29 de la page 238, indiquons une méthode de calcul du polynôme caractéristique d'une matrice $A \in M_n(\mathbb{C})$. Certes, $\chi_A = \det(XI_n - A)$ est un déterminant. Mais le calcul de ce déterminant par une méthode du type « pivot de Gauß » (*cf.* le cours de L1) est malaisé, car il faut travailler dans le corps de fractions rationnelles $\mathbb{C}(X)$. Une méthode plus efficace, la *méthode de Faddeev-Le Verrier*, est basée sur la proposition ci-après. Développons χ_A :

$$\chi_A \;=\; X^n + a_1X^{n-1} + a_2X^{n-2} + \cdots + a_{n-1}X + a_n. \tag{18}$$

Il s'agit donc de calculer $a_1, \ldots, a_n$.

Proposition 33. Définissons une suite $(u_k)_{k\geqslant 1}$ de nombres complexes et une suite $(M_k)_{k\geqslant 1}$ d'éléments de $M_n(\mathbb{C})$ en posant $M_1 := A$ et, pour tout entier $k \geqslant 1$:

$$u_k := \frac{\mathrm{Tr}(M_k)}{k} \quad \text{et} \quad M_{k+1} := (M_k - u_kI_n)A. \tag{19}$$

Alors $a_k = -u_k$ pour $k = 1, \ldots, n$.

Démonstration. Elle utilise les *formules de Newton*, *cf.* le module II.5 de [L1] ou le module I.7 de ce volume (théorème 15 de la page 332). Prouvons d'abord, par récurrence sur l'entier $k \geqslant 1$, la formule :

$$M_k \;=\; A^k - \sum_{j=1}^{k-1} u_{k-j}A^j \tag{20}$$

Le cas $k = 1$ est clair, car $M_1 = A$. Supposons que la formule (20) soit vraie pour un entier $k \geqslant 1$. La seconde formule (19) donne, en effectuant le changement d'in-

dice $h := j+1$:

$$M_{k+1} = \left(A^k - \sum_{j=1}^{k-1} u_{k-j}A^j\right)A - u_kA = A^{k+1} - \sum_{h=1}^{k} u_{k+1-h}A^h,$$

d'où la formule (20) à l'ordre $k+1$.
Pour tout $h \in \mathbb{N}$, posons $s_h := \mathrm{Tr}(A^h)$. Soit $k \in \mathbb{N}^*$. Dans l'égalité (20) à l'ordre k, égalons les traces des deux membres. Compte tenu de la première formule (19), il vient :

$$ku_k = s_k - \sum_{j=1}^{k-1} u_{k-j}s_j. \tag{21}$$

Calculons ensuite les s_k en fonction des valeurs propres de A. Le corps de base étant $\mathbb{C}$, la matrice A est semblable à une matrice triangulaire supérieure T, dont les termes diagonaux seront notés $\lambda_1, \dots, \lambda_n$. Alors $\chi_A = \prod_{j=1}^{n}(X - \lambda_j)$. Soit $k \in \mathbb{N}^*$. La matrice A^k est semblable à T^k, qui est une matrice triangulaire supérieure, de termes diagonaux $\lambda_1^k, \dots, \lambda_n^k$. Ainsi $s_k = \mathrm{Tr}(A^k) = \mathrm{Tr}(T^k) = \sum_{j=1}^{n} \lambda_j^k$ est la $k^{ème}$ somme de Newton de $\lambda_1, \dots, \lambda_n$. Les formules de Newton peuvent s'écrire ainsi, pour $k = 1, \dots, n$:

$$-ka_k = s_k + \sum_{j=1}^{k-1} a_j s_{k-j} = s_k + \sum_{j=1}^{k-1} a_{k-j}s_j.$$

Ainsi les suites $(u_k)_{1\leqslant k\leqslant n}$ et $(-a_k)_{1\leqslant k\leqslant n}$ vérifient la même relation de récurrence. Puisque $u_1 = \mathrm{Tr}(A) = -a_1$, ces deux suites sont égales. ∎

Le calcul des coefficients de χ_A se ramène donc à des opérations « rationnelles », à savoir produits de matrices et calculs de traces. Noter que les λ_j ne servent que dans la preuve ci-dessus, et pas dans le calcul effectif de χ_A.

3 Réduction : résultats généraux

3.1 Sous-espaces caractéristiques

Supposons que u soit *scindé*. Nous allons étudier une décomposition de E généralisant la décomposition spectrale, donnée dans la section 2.3 dans le cas où u est diagonalisable.
En notant $\{\lambda_1, \dots, \lambda_r\}$ le spectre de u, partons de la factorisation de χ_u :

$$\chi_u := (X-\lambda_1)^{n_1}(X-\lambda_2)^{n_2}\cdots(X-\lambda_r)^{n_r}. \tag{22}$$

À chaque valeur propre λ_i de u est associé l'espace propre $E_{\lambda_i}(u)$. Il s'avère qu'un autre sous-espace va jouer un rôle important :

Définition 8. Avec les hypothèses et notations ci-dessus, soit $i \in [\![1, r]\!]$. Le noyau de $(u - \lambda_i \mathrm{Id}_E)^{n_i}$ est appelé *sous-espace caractéristique* de u associé à la valeur propre λ_i, on le notera $E^{\lambda_i}(u)$. Ainsi $E^{\lambda_i}(u) := \mathrm{Ker}\big((u - \lambda_i \mathrm{Id}_E)^{n_i}\big)$.

Chaque sous-espace caractéristique $E^{\lambda_i}(u) = \mathrm{Ker}\big((u-\lambda_i \mathrm{Id}_E)^{n_i}\big)$ contient le sous-espace propre correspondant $E_{\lambda_i}(u) = \mathrm{Ker}(u-\lambda_i \mathrm{Id}_E)$. En outre, ce sous-espace est stable par u, en vertu de la proposition 11 de la page 224.

Théorème et définition 34. Conservons les hypothèses et notations ci-dessus.

1. Pour tout $i \in [\![1, r]\!]$, la dimension de $E^{\lambda_i}(u)$ est égale à n_i.
2. L'espace E est somme directe des sous-espaces caractéristiques de u :

$$E \;=\; E^{\lambda_1}(u) \oplus E^{\lambda_2}(u) \oplus \cdots \oplus E^{\lambda_r}(u). \tag{23}$$

Les projecteurs $\pi_1, \ldots, \pi_r \in \mathcal{L}(E)$ associés à cette décomposition de E en somme directe sont appelés *projecteurs spectraux* de u, ce sont des polynômes en u.

Démonstration. Puisque les λ_i sont distincts, les polynômes $(X-\lambda_i)^{n_i}$ sont premiers entre eux deux à deux. Le produit de ces polynômes étant χ_u, qui annule u en vertu du théorème de Cayley-Hamilton, l'assertion 2 résulte du lemme des noyaux, qui prouve aussi que les π_i sont des polynômes en u.

Pour tout i, notons u_i l'endomorphisme de $E^{\lambda_i}(u)$ induit par u. Par définition même, $(u_i - \lambda_i \mathrm{Id})^{n_i} = 0$. Ainsi $(X-\lambda_i)^{n_i}$ annule u_i, donc μ_{u_i} est une puissance de $X-\lambda_i$. En particulier μ_{u_i} est scindé. La remarque de la page 240 montre alors que $\chi_{u_i} = (X-\lambda_i)^{d_i}$, où $d_i := \dim\big(E^{\lambda_i}(u)\big)$ (noter que λ_i est la seule racine de χ_{u_i}). Vu la décomposition (23) et la proposition 18 de la page 227, χ_u est le produit des $\chi_{u_i} = (X-\lambda_i)^{d_i}$. En comparant avec (22), il vient $d_i = n_i$ pour $i = 1, \ldots, r$. ∎

Cette décomposition généralise effectivement la décomposition spectrale précédente. Pour tout i, on a en effet $\dim E_{\lambda_i}(u) \leqslant n_i = \dim E^{\lambda_i}(u)$. D'après le théorème 32 de la page 241, u est donc diagonalisable si, et seulement si, $E_{\lambda_i}(u) = E^{\lambda_i}(u)$ pour tout i.

Indiquons une méthode de calcul des projecteurs spectraux π_i, généralisant celle que nous avons donnée dans le cas diagonalisable. On part d'un polynôme (non nul) annulant u. Pour alléger les notations, choisissons χ_u, mais μ_u conviendrait aussi. La décomposition de la fraction rationnelle $1/\chi_u$ en éléments simples conduit à une égalité de la forme :

$$\frac{1}{\chi_u} \;=\; \sum_{i=1}^{r} \frac{B_i}{(X-\lambda_i)^{n_i}}, \tag{24}$$

avec $\deg B_i \leqslant n_i - 1$ pour tout i. Si $i \in [\![1, r]\!]$, notons T_i le produit des $(X-\lambda_j)^{n_j}$ pour $j \neq i$, de sorte que $\chi_u = (X-\lambda_i)^{n_i} T_i$. Multiplions l'égalité précédente par χ_u :

$$1 \;=\; \sum_{i=1}^{r} B_i(X) T_i(X). \tag{25}$$

Proposition 35. Conservons les hypothèses et notations ci-dessus. Pour tout $i \in [\![1, r]\!]$, le projecteur spectral π_i est donné par la formule : $\pi_i = (B_i T_i)(u)$.

Démonstration. Procédons de manière indirecte. *Définissons* en fait chaque π_i par la formule de l'énoncé, et posons $V_i := \mathrm{Im}\,\pi_i$. En substituant u à X dans l'égalité (25), on

voit que la somme des π_i est Id_E. Soient ensuite $i, j \in [\![1, r]\!]$, $i \neq j$. Le polynôme $T_i T_j$ est multiple de χ_u (le vérifier), il annule donc u, et ainsi $\pi_i \circ \pi_j = 0$. L'exercice 1 de la page 215 montre alors que $E = V_1 \oplus V_2 \oplus \cdots \oplus V_r$, et que les π_i sont les projecteurs associés à cette décomposition. Il suffit donc de vérifier que $V_i = E^{\lambda_i}(u)$, ou encore (pourquoi ?) $V_i \subset E^{\lambda_i}(u)$ pour tout i. Comme $E^{\lambda_i}(u)$ est le noyau de $(u - \lambda_i \mathrm{Id}_E)^{n_i}$ et V_i est l'image de π_i, cela revient à voir que $(u - \lambda_i \mathrm{Id}_E)^{n_i} \circ \pi_i = 0$. Cela résulte de l'égalité $\chi_u(u) = 0$: $\big((X - \lambda_i)^{n_i} B_i T_i\big)(u) = (B_i \chi_u)(u) = B_i(u) \circ \chi_u(u) = 0$. ∎

Exercice 10.

À l'aide de la méthode précédente, calculer les *matrices spectrales* de la matrice $A \in M_3(\mathbb{C})$ de l'exemple de la page 230. Par définition, ces matrices sont les matrices, dans la base canonique de $\mathbb{C}^3$, de l'endomorphisme de $\mathbb{C}^3$ canoniquement associé à A.

Solution. Nous avons vu que $\chi_A = (X + 1)(X - 1)^2$. Posons $\lambda_1 := -1$ et $\lambda_2 := 1$. La décomposition de $1/\chi_A$ en éléments simples est (le vérifier) :

$$\frac{1}{(X+1)(X-1)^2} = \frac{1}{4}\left[\frac{1}{X+1} + \frac{2}{(X-1)^2} - \frac{1}{X-1}\right] = \frac{1}{4}\left[\frac{1}{X+1} + \frac{3-X}{(X-1)^2}\right].$$

D'où $B_1 = 1/4$, $B_2 = (3 - X)/4$ et $T_1 = (X - 1)^2$, $T_2 = X + 1$. Les matrices spectrales Π_1, Π_2 associées à λ_1, λ_2 respectivement sont donc :

$$\Pi_1 := (A - I_3)^2/4 = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; \Pi_2 = (A + I_3)(3I_3 - A)/4 = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

À titre de vérification, on peut constater que $\Pi_1 + \Pi_2 = I_3$.

Comme illustration de la notion de sous-espace caractéristique, intéressons-nous aux suites définies par une relation de récurrence linéaire. On se donne un entier $k \geqslant 1$ et des nombres complexes $a_0, \ldots, a_{k-1}, a_0 \neq 0$. Les suites (u_n) vérifiant la relation de récurrence linéaire d'ordre k :

$$u_{n+k} := a_{k-1} u_{n+k-1} + \cdots + a_1 u_{n+1} + a_0 u_n \tag{26}$$

forment un sous-espace vectoriel E de $\mathbb{C}^{\mathbb{N}}$, espace des suites à termes complexes. Puisque $a_0 \neq 0$, la relation de récurrence (26) est vraiment d'ordre k. Lorsque $k := 2$, ces suites ont été étudiées dans le module IV.6 de [L1]. Cherchons à expliciter une base de E.

Indiquons une première méthode, basée sur une remarque simple : si $t_0, \ldots, t_{k-1} \in \mathbb{C}$ sont donnés, il existe une unique suite $(u_n)_{n\in\mathbb{N}} \in E$ telle que $u_0 := t_0, \ldots, u_{k-1} := t_{k-1}$. On dit que la suite (u_n) est définie par la relation de récurrence (26) et les *conditions initiales* $u_0 := t_0, \ldots, u_{k-1} := t_{k-1}$. En d'autres termes, $(u_n) \mapsto (u_0, \ldots, u_{k-1})$ est un isomorphisme φ de E sur $\mathbb{C}^k$. Cela montre déjà que E est de dimension finie, à savoir k. De plus, $(e_1, \ldots, e_k)$ désignant la base canonique de $\mathbb{C}^k$, posons $\varepsilon_i := \varphi^{-1}(e_{i+1})$ pour $i = 0, \ldots, k - 1$. Alors $\mathcal{B} := (\varepsilon_0, \ldots, \varepsilon_{k-1})$ est une base de E. Par exemple ε_0 est la seule suite $(u_n) \in E$ telle que $u_0 = 1$ et $u_1 = u_2 = \cdots = u_{k-1} = 0$. Mais ε_0 n'est pas connue « explicitement » : il n'y a pas de formule évidente donnant u_n en fonction de n.

Voici une autre méthode. Notons $T \in \mathcal{L}(\mathbb{C}^{\mathbb{N}})$ l'opérateur de « décalage » : pour toute suite $u := (u_n)_{n\geqslant 0}$, on pose $T(u) := (u_{n+1})_{n\geqslant 0}$.
Évidemment $T^j(u) = (u_{n+j})_{n\geqslant 0}$ pour tout $j \in \mathbb{N}$. Si $Q := b_r X^r + \cdots + b_1 X + b_0 \in \mathbb{C}[X]$, on en déduit que $Q(u)$ est la suite $(b_r u_{n+r} + \cdots + b_1 u_{n+1} + b_0 u_n)_{n\geqslant 0}$. En posant $P := X^k - (a_{k-1}X^{k-1} + \cdots + a_1 X + a_0)$, il en résulte que E *est le noyau de* $P(T)$. En particulier, E est stable par T ; notons $D \in \mathcal{L}(E)$ l'endomorphisme induit par T.
Déterminons la matrice A de D dans la base $\mathcal{B}$. Soit $j \in [\![0, k-1]\!]$, écrivons $\varepsilon_j := (u_n)$ et $D(\varepsilon_j) := (v_n)$. Ainsi $v_n = u_{n+1}$ pour tout $n \in \mathbb{N}$, donc $(v_0, \ldots, v_{k-1}) = (u_1, \ldots, u_k)$. Par définition de ε_j, $u_j = 1$ et $u_i = 0$ pour tout $i \in [\![0, k-1]\!]$ distinct de j. La relation (26) donne donc :

$$v_{k-1} \;=\; u_k \;=\; a_{k-1}u_{k-1} + \cdots + a_1 u_1 + a_0 u_0 \;=\; a_j.$$

Si $j := 0$, on a $v_{k-1} = u_k = a_0$ et $v_i = u_{i+1} = 0$ pour $i = 0, \ldots, k-2$, ce qui montre que $D(\varepsilon_0) = a_0 \varepsilon_{k-1}$. Si $j \geqslant 1$, on a $v_{j-1} = u_j = 1$ et $v_i = u_{i+1} = 0$ pour tout $i \in [\![0, k-2]\!]$ distinct de $j-1$, ce qui montre que $D(\varepsilon_j) = \varepsilon_{j-1} + a_j \varepsilon_{k-1}$. D'où la matrice A :

$$A \;=\; \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ a_0 & a_1 & a_2 & \ldots & a_{k-2} & a_{k-1} \end{pmatrix}.$$

La matrice A est ainsi la transposée de la matrice compagnon B de P, ce qui montre que $\chi_D = \mu_D = P$. Factorisons maintenant P :

$$P = (X - \alpha_1)^{k_1} (X - \alpha_2)^{k_2} \cdots (X - \alpha_m)^{k_m},$$

où $\alpha_1, \ldots, \alpha_m \in \mathbb{C}$ sont distincts et $k_1, \ldots, k_m \in \mathbb{N}^*$. Puisque $a_0 \neq 0$, les α_i sont non nuls. Soit $i \in [\![1, m]\!]$. La suite géométrique $(\alpha_i^n)_{n\geqslant 0}$ vérifie la relation de récurrence (26). Si $n \in \mathbb{N}$, on a en effet : $\alpha_i^{n+k} - (a_{k-1}\alpha_i^{n+k-1} + \cdots + a_1 \alpha_i^{n+1} + a_0 \alpha_i^n) = \alpha_i^n P(\alpha_i) = 0$. De plus, l'image de $(\alpha_i^n)_{n\geqslant 0}$ par D est la suite $(\alpha_i^{n+1})_{n\geqslant 0} = \alpha_i (\alpha_i^n)_{n\geqslant 0}$, *i.e.* $(\alpha_i^n)_{n\geqslant 0}$ est un vecteur propre de D associé à la valeur propre α_i. On peut généraliser :

Lemme 36. Soient $i \in [\![1, m]\!]$ et $Q \in \mathbb{C}[X]$ un polynôme de degré $d < k_i$. La suite $u := \big(Q(n)\alpha_i^n\big)_{n\geqslant 0}$ appartient à E, plus précisément $u \in \operatorname{Ker}\big((D - \alpha_i \operatorname{Id}_E)^{d+1}\big)$.

Démonstration. D'abord, si $r \in [\![0, k_i]\!]$, $(X - \alpha_i)^r$ divise P, donc $\operatorname{Ker}\big((T - \alpha_i \operatorname{Id})^r\big)$ est contenu dans $\operatorname{Ker}\big(P(T)\big) = E$, d'où $\operatorname{Ker}\big((T - \alpha_i \operatorname{Id})^r\big) = \operatorname{Ker}\big((D - \alpha_i \operatorname{Id})^r\big)$. Raisonnons alors par récurrence sur d. Si $d := 0$, u est multiple de la suite $(\alpha_i^n)_{n\geqslant 0}$, donc $u \in \operatorname{Ker}(D - \alpha_i \operatorname{Id}_E)$. Supposons que $Q \in \mathbb{C}[X]$ soit de degré $d \in [\![1, k_i - 1]\!]$ et que la conclusion soit vraie pour tout polynôme de degré strictement inférieur à d. On voit facilement que $R := Q(X+1) - Q(X)$ est de degré $d-1$. Posons

donc $u := \big(Q(n)\alpha_i^n\big)_{n\geqslant 0} := (u_n)$ et $v := (T-\alpha_i\,\mathrm{Id})(u) := (v_n)$. Pour tout n, il vient :

$$v_n = u_{n+1} - \alpha_i u_n = Q(n+1)\alpha_i^{n+1} - \alpha_i\big[Q(n)\alpha_i^n\big] = \alpha_i\big[R(n)\alpha_i^n\big].$$

D'après l'hypothèse de récurrence, $v \in \mathrm{Ker}\big((D-\alpha_i\,\mathrm{Id}_E)^d\big) = \mathrm{Ker}\big((T-\alpha_i\,\mathrm{Id})^d\big)$. Ainsi $(T-\alpha_i\,\mathrm{Id})^{d+1}(u) = (T-\alpha_i\,\mathrm{Id})^d(v) = 0$. Puisque $d+1 \leqslant k_i$, il s'ensuit que $u \in E$ et $u \in \mathrm{Ker}\big((D-\alpha_i\,\mathrm{Id}_E)^{d+1}\big)$. ∎

Nous pouvons maintenant donner une base « explicite » de E :

Proposition 37.
Gardons les hypothèses et notations ci-dessus. Les suites $\big(n^j\alpha_i^n\big)_{n\geqslant 0}$, où $i \in [\![1,m]\!]$ et $j \in [\![0,k_i-1]\!]$, forment une base de l'espace vectoriel E.

Démonstration. Fixons un indice $i \in [\![1,m]\!]$. L'espace $V_i := \mathrm{Ker}\big((D-\alpha_i\,\mathrm{Id}_E)^{k_i}\big)$ est le sous-espace caractéristique de D associé à la valeur propre α_i, donc $\dim V_i = k_i$. Pour tout $j \in [\![0,k_i-1]\!]$, posons $w_j = \big(n^j\alpha_i^n\big)_{n\geqslant 0}$. D'après le lemme précédent, w_j appartient à $\mathrm{Ker}\big((D-\alpha_i\,\mathrm{Id}_E)^{j+1}\big)$, *a fortiori* $w_j \in \mathrm{Ker}\big((D-\alpha_i\,\mathrm{Id}_E)^{k_i}\big) = V_i$. Montrons que $(w_0,\dots,w_{k_i-1})$ est une base de V_i.
Il suffit de montrer que cette famille est libre. Soient $c_0,\dots,c_{k_i-1} \in \mathbb{C}$ tels que $c_0w_0 + \cdots + c_{k_i-1}w_{k_i-1} = 0$. Pour tout $n \in \mathbb{N}$, on a donc $\sum_{j=0}^{k_i-1} c_j n^j\alpha_i^n = 0$, c'est-à-dire $\sum_{j=0}^{k_i-1} c_j n^j = 0$, puisque $\alpha_i \neq 0$. Lorsque n décrit $[\![1,k_i]\!]$, on obtient un système linéaire homogène de k_i équations à k_i inconnues $c_0,\dots,c_{k_i-1}$. La matrice de ce système est $(n^j)_{1\leqslant n\leqslant k_i,\,0\leqslant j\leqslant k_i-1}$, son déterminant est le déterminant de Vandermonde de $1,2,\dots,k_i$. Ce déterminant n'étant pas nul, les c_j sont tous nuls.
Nous savons que E est somme directe des sous-espaces caractéristiques $V_1,\dots,V_m$ de D. En juxtaposant les bases de chaque V_i obtenues ci-dessus, on obtient donc une base de E, elle est formée des suites indiquées dans l'énoncé. ∎

Considérons par exemple les suites (u_n) vérifiant la relation de récurrence linéaire d'ordre 3 : $u_{n+3} = 7u_{n+2} - 16u_{n+1} + 12u_n$.
Ici $P := X^3-7X^2+16X-12 = (X-2)^2(X-3)$. Compte tenu de la proposition, la forme générale d'une suite vérifiant la relation de récurrence ci-dessus est : $u_n = 2^n(an+b)+3^nc$, où $a,b,c \in \mathbb{C}$ sont fixés arbitrairement.

3.2 Endomorphismes nilpotents, matrices nilpotentes

Rappelons qu'un élément x d'un anneau A est dit *nilpotent* s'il existe $k \in \mathbb{N}^*$ tel que $x^k = 0$. Étant donné un tel k, on a $1 = 1-x^k = (1-x)(1+x+x^2+\cdots+x^{k-1})$, ce qui montre que $1-x$ est inversible, son inverse étant $1+x+x^2+\cdots+x^{k-1}$. Le plus petit entier $k \geqslant 1$ tel que $x^k = 0$ est appelé *indice de nilpotence* de x. Si $x,y \in A$ commutent et sont nilpotents, $x+y$ est nilpotent : cela résulte aisément de la formule du binôme (*cf.* un exercice du module II.2 de [L1]). Un endomorphisme u d'un espace vectoriel E est donc

nilpotent s'il existe $k \in \mathbb{N}^*$ tel que $u^k = 0$, de même une matrice $A \in M_n(K)$ est nilpotente s'il existe $k \in \mathbb{N}^*$ tel que $A^k = 0$. Observons que, si u est nilpotent, son polynôme minimal est X^m, où m est l'indice de nilpotence de u. En effet X^m annule u mais aucun de ses diviseurs unitaires autres que X^m, à savoir $1, X, \ldots, X^{m-1}$, n'annule u.
Ces éléments nilpotents seront importants dans les sections suivantes. En voici une caractérisation simple :

Proposition 38. Soit u un endomorphisme d'un espace vectoriel E de dimension finie $n \geqslant 1$. Pour que u soit nilpotent, il faut, et il suffit, que son polynôme caractéristique soit égal à X^n. On a la même conclusion pour une matrice carrée.

Démonstration. Supposons que u soit nilpotent, et soit $k \in \mathbb{N}^*$ tel que $u^k = 0$. D'abord u a une seule valeur propre, à savoir 0 (le vérifier). Puisque u est annulé par le polynôme scindé X^k, la remarque de la page 240 montre qu'il est trigonalisable. Il en résulte que χ_u est scindé et a pour seule racine 0, d'où $\chi_u = X^n$. Inversement, si $\chi_u = X^n$, le théorème de Cayley-Hamilton montre que $u^n = 0$, donc u est nilpotent. ∎

Il en résulte qu'une matrice $A \in M_n(K)$ est nilpotente si, et seulement si, $A^n = 0$. La suffisance est claire, mais la nécessité n'était pas évidente *a priori*.

Proposition 39. La trace d'un endomorphisme nilpotent (en dimension finie) est nulle, et l'on a la même conclusion pour une matrice carrée.

Démonstration. Supposons que u soit nilpotent, de sorte que $\chi_u = X^n$. La formule (7) de la page 223 montre alors, par identification des coefficients de X^{n-1}, que $\mathrm{Tr}(u) = 0$. ∎

Exercice 11.
Montrer qu'une matrice $A \in M_n(\mathbb{C})$ est nilpotente si, et seulement si, $\mathrm{Tr}(A^k) = 0$ pour $k = 1, \ldots, n$.
Solution. La nécessité résulte de la proposition précédente. Pour la suffisance, c'est très simple, si l'on utilise la proposition 33 de la page 241 ainsi que sa preuve. Reprenons en effet les notations de cette proposition. Les formules (20) de la page 241 montrent que $\mathrm{Tr}(M_k) = 0$ pour $k = 1, \ldots, n$. Ainsi $u_1 = u_2 = \cdots = u_n = 0$ (*cf.* la première formule (19) de la page 241). La proposition 33 de la page 241 montre alors que $a_1 = a_2 = \cdots = a_n = 0$. Autrement dit $\chi_A = X^n$, et donc A est nilpotente.

Voici une autre caractérisation des matrices nilpotentes :

Proposition 40. Une matrice $A \in M_n(K)$ est nilpotente si, et seulement si, elle est semblable à une matrice triangulaire supérieure *stricte*, c'est-à-dire à diagonale nulle.

Démonstration. Si A est nilpotente, $\chi_A = X^n$ est scindé. Soit T une matrice triangulaire supérieure semblable à A. Les valeurs propres de T sont ses termes diagonaux. Mais 0 est la seule valeur propre de A, donc de T, et ainsi la diagonale de T est nulle. La réciproque est claire, car le polynôme caractéristique d'une matrice triangulaire stricte est X^n. ∎

Remarque. Si u est à la fois diagonalisable et nilpotent, il est nul. C'est immédiat, en considérant une base de E dans laquelle la matrice de u est diagonale.

Le théorème qui va suivre montre que chaque matrice nilpotente est semblable à une matrice assez simple, dont chaque coefficient vaut 0 ou 1, résultat beaucoup plus précis que la proposition 40 de la page précédente. Ce théorème admis dans ce volume, est démontré dans [MPA1]. Introduisons d'abord une notation. Pour tout entier $r \geqslant 2$, posons :

$$J_r := \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \in M_r(K). \tag{27}$$

Les coefficients de J_r sont nuls, sauf ceux de la diagonale immédiatement en-dessous de la diagonale principale, qui valent 1. La matrice J_r est la matrice compagnon du polynôme X^r (*cf.* la section 2.1). Le polynôme minimal et le polynôme caractéristique de J_r sont donc tous deux égaux à X^r. C'est encore vrai pour $r = 1$, en posant $J_1 := 0 \in M_1(K)$.

Exercice 12.

Supposons que la matrice représentant u dans une certaine base $\mathcal{B} := (e_1, \ldots, e_n)$ de E soit égale à J_n. Pour tout $k \in \mathbb{N}$, montrer que la dimension du noyau de u^k est $\min(k, n)$.

Solution. Si $k \geqslant n$, on a $\dim(\operatorname{Ker} u^k) = \dim E = n = \min(k, n)$, car $u^k = 0$ (déjà $u^n = 0$). Supposons $k \leqslant n-1$. Par définition de J_n, $u(e_n) = 0$ et $u(e_i) = e_{i+1}$ pour $i = 1, \ldots, n-1$. Si $x := \sum_{i=1}^{n} \lambda_i e_i \in E$, on a donc $u(x) = \sum_{i=1}^{n-1} \lambda_i e_{i+1}$ puis, par récurrence sur $j \in [\![0, n-1]\!]$: $u^j(x) = \sum_{i=1}^{n-j} \lambda_i e_{i+j}$. Si $j \in [\![0, n-1]\!]$, on en déduit que $x \in \operatorname{Ker} u^j$ si, et seulement si, $\lambda_i = 0$ pour $i = 1, \ldots, n-j$.
Ainsi $\operatorname{Ker} u^j = \operatorname{Vect}(e_{n-j+1}, \ldots, e_n)$ est de dimension j.
Pour conclure, il suffit de prendre $j := k$.

Théorème 41 (Réduction des endomorphismes nilpotents).
Supposons que u soit *nilpotent*. Il existe une suite finie et décroissante d'entiers strictement positifs $(r_1, \ldots, r_m)$ et une base $\mathcal{B}$ de E telles que $\operatorname{Mat}_{\mathcal{B}}(u)$ soit égale à la matrice diagonale par blocs $\operatorname{Diag}(J_{r_1}, \ldots, J_{r_m})$. La suite $(r_1, \ldots, r_m)$ est de plus *unique*, et $\sum_{i=1}^{m} r_i = n$.

Ce théorème sera admis. Nous laissons le lecteur en déduire la version matricielle :

Théorème 42. Soit $A \in M_n(K)$ une matrice nilpotente. Il existe une *unique* suite finie et décroissante d'entiers strictement positifs $(r_1, \ldots, r_m)$ telle que A soit semblable à la matrice diagonale par blocs $\operatorname{Diag}(J_{r_1}, \ldots, J_{r_m})$.

Exemple. Prenons $n := 4$. Il y a cinq suites finies et décroissantes d'entiers strictement positifs de somme 4, à savoir $(4), (3,1), (2,2), (2,1,1), (1,1,1,1)$. Il y a donc, à similitude près, cinq matrices nilpotentes dans $M_4(K)$, à savoir $A_1 := J_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$,

$A_2 := \mathrm{Diag}(J_3, J_1) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$, $A_3 := \mathrm{Diag}(J_2, J_2) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$,

$A_4 := \mathrm{Diag}(J_2, J_1, J_1) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ et $A_5 := \mathrm{Diag}(J_1, J_1, J_1, J_1) = 0$.

Le polynôme caractéristique de chacune de ces matrices est X^4. Pour les polynômes minimaux, appliquons la proposition 21 de la page 228 et la proposition 19 de la page 227. Il vient :

$$\mu_{A_1} = X^4,\ \mu_{A_2} = X^3,\ \mu_{A_3} = \mu_{A_4} = X^2,\ \mu_{A_5} = X.$$

Deux matrices semblables ont même polynôme caractéristique et même polynôme minimal. La réciproque est fausse : les matrices A_3 et A_4 ont même polynôme caractéristique X^4 et même polynôme minimal X^2, mais elles ne sont pas semblables. C'est d'ailleurs évident, car les rangs de A_3 et A_4 sont 2 et 1 respectivement.

3.3 Décomposition de Dunford

La décomposition faisant l'objet de cette section n'est pas une décomposition de l'espace E en somme directe. Il s'agit de décomposer, de façon appropriée, l'endomorphisme u supposé scindé comme somme de deux endomorphismes, l'un étant diagonalisable et l'autre nilpotent. Une application importante de cette décomposition est le calcul de l'exponentielle d'une matrice carrée à coefficients réels ou complexes (*cf.* le module II.2).

Théorème et définition 43. On suppose que u est scindé. Il existe un unique couple $(d, \nu) \in \mathcal{L}(E)^2$ vérifiant les conditions suivantes : $u = d + \nu$, d est diagonalisable, ν est nilpotent, et $d \circ \nu = \nu \circ d$. De plus, d et ν sont des polynômes en u. La décomposition $u = d + \nu$ est appelée *décomposition de Dunford* de u.

Démonstration. Pour l'existence de (d, ν), appliquons le théorème 34 de la page 243, dont nous reprenons les notations. Puisque E est somme directe des $E^{\lambda_i}(u)$, il existe un unique $d \in \mathcal{L}(E)$ dont la restriction à chaque $E^{\lambda_i}(u)$ soit $\lambda_i \,\mathrm{Id}_{E^{\lambda_i}(u)}$. En fait, $d = \sum_{i=1}^{r} \lambda_i \pi_i$, *cf.* les restrictions à chaque $E^{\lambda_i}(u)$; cela montre que d est un polynôme en u : $d \in K[u]$. Ensuite d est diagonalisable car, si $i \in [\![1, r]\!]$, tout vecteur non

nul de $E^{\lambda_i}(u)$ est vecteur propre de d, associé à λ_i ; par juxtaposition de bases de chacun des $E^{\lambda_i}(u)$, on obtient donc une base de E formée de vecteurs propres de d.

Posons $\nu := u - d \in K[u]$. Évidemment, $u = d + \nu$ et $d \circ \nu = \nu \circ d$. Soit $i \in [\![1, r]\!]$. D'abord ν laisse stable $E^{\lambda_i}(u)$, car c'est le cas pour u et d. Notons ensuite que ν et $u - \lambda_i \operatorname{Id}_E$ ont même restriction à $E^{\lambda_i}(u)$. Par définition de $E^{\lambda_i}(u)$, la restriction de ν^{n_i} à $E^{\lambda_i}(u)$ est donc nulle. En posant $m := \max(n_1, \dots, n_r)$, il en résulte que ν^m est nul, car sa restriction à chaque $E^{\lambda_i}(u)$ l'est. Ainsi ν est nilpotent, d'où l'existence de (d, ν).

Pour l'unicité, soit $(d', \nu') \in \mathcal{L}(E)^2$ un autre couple tel que $u = d' + \nu'$ et $d' \circ \nu' = \nu' \circ d'$, on suppose que d' est diagonalisable et que ν' est nilpotent. D'abord $d' + \nu' = d + \nu$, c'est-à-dire $d - d' = \nu' - \nu$. Il s'agit de montrer que $d = d'$. L'endomorphisme ν' commute avec lui-même et avec d', il commute donc avec $u = d' + \nu'$, et aussi avec $\nu \in K[u]$. Puisque ν et ν' sont nilpotents, $\nu' - \nu$ est alors nilpotent. De même, d' commute avec d. Puisque d et d' sont diagonalisables, l'exercice 9 de la page 237 montre qu'ils sont simultanément diagonalisables, et il en résulte que $d - d'$ est diagonalisable. Ainsi $d - d' = \nu' - \nu$ est diagonalisable et nilpotent, il est donc nul (*cf.* la remarque de la page 248). D'où $d = d'$. ∎

Voici la traduction matricielle de ce théorème :

Théorème et définition 44. On suppose que A est scindée. Il existe une unique matrice *diagonalisable* $D \in M_n(K)$ et une unique matrice *nilpotente* $N \in M_n(K)$ telles que $A = D + N$ et $DN = ND$. De plus, D et N sont des polynômes en A. La décomposition $A = D + N$ est appelée *décomposition de Dunford* de A.

Démonstration. Soit $(D, N) \in M_n(K)^2$. Notons u, d, ν les endomorphismes de K^n canoniquement associés à A, D, N respectivement. Alors $(A = D+N) \iff (u = d+\nu)$, $(DN = ND) \iff (d \circ \nu = \nu \circ d)$, et $\chi_A = \chi_u$. Ensuite D est diagonalisable si, et seulement si, d est diagonalisable, de même N est nilpotente si, et seulement si, ν est nilpotent. Il suffit donc d'appliquer le théorème précédent à u. ∎

Exercice 13.

Déterminer la décomposition de Dunford $A = D + N$ de la matrice $A \in M_3(\mathbb{Q})$ de l'exemple de la page 230, *cf.* aussi l'exercice 10 de la page 244.

Solution. Nous avons vu que $\chi_A = (X + 1)(X - 1)^2$. Dans l'exercice cité, nous avons calculé les matrices spectrales Π_1, Π_2 associées aux valeurs propres $-1, 1$ respectivement. La preuve du théorème 43 de la page précédente montre que $D = -\Pi_1 + \Pi_2 = I_3 - 2\Pi_1$ (car $\Pi_1 + \Pi_2 = I_3$). D'où $D = I_3 - 2 \begin{pmatrix} 1 & -1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 2 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, et l'on en déduit :

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

La méthode illustrée dans l'exercice précédent nécessite le calcul des matrices spectrales, en particulier il faut connaître les valeurs propres λ_i de A. Si par exemple $K := \mathbb{C}$ et $n := 100$, le calcul exact des λ_i n'est pas possible. Supposons en revanche que $A \in M_{100}(\mathbb{Q})$. Comme la proposition suivante va le montrer, D et N sont à coefficients dans $\mathbb{Q}$ et peuvent être calculées « rationnellement » à partir des coefficients de A.

Proposition 45. Conservons les hypothèses et notations du théorème précédent. On suppose de plus que K est de caractéristique 0 et que les coefficients de A appartiennent à un certain sous-corps L de K. Alors les coefficients de D et N appartiennent à L.

Démonstration. Démontrons à nouveau l'existence du couple (D, N). Factorisons $P := \chi_N$:

$$P = (X-\lambda_1)^{n_1}(X-\lambda_2)^{n_2}\cdots(X-\lambda_r)^{n_r},$$

où $r \in \mathbb{N}^*$, $\lambda_1, \ldots, \lambda_r \in K$ sont distincts, et $n_1, \ldots, n_r \in \mathbb{N}^*$. Considérons le polynôme $T := (X-\lambda_1)(X-\lambda_2)\cdots(X-\lambda_r)$. Il a r racines simples, qui ne sont pas racines de T', de sorte que $T \wedge T' = 1$. Il existe donc $U, V \in K[X]$ tels que $UT + VT' = 1$. Par ailleurs T divise P et, si l'on pose $m := \max(n_1, \ldots, n_r)$, P divise T^m. Comme $P(A) = 0$ (théorème de Cayley-Hamilton), il en résulte que $T(A)^m = 0$, et donc $T(A)$ est nilpotente.

Nous allons travailler dans la sous-algèbre $R := K[A]$ de $M_n(K)$, qui est commutative, comme image du morphisme $Q \mapsto Q(A)$ de $K[X]$ dans $M_n(K)$. L'idée est de chercher une matrice $D \in R$ vérifiant $T(D) = 0$ et telle que $N := A - D$ soit nilpotente. Supposons que D soit une telle matrice. D'abord D est diagonalisable, puisque T est scindé à racines simples et annule D. Ensuite N est nilpotente par hypothèse et $A = D + N$ par construction. Enfin $D, N \in R$ commutent car R est commutative.

Le problème est donc de construire dans R une racine D de l'équation $T(D) = 0$. Inspirons-nous de la méthode de Newton pour résoudre dans $\mathbb{R}$ une équation de la forme $f(x) = 0$ (*cf.* le module II.8, section 2.2.3 de la page 903). Définissons une suite $(M_k)_{k\geqslant 0}$ d'éléments de R à l'aide de son premier terme $M_0 := A$ et de la relation de récurrence suivante :

$$M_{k+1} := M_k - T(M_k)T'(M_k)^{-1}. \tag{28}$$

Il nous faut justifier cette définition : pourquoi les $T'(M_k)$ sont-elles inversibles ? Supposons que, pour un certain $k \in \mathbb{N}$, nous ayons construit un élément M_k de R de sorte que :

a) d'une part $A - M_k$ soit nilpotente ;

b) d'autre part $T(M_k)$ soit multiple (dans R) de $T(A)^{2^k}$.

Tel est le cas si $k = 0$, puisque $M_0 = A$. L'égalité $UT + VT' = 1$ implique la suivante :

$$V(M_k)T'(M_k) = I_n - U(M_k)T(M_k).$$

Comme $T(M_k)$ est nilpotente (parce que $T(A)$ l'est) et R est commutative, on en déduit que $T'(M_k)$ est inversible. Nous pouvons donc définir M_{k+1} par la formule (28). Comme $A - M_k$ et $T(M_k)T'(M_k)^{-1} = M_k - M_{k+1}$ sont nilpotentes et commutent, leur somme $A - M_{k+1}$ est nilpotente. Il reste à évaluer $T(M_{k+1})$. Puisque R est une algèbre commutative sur le corps K de caractéristique 0, on peut appliquer à T la formule de Taylor à l'ordre 1 au point M_k. Il existe donc $Z \in R$ tel que :

$$\begin{aligned} T(M_{k+1}) &= T(M_k) + (M_{k+1} - M_k)T'(M_k) + Z(M_{k+1} - M_k)^2 \\ &= T(M_k) - \big[T(M_k)T'(M_k)^{-1}\big]T'(M_k) + Z\big[T(M_k)T'(M_k)^{-1}\big]^2 \\ &= Z\big[T(M_k)T'(M_k)^{-1}\big]^2. \end{aligned}$$

Ainsi $T(M_{k+1})$ est multiple de $T(M_k)^2$, donc de $T(A)^{2^{k+1}}$, vu la condition b). Cela achève la construction de la suite $(M_k)_{k\geqslant 0}$ et montre que les propriétés a) et b) sont vérifiées pour tout k.

Soit maintenant k un entier naturel (par exemple le plus petit) tel que $2^k \geqslant m$. Posons $D := M_k$. D'une part $T(D)$ est multiple de $T(A)^{2^k}$, donc de $T(A)^m = 0$, *i.e.* $T(D) = 0$. D'autre part $N := A - D$ est nilpotente, vu la condition a). Le couple (D, N) répond ainsi à la question.

Venons-en à la dernière assertion de l'énoncé. On suppose que $A \in M_n(L)$, L étant un sous-corps de K, et l'on veut montrer que $D, N \in L[A]$. Tout d'abord $P = \chi_A$ appartient à $L[X]$. Par contre il n'y a aucune raison pour que les λ_i appartiennent à L (penser au cas où $K := \mathbb{C}$ et $L := \mathbb{Q}$). Montrons cependant que $T \in L[X]$. Partons de la formule de dérivation évidente :

$$P' = \sum_{i=1}^{r} n_i(X-\lambda_1)^{n_1}(X-\lambda_2)^{n_2}\cdots(X-\lambda_i)^{n_i-1}\cdots(X-\lambda_r)^{n_r}.$$

Pour tout $i \in [\![1, r]\!]$, seul le $i^{ème}$ terme du second membre de cette formule n'est pas multiple de $(X-\lambda_i)^{n_i}$ (en effet $n_i 1_K \neq 0$, car K est de caractéristique 0). Le pgcd de P et P' est alors $H := \prod_{i=1}^{r}(X-\lambda_i)^{n_i-1}$. De plus $H \in L[X]$ car $P, P' \in L[X]$. Comme $P = TH$, T est le quotient dans la division euclidienne de P par H, donc $T \in L[X]$. Considérons la L-algèbre $S := L[A]$. Les formules (28) montrent maintenant que $M_h \in S$ pour tout $h \in \mathbb{N}$, en particulier $D = M_k \in S$, puis $N = A - D \in S$. ∎

Remarque. La preuve ci-dessus suggère un algorithme donnant D et N (*cf.* l'exercice I.5.26 de la page 259). Cet algorithme est « rationnel », *i.e.* n'utilise que les quatre opérations dans L. On calcule d'abord T par la formule $T = P/(P \wedge P')$, utilisant l'algorithme d'Euclide, puis on construit les M_h à l'aide des formules (28). Cet algorithme est aussi assez rapide. Supposons ainsi que $K := \mathbb{C}$, $L := \mathbb{Q}$ et $n := 100$. Nous n'avons heureusement pas besoin de calculer les λ_i, qui sont les racines d'un polynôme de degré 100 ! Avec les notations de la preuve ci-dessus, on a $m \leqslant 100$, ce qui permet de prendre $k := 7$. Il y a donc au plus sept itérations à effectuer.

Exercice 14.

À l'aide de la méthode précédente, retrouver les résultats de l'exercice 13 de la page 250.

Solution. Reprenons les notations de la preuve de la proposition 45 de la page précédente. Ici $T := X^2 - 1$ et $T(A)^2 = 0$, car $m := 2$.

Ainsi $k = 1$ convient, d'où $D = A - T(A)T'(A)^{-1}$, soit :

$$D = A - (A^2 - I_2)(2A)^{-1} = (A + A^{-1})/2 \qquad \text{et} \qquad N = (A - A^{-1})/2.$$

On vérifie que $A^{-1} = \begin{pmatrix} -1 & 2 & -5 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix}$, et l'on retrouve les résultats de l'exercice cité :

$$D = \begin{pmatrix} -1 & 2 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{et} \qquad N = \begin{pmatrix} 0 & 0 & 3 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}.$$

Une première application de la décomposition de Dunford est de permettre le calcul des puissances de A. D'abord, nous savons calculer les puissances de D, *cf.* la section 2.3. Ensuite, soit $p \in \mathbb{N}^*$ le plus petit entier tel que $N^p = 0$ (en fait, $p \leqslant n$, nous l'avons vu). Le calcul des puissances de N se limite au calcul de N^k pour $k = 2, \dots, p-1$. Le point crucial est que l'égalité $DN = ND$ permet d'appliquer alors la formule du binôme : pour tout $k \in \mathbb{N}$, on obtient :

$$A^k = \sum_{i=0}^{\min(k,p-1)} \binom{k}{i} N^i D^{k-i}.$$

Soient $D' = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ et $N' := \begin{pmatrix} 0 & 2 & 1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}$.

On a encore $A = D' + N'$, D' est diagonalisable (!) et N' est nilpotente (proposition 40 de la page 247). Cependant $A = D' + N'$ *n'est pas* la décomposition de Dunford de A, car $N'D' \neq D'N'$. Cela vient de l'unicité de la décomposition de Dunford ou d'un calcul direct. On ne peut donc pas appliquer la formule du binôme pour calculer les $(D' + N')^k$.

3.4 Réduction de Jordan

Pour un endomorphisme *scindé*, le théorème de réduction « le plus précis » est le théorème de Jordan ci-après. Ce théorème est démontré dans [MPA1] (comme cas particulier de résultats plus généraux). Nous l'admettons dans ce volume, en nous contentant de l'illustrer.
Introduisons une notation. Si $\lambda \in K$ et $r \in \mathbb{N}^*$, nous poserons :

$$J_r(\lambda) := \begin{pmatrix} \lambda & 0 & \cdots & 0 & 0 \\ 1 & \lambda & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \lambda \end{pmatrix} = \lambda I_r + J_r \in M_r(K). \tag{29}$$

Nous dirons que $J_r(\lambda)$ est le *bloc de Jordan de taille r associé à λ*. Calculons le polynôme caractéristique et le polynôme minimal d'un tel bloc A. La matrice $XI_r - JA$ est triangulaire inférieure et ses termes diagonaux valent $X - \lambda$, donc son déterminant χ_A vaut $(X - \lambda)^r$. Soient ensuite $Q \in K[X]$ et $R := Q(X - \lambda)$. En substituant A à X, il vient : $R(A) = Q(A - \lambda I_r) = Q(J_r)$. Ainsi $R(A) = 0$ si, et seulement si, $Q(J_r) = 0$, *i.e.* si, et seulement si, Q est multiple du polynôme minimal de J_r, à savoir X^r, ce qui revient à dire que $R = Q(X - \lambda)$ est multiple de $(X - \lambda)^r$. Ainsi $\mu_A = (X - \lambda)^r$. Le polynôme caractéristique et le polynôme minimal de $J_r(\lambda)$ sont donc tous deux égaux à $(X - \lambda)^r$.
Voici le théorème de réduction visé, sous forme « géométrique », puis matricielle :

Théorème 46 (Jordan).
Supposons que u soit scindé. Il existe une base $\mathcal{B}$ de E, un entier $m \geqslant 1$ et deux suites $(r_1, \dots, r_m) \in \mathbb{N}^{*m}$, $(\lambda_1, \dots, \lambda_m) \in K^m$ tels que la matrice de u dans la base $\mathcal{B}$ soit égale à la matrice diagonale par blocs $\operatorname{Diag}\big(J_{r_1}(\lambda_1), \dots, J_{r_m}(\lambda_m)\big)$.
De plus, la suite $\big((r_1, \lambda_1), \dots, (r_m, \lambda_m)\big)$ est unique, *à l'ordre près*.

Théorème 47 (Réduction de Jordan). Soit $A \in M_n(K)$ une matrice scindée.
Il existe un entier $m \geqslant 1$ et deux suites $(r_1, \dots, r_m) \in \mathbb{N}^{*m}$, $(\lambda_1, \dots, \lambda_m) \in K^m$ tels que A soit semblable à la matrice $\mathrm{Diag}\big(J_{r_1}(\lambda_1), \dots, J_{r_m}(\lambda_m)\big)$. Une telle matrice diagonale par blocs est appelée *réduite de Jordan* de A. De plus, la suite $\big((r_1, \lambda_1), \dots, (r_m, \lambda_m)\big)$ est unique, *à l'ordre près*.

Remarque. Soit $A' := \mathrm{Diag}\big(J_{r_1}(\lambda_1), \dots, J_{r_m}(\lambda_m)\big)$ une réduite de Jordan de A, comme dans le théorème. Il est important de noter que $\lambda_1, \dots, \lambda_m$ ne sont pas nécessairement distincts : il peut y avoir plusieurs blocs associés au même scalaire. Puisque A et A' sont semblables, elles ont même polynôme caractéristique et même polynôme minimal. Les calculs ci-dessus et la proposition 19 de la page 227 montrent que $\mu_A = \mu_{A'}$ est le ppcm des $(X - \lambda_i)^{r_i}$ et que $\chi_A = \chi_{A'}$ est le produit des $(X - \lambda_i)^{r_i}$. En particulier un scalaire α est valeur propre de A si, et seulement s'il est égal à l'un des λ_i, la multiplicité de α comme valeur propre de A étant alors la somme des r_i pour tous les $i \in [\![1, m]\!]$ tels que $\lambda_i = \alpha$.
La matrice A' est triangulaire inférieure, car c'est le cas pour chaque bloc de Jordan.
Posons ensuite $D' := \mathrm{Diag}(\lambda_1 I_{r_1}, \dots, \lambda_m I_{r_m})$ et $N' := \mathrm{Diag}(J_{r_1}, \dots, J_{r_m})$. Évidemment $A' = D' + N'$, D' est diagonale, N' est nilpotente, $D'N' = N'D'$. Ainsi $A' = D' + N'$ est la décomposition de Dunford de A'. La décomposition de Dunford de A s'en déduit : si $P \in GL_n(K)$ est telle que $A = PA'P^{-1}$, cette décomposition est $A = D + N$, où $D := PD'P^{-1}$ et $N := PN'P^{-1}$.

Corollaire 48. Supposons que le corps K soit algébriquement clos. Deux matrices $A, B \in M_n(K)$ sont semblables si, et seulement si, elles ont même réduite de Jordan, à l'ordre des blocs près.

Démonstration. C'est une conséquence immédiate du théorème précédent, puisque tout polynôme non constant à coefficients dans K est scindé. ∎

Exemple. Soit $A \in M_2(\mathbb{C})$. D'après le théorème précédent, ou bien A est semblable à une matrice diagonale $\mathrm{Diag}(\lambda, \mu)$ (où $\lambda, \mu \in K$ sont distincts ou non), auquel cas A est diagonalisable, ou bien A est semblable à un bloc de Jordan $J_2(\lambda) = \begin{pmatrix} \lambda & 0 \\ 1 & \lambda \end{pmatrix}$, $\lambda \in K$. Dans ce dernier cas, A n'est pas diagonalisable : cela se déduit de la remarque ci-dessus, ou, plus directement, du fait que $\mu_A = (X - \lambda)^2$ ait une racine double.

Exercice 15.

Déterminer une réduite de Jordan A' de la matrice $A := \begin{pmatrix} -1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix} \in M_3(\mathbb{Q})$ de l'exemple de la page 230, ainsi qu'une matrice $P \in GL_3(\mathbb{Q})$ telle que $A' = P^{-1}AP$.

Solution. Dans l'exemple cité, nous avons montré que $\chi_A = (X+1)(X-1)^2$ (donc χ_A est scindé sur $\mathbb{Q}$), et que A n'est pas diagonalisable. Il y a donc pour A' une seule possibilité, à l'ordre des blocs près, à savoir $A' := \mathrm{Diag}\big(J_1(-1), J_2(1)\big)$, *i.e.* $A' = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$.

Pour déterminer P, il faut travailler un peu plus.
Soient $\mathcal{B} := (e_1, e_2, e_3)$ la base canonique de $\mathbb{Q}^3$ et $u \in \mathcal{L}(\mathbb{Q}^3)$ l'endomorphisme canoniquement associé à A. L'espace propre de u associé à -1 est évidemment la droite $\mathbb{Q}\varepsilon_1$, où $\varepsilon_1 := e_1$. On vérifie que l'espace propre de u associé à 1 est la droite $\mathbb{Q}\varepsilon_3$, où $\varepsilon_3 := 3(e_1 + e_2)$. Le calcul donne ensuite : $(A - I_3)^2 = \begin{pmatrix} 4 & -4 & 4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, donc le sous-espace caractéristique de u associé à la valeur propre 1 est le plan V d'équation $x_1 - x_2 + x_3 = 0$. L'endomorphisme v de V induit par $u - \mathrm{Id}$ vérifiant $v^2 = 0$ et $v \neq 0$, le noyau et l'image de v sont tous deux égaux à $\mathbb{Q}\varepsilon_3$. Le plan V possède donc une base de la forme $(\varepsilon_2, \varepsilon_3)$, où $\varepsilon_2 \in V$ est tel que $v(\varepsilon_2) = \varepsilon_3$; on peut par exemple prendre $\varepsilon_2 := e_2 + e_3$. Ainsi $\mathcal{C} := (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ est une base de $\mathbb{Q}^3$, et l'on a $u(\varepsilon_1) = -\varepsilon_1$, $u(\varepsilon_2) = \varepsilon_2 + \varepsilon_3$, $u(\varepsilon_3) = \varepsilon_3$, donc $\mathrm{Mat}_{\mathcal{C}}(u) = A'$. En conclusion, $A' := P^{-1}AP$, où P est la matrice de passage de $\mathcal{B}$ à $\mathcal{C}$, *i.e.* $P := \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 3 \\ 0 & 1 & 0 \end{pmatrix}$.

Exercice 16.

Soit $P \in \mathbb{C}[X]$ un polynôme unitaire de degré $k \geqslant 1$. Déterminer une réduite de Jordan de la matrice compagnon de P, en fonction des racines de P et de leurs multiplicités.

Solution. Soit $A \in M_k(\mathbb{C})$ la matrice compagnon de P. Nous savons que $\chi_A = \mu_A = P$. Écrivons maintenant la factorisation de P : $P = (X - \alpha_1)^{k_1}(X - \alpha_2)^{k_2} \cdots (X - \alpha_m)^{k_m}$, où $\alpha_1, \ldots, \alpha_m \in \mathbb{C}$ sont distincts et $k_1, \ldots, k_m \in \mathbb{N}^*$. Ainsi $\sum_{i=1}^{m} k_i = k$. Considérons *a priori* une réduite de Jordan $A' := \mathrm{Diag}\big(J_{r_1}(\lambda_1), \ldots, J_{r_{m'}}(\lambda_{m'})\big)$ de A. D'après la remarque de la page ci-contre, P est à la fois le produit et le ppcm de $(X - \lambda_1)^{r_1}, \ldots, (X - \lambda_{m'})^{r_{m'}}$, ce qui n'est possible que si $\lambda_1, \ldots, \lambda_{m'}$ sont deux à deux distincts. Vu la factorisation de P, et quitte à renuméroter les α_i, on a donc $m' = m$ et, pour tout $i \in [\![1, m]\!]$, $\lambda_i = \alpha_i$ et $r_i = k_i$. Voici la conclusion : $A' := \mathrm{Diag}\big(J_{k_1}(\alpha_1), \ldots, J_{k_m}(\alpha_m)\big)$ est une réduite de Jordan de A. Comme $\mu_A = P$, A est diagonalisable si, et seulement si, tous les k_i valent 1.

Exercice type corrigé

I.5.0 Soit E un espace vectoriel de dimension finie $n \geqslant 1$ sur $\mathbb{C}$ et u, v deux endomorphismes de E. On suppose que $u^2 = v^2 = -\mathrm{Id}_E$ et $u \circ v + v \circ u = 0$. Montrer que n est pair, disons $n := 2m$, et qu'il existe une base $\mathcal{B}$ de E telle que :

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} iI_m & 0 \\ 0 & -iI_m \end{pmatrix} \quad \text{et} \quad \mathrm{Mat}_{\mathcal{B}}(v) = \begin{pmatrix} 0 & -I_m \\ I_m & 0 \end{pmatrix}.$$

Pour cela, montrer que u et v sont diagonalisables, puis étudier leurs sous-espaces propres.

Solution. L'endomorphisme u est diagonalisable, car annulé par le polynôme $X^2 + 1$, scindé à racines simples. Par ailleurs, puisque $u^2 = -\mathrm{Id}_E$, le spectre de u est inclus dans $\{-i, i\}$. Il en résulte que $E = E_i(u) \oplus E_{-i}(u)$. La situation pour v est analogue.
Soient $\varepsilon \in \{-1, 1\}$ et $x \in E_{\varepsilon i}(u)$, de sorte que $u(x) = \varepsilon i x$. L'égalité $u \circ v + v \circ u = 0$ implique : $u\big(v(x)\big) = -v\big(u(x)\big) = -v(\varepsilon i x) = -\varepsilon i v(x)$, d'où $v(x) \in E_{-\varepsilon i}(u)$.

Ainsi $v\big(E_i(u)\big) \subset E_{-i}(u)$ et $v\big(E_{-i}(u)\big) \subset E_i(u)$. Mais v est inversible, puisque $v^2 = -\mathrm{Id}_E$. La restriction w de v à $E_i(u)$ est donc une application linéaire *injective* de $E_i(u)$ dans $E_{-i}(u)$, d'où $\dim\big(E_i(u)\big) \leqslant \dim\big(E_{-i}(u)\big)$. En échangeant les rôles de i et $-i$, on voit donc que $E_i(u)$ et $E_{-i}(u)$ ont une même dimension m, donc w est un isomorphisme de $E_i(u)$ sur $E_{-i}(u)$. De plus, $n = 2m$, car $E = E_i(u) \oplus E_{-i}(u)$, donc n est pair.
Soit $(e_1, \dots, e_m)$ une base de $E_i(u)$. Définissons des vecteurs $e_{m+1}, \dots, e_n$ en posant pour tout $k \in [\![1, m]\!]$: $e_{m+k} := v(e_k)$. Puisque w est un isomorphisme, $(e_{m+1}, \dots, e_n)$ est une base de $E_{-i}(u)$. Par juxtaposition, $\mathcal{B} := (e_1, \dots, e_n)$ est une base de E, adaptée à la décomposition $E = E_i(u) \oplus E_{-i}(u)$. Montrons que $\mathcal{B}$ répond à la question.
Soit $k \in [\![1, m]\!]$. Puisque $e_k \in E_i(u)$ et $e_{m+k} \in E_{-i}(u)$, on a $u(e_k) = ie_k$ et $u(e_{m+k}) = -ie_{m+k}$. On en déduit l'égalité $\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} iI_m & 0 \\ 0 & -iI_m \end{pmatrix}$.
Passons à v. Pour tout $k \in [\![1, m]\!]$, $v(e_k) = e_{m+k}$ par définition de e_{m+k}. Comme $v^2 = -\mathrm{Id}_E$, on a aussi $v(e_{m+k}) = v^2(e_k) = -e_k$. Cela prouve l'égalité $\mathrm{Mat}_{\mathcal{B}}(v) = \begin{pmatrix} 0 & -I_m \\ I_m & 0 \end{pmatrix}$.
Le lecteur constatera que la réciproque de ce que nous venons de prouver est immédiate.

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.5.1 Si $p, q \in \mathbb{N}^*$ et $A \in M_{p,q}(K), B \in M_{q,p}(K)$, montrer que $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

I.5.2 $*$ Soient K un corps de caractéristique nulle et $M \in M_n(K)$.

1. Montrer que, si la trace de M est nulle, M est semblable à une matrice dont tous les termes diagonaux sont nuls.
2. Soit $D = \mathrm{Diag}(1, 2, \dots, n)$. Déterminer le noyau et l'image de l'application linéaire $A \mapsto DA - AD$ de $M_n(K)$ dans $M_n(K)$.
3. En déduire que, si la trace de M est nulle, il existe deux matrices $A, B \in M_n(K)$ telles que $M = AB - BA$ (la réciproque est évidente).

I.5.3 En utilisant la méthode de Faddeev-Le Verrier, écrire un programme donnant le polynôme caractéristique de toute matrice $A \in M_n(\mathbb{C})$.

I.5.4 Soient $p, q \in \mathbb{N}^*$.
Considérons des matrices $A \in M_p(K)$, $B \in M_{p,q}(K)$, $C \in M_{q,p}(K)$, $D \in M_q(K)$, et posons $M := \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in M_{p+q}(K)$. On suppose que A est inversible. Montrer que M est de rang p si, et seulement si, $D = CA^{-1}B$.

I.5.5 On suppose ici que K est un sous-corps d'un corps L. Soit $A \in M_n(K)$. Montrer que le polynôme minimal de A est le même, que l'on considère A comme élément de $M_n(K)$ ou comme élément de $M_n(L)$. On pourra utiliser l'exercice II.7.2 de [L1] portant sur les systèmes linéaires.

Exercices sur la section 2

I.5.6 Résoudre dans $M_3(\mathbb{C})$ l'équation $M^2 = A$, où $A := \begin{pmatrix} 11 & -5 & 5 \\ -5 & 3 & -3 \\ 5 & -3 & 3 \end{pmatrix}$.

I.5.7 Soit u un endomorphisme diagonalisable d'un espace vectoriel E de dimension finie : E est somme directe des espaces propres $E_\lambda(u)$, λ décrivant $\mathrm{sp}(u)$. Montrer que les sous-espaces de E stables par u sont exactement les sommes $\sum_{\lambda\in\mathrm{sp}(u)} V_\lambda$ où, pour tout $\lambda \in \mathrm{sp}(u)$, V_λ est un sous-espace vectoriel arbitraire de $E_\lambda(u)$.

I.5.8 Soit $A := \mathrm{Diag}(A_1, \ldots, A_m)$ une matrice diagonale par blocs.
Montrer que A est diagonalisable si, et seulement si, chaque A_i est diagonalisable.

I.5.9 Soit $A \in E := M_n(K)$ une matrice fixée. Montrer que $u : M \mapsto AM$ est un endomorphisme de E. Montrer que u est nilpotent (resp. diagonalisable) si, et seulement si, la matrice A possède la même propriété.

I.5.10 $*$ Soit E un $\mathbb{C}$-espace vectoriel de dimension finie $n \geqslant 1$. Montrer qu'un endomorphisme u de E est diagonalisable si, et seulement si, tout sous-espace de E stable par u possède un supplémentaire lui aussi stable par u.

I.5.11 Soient E un $\mathbb{C}$-espace vectoriel de dimension finie $n \geqslant 1$ et $u \in \mathcal{L}(E)$. Soit p un entier strictement positif tel que u^p soit diagonalisable. Montrer que u est diagonalisable si, et seulement si, u et u^p ont même noyau.

I.5.12 Soient $A \in M_n(\mathbb{C})$ et $M := \begin{pmatrix} 0 & A \\ I_n & 0 \end{pmatrix} \in M_{2n}(\mathbb{C})$. Montrer que M est diagonalisable si, et seulement si, A est inversible et diagonalisable (étudier les éléments propres de M).

I.5.13 Soient $A \in M_n(\mathbb{C})$ et $M := \begin{pmatrix} A & A \\ 0 & A \end{pmatrix} \in M_{2n}(\mathbb{C})$.
Montrer que M est diagonalisable si, et seulement si, A est nulle.

I.5.14 Soient $A = (a_{i,j}) \in M_p(\mathbb{C})$ et $B \in M_n(\mathbb{C})$ deux matrices diagonalisables (avec $p \geqslant 1$ et $n \geqslant 1$). On considère la matrice $M \in M_{pn}(\mathbb{C})$ ci-dessous, définie par blocs :

$$M := \begin{pmatrix} a_{1,1}B & a_{1,2}B & \ldots & a_{1,p}B \\ a_{2,1}B & a_{2,2}B & \ldots & a_{2,p}B \\ \vdots & \vdots & & \vdots \\ a_{p,1}B & a_{p,2}B & \ldots & a_{p,p}B \end{pmatrix}.$$

Montrer que M (appelée *produit de Kronecker* de A et B) est diagonalisable.

I.5.15 Soient n, p, q trois entiers tels que $1 \leqslant p, q \leqslant n$ et $t \in \mathbb{C}$. Notons $A \in M_n(\mathbb{C})$ la matrice dont tous les coefficients valent 1, sauf peut-être le coefficient d'indices (p, q), qui vaut t.
À quelles conditions A est-elle diagonalisable ?

I.5.16 Trigonaliser la matrice $A := \begin{pmatrix} -3 & -3 & 2 \\ 1 & 1 & -2 \\ 2 & 4 & -4 \end{pmatrix} \in M_3(\mathbb{C})$, *i.e.* expliciter une matrice $P \in GL_3(\mathbb{C})$ telle que $P^{-1}AP$ soit triangulaire supérieure.
Calculer alors les puissances A^k de A, $k \in \mathbb{N}$.

I.5.17 Soit $A := \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \in M_3(\mathbb{Q})$. Calculer A^{1789} de plusieurs manières.

I.5.18 Soit $A \in M_n(\mathbb{C})$, écrivons $\chi_A := \prod_{j=1}^{n} (X - \lambda_j)$. Montrer que, pour tout $P \in \mathbb{C}[X]$, le polynôme caractéristique de la matrice $P(A)$ est $\prod_{j=1}^{n} \big(X - P(\lambda_j)\big)$.

I.5.19 Soient $P \in K[X]$ un polynôme unitaire de degré $n \geqslant 1$ et A l'algèbre quotient $K[X]/(P)$. Notons $x \in A$ la classe de X modulo P et u l'endomorphisme $y \mapsto xy$ du K-espace vectoriel A. Montrer que la matrice de u dans une base convenable de A est la matrice compagnon C_P de P. En déduire le polynôme minimal de C_P.

I.5.20 $*$ Soient E un espace vectoriel de dimension finie $n \geqslant 1$ et $(u_i)_{i\in I}$ une famille d'endomorphismes trigonalisables de E commutant deux à deux. Montrer que les u_i sont *simultanément trigonalisables*, *i.e.* qu'il existe une base $\mathcal{B}$ de E telle que, pour tout $i \in I$, $\mathrm{Mat}_{\mathcal{B}}(u_i)$ soit triangulaire supérieure.

I.5.21 Soient E un espace vectoriel de dimension finie $n \geqslant 1$ sur $\mathbb{C}$ et $u, v \in \mathcal{L}(E)$ tels que $u \circ v = v \circ u$. On suppose que v est nilpotent. Montrer qu'alors $\det(u + v) = \det(u)$.

I.5.22 Soient $A, B \in M_n(\mathbb{C})$. Montrer que l'égalité $AB - BA = I_n$ est impossible. Si $AB - BA = B$, montrer que B est nilpotente.

I.5.23 $*$ Soit $n \in \mathbb{N}^*$. Pour toute permutation $s \in \mathfrak{S}_n$ de $[\![1, n]\!]$, notons $P_s \in GL_n(\mathbb{C})$ la matrice de permutation correspondante : notant $\mathcal{B} := (e_1, \ldots, e_n)$ la base canonique de $\mathbb{C}^n$, $P_s := \mathrm{Mat}_{\mathcal{B}}(u_s)$, où $u_s \in \mathcal{L}(\mathbb{C}^n)$ est défini par $u_s(e_j) := e_{s(j)}$ pour $j = 1, \ldots, n$.

1. Soit $s \in \mathfrak{S}_n$. Montrer sans calculs que P_s est diagonalisable.
2. Diagonaliser P_s lorsque s est la permutation circulaire $(1\ 2\ \cdots\ n)$.
3. Soit $s \in \mathfrak{S}_n$ une permutation quelconque. Expliciter une base de $\mathbb{C}^n$ formée de vecteurs propres de u_s, à l'aide de la décomposition de s en produit de cycles à supports disjoints.

I.5.24 $*$ Soit K un corps fini à q éléments. On sait que q est de la forme $q := p^r$, où p est la caractéristique de K et $r \in \mathbb{N}^*$. Soient E un K-espace vectoriel de dimension finie $n \geqslant 1$ et $u \in \mathcal{L}(E)$. Montrer que u est diagonalisable si, et seulement si, $u^q = u$.

I.5.25 $**$ Soit n un entier naturel *impair*.

On se propose de calculer la *somme de Gauß* $G(n) := \sum_{k=0}^{n-1} \exp(2\pi i k^2/n)$. Soient $z := \exp(2\pi i/n)$ et A la matrice de Vandermonde de $1, z, z^2, \dots, z^{n-1}$, *i.e.* $A := \left(z^{(j-1)(k-1)}\right)_{1\leqslant j,k\leqslant n} \in M_n(\mathbb{C})$. Noter que $G(n) = \mathrm{Tr}(A)$.

1. Pour tout $r \in \mathbb{Z}$, calculer $\sum_{k=1}^{n} z^{kr}$.

2. Calculer A^2 et $\mathrm{Tr}(A^2)$. Montrer que $A^4 = n^2 I_n$, et en déduire que A est diagonalisable.

3. Montrer que $\chi_A := (X-\sqrt{n})^a (X+\sqrt{n})^b (X-i\sqrt{n})^c (X+i\sqrt{n})^d$, où $(a,b,c,d) \in \mathbb{N}^4$.

4. Montrer que $\det(A) := i^{n(n-1)/2} t$, où t est un nombre réel strictement positif.

5. Montrer que $G(n)\overline{G(n)} = \sum_{j=1}^{n} \left(\sum_{h=1}^{n} z^{(j+h)^2-h^2} \right) = n$. En déduire que $G(n)$ vaut $\sqrt{n}$ ou $i\sqrt{n}$, suivant que n soit congru à 1 ou à 3 modulo 4 $\big($déterminer d'abord $(a,b,c,d)\big)$.

Exercices sur la section 3

I.5.26 Soit $n \in \mathbb{N}^*$.
Écrire un programme donnant la décomposition de Dunford d'une matrice $A \in M_n(\mathbb{C})$.

I.5.27 Soient $n \in \mathbb{N}^*$ et $A \in GL_n(K)$ une matrice scindée. Montrer que A s'écrit de manière unique comme produit de deux matrices qui commutent, dont l'une est diagonalisable et l'autre est la somme de I_n et d'une matrice nilpotente.

I.5.28 Soit $n \in \{1,2,3\}$. Montrer que deux matrices *scindées* $A, B \in M_n(K)$ sont semblables si, et seulement si, elles ont même polynôme caractéristique et même polynôme minimal.

I.5.29 Soit $A := \begin{pmatrix} 1 & -5 & 4 & -3 \\ 4 & 11 & -8 & 4 \\ 6 & 7 & -5 & 1 \\ 4 & -10 & 8 & -7 \end{pmatrix} \in M_4(\mathbb{C})$. Déterminer les sous-espaces caractéristiques et les sous-espaces propres de A, *i.e.* de l'endomorphisme de $\mathbb{C}^4$ canoniquement associé à A. Déterminer la décomposition de Dunford de A. Expliciter enfin une réduite de Jordan de A.

I.5.30 Soit $A \in M_n(\mathbb{C})$, $n \geqslant 1$. À partir d'une réduite de Jordan de A, expliciter le polynôme minimal de A et calculer les dimensions des espaces propres de A.

I.5.31 Soit $A \in M_n(\mathbb{C})$. Montrer que $\chi_A = \mu_A$ si, et seulement si, A est semblable à la matrice compagnon d'un polynôme unitaire non constant.

I.5.32 Soient $k \in \mathbb{N}^*$ et $a_0, a_1, \dots, a_{k-1} \in \mathbb{C}$, $a_0 \neq 0$.
Trouver des conditions nécessaires et suffisantes pour que toute suite complexe (u_n) vérifiant la relation de récurrence linéaire $u_{n+k} = a_{k-1}u_{n+k-1} + \cdots + a_1 u_{n+1} + a_0 u_n$ soit bornée.

I.5.33 $*$ Soient E un $\mathbb{C}$-espace vectoriel de dimension finie $n \geqslant 1$ et $u \in \mathcal{L}(E)$. Trouver une condition nécessaire et suffisante pour que E n'ait qu'un nombre fini de sous-espaces stables par u.

I.5.34 $*$ Étant donnés un entier $n \geqslant 2$ et deux nombres complexes non nuls a, b, on considère la matrice :

$$A = \begin{pmatrix} 0 & b & 0 & \dots & 0 & 0 \\ a & 0 & b & \dots & 0 & 0 \\ 0 & a & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & b \\ 0 & 0 & 0 & \dots & a & 0 \end{pmatrix} \in M_n(\mathbb{C}).$$

Montrer que A est diagonalisable, et la diagonaliser. On pourra utiliser certaines suites définies par des relations de récurrence linéaires d'ordre 2.

I.5.35 $*$ Montrer que deux matrices $A, B \in M_n(\mathbb{C})$ sont semblables si, et seulement si, pour tout couple $(\lambda, k) \in \mathbb{C} \times \mathbb{N}$, $(A - \lambda I_n)^k$ et $(B - \lambda I_n)^k$ ont même rang.

I.5.36 $*$ Soit $A \in M_n(\mathbb{C})$. Posons $C(A) := \big\{M \in M_n(\mathbb{C}) \mid AM = MA\big\}$.
Montrer que $C(A)$ est une sous-algèbre de $M_n(\mathbb{C})$ contenant $\mathbb{C}[A]$.
Prouver l'inégalité $\dim\big(C(A)\big) \geqslant n$. L'algèbre $C(A)$ est appelée *commutant* de A.

I.5.37 Soit $A \in E := M_n(\mathbb{C})$ une matrice fixée. Montrer que $u : M \mapsto AM - MA$ est un endomorphisme de E. Montrer que u est nilpotent lorsque A est nilpotente, et que u est diagonalisable si, et seulement si, la matrice A est diagonalisable.

I.5.38 $*$ Soit $n \in \mathbb{N}^*$. Déterminer les couples (a, M) formés d'un nombre complexe $a \notin \{0, 1\}$ et d'une matrice $M \in M_n(\mathbb{C})$ tels que aM soit semblable à M (distinguer deux cas, suivant que M soit ou ne soit pas nilpotente).

Groupes classiques I.6

Ce module n'est pas centré sur une théorie, des concepts, des méthodes, mais sur des *objets* : le groupe linéaire et certains de ses sous-groupes les plus remarquables. Ces groupes sont apparus lorsque Felix Klein, dans le *programme d'Erlangen*, a fait remarquer que toute géométrie (euclidienne, affine, projective ...) pouvait être caractérisée par un groupe de transformations qui préservait ses caractéristiques. Klein a même renversé la perspective en affirmant qu'une géométrie était définie comme ce qui était préservé par un tel groupe. Depuis le livre éponyme de Hermann Weyl, ces groupes ont été baptisés *groupes classiques*. Ils sont d'un intérêt fondamental pour les mathématiciens de toute obédience, mais également pour les physiciens : en effet, ce qu'il y a eu de plus novateur dans la physique du XX$^{\text{e}}$ siècle (relativité, physique quantique) est décrit par la géométrie des groupes classiques.
L'étude comporte un volet généraliste : sur un corps K quelconque (même fini !), on étudie algébriquement et géométriquement le groupe linéaire et le groupe spécial linéaire. *Toutes* les ressources de l'algèbre générale (y compris du module suivant), de l'algèbre linéaire et de la géométrie sont mises en jeu. L'adoption d'un point de vue géométrique signifie essentiellement que l'on s'intéresse à l'action de ces groupes sur divers ensembles. Lorsque l'on spécialise l'étude au cas des corps $\mathbb{R}$ et $\mathbb{C}$, on utilise même de l'analyse (topologie, séries entières, calcul différentiel). Enfin, lorsque l'on s'intéresse au groupe orthogonal et au groupe unitaire, on a, bien entendu, besoin de l'algèbre bilinéaire. Notons que l'étude du *groupe symplectique*, essentielle en géométrie et en mécanique, mais plus délicate, est reportée dans [MPA1], chapitre d'algèbre multilinéaire.

1 Le groupe linéaire $GL_n(K)$

À titre de mise en route, considérons un espace vectoriel E de dimension finie n sur le corps K. Pour tout choix d'une base $\mathcal{B}$ de E, on définit un isomorphisme $\varphi_{\mathcal{B}}$ du groupe $\mathcal{GL}(E)$ sur le groupe $GL_n(K)$: à tout automorphisme f de E, on associe sa matrice $\varphi_{\mathcal{B}}(f) \in GL_n(K)$ dans la base $\mathcal{B}$.

Exercice 1.
Soit $\mathcal{B}'$ une autre base de E. Décrire l'automorphisme $\varphi_{\mathcal{B}'} \circ \varphi_{\mathcal{B}}^{-1}$ de $GL_n(K)$ à l'aide de la matrice de passage P de $\mathcal{B}$ à $\mathcal{B}'$.
Solution. Si M et M' sont respectivement les matrices de l'automorphisme f dans les bases $\mathcal{B}$ et $\mathcal{B}'$, on a $M' = P^{-1}MP$.

Ainsi, $\varphi_{\mathcal{B}'}(f) = P^{-1}\varphi_{\mathcal{B}}(f)P$ et $\varphi_{\mathcal{B}'} \circ \varphi_{\mathcal{B}}^{-1}(M) = P^{-1}MP$: l'automorphisme $\varphi_{\mathcal{B}'} \circ \varphi_{\mathcal{B}}^{-1}$ de $GL_n(K)$ est l'automorphisme *intérieur* $M \mapsto P^{-1}MP$.

Dans la plupart des cas, nous donnerons de nos résultats soit une preuve « matricielle », concernant $GL_n(K)$ ou $SL_n(K)$, soit une preuve « géométrique », concernant $\mathcal{GL}(E)$ ou $\mathcal{SL}(E)$, considérant la transposition de l'une à l'autre formulation comme allant de soi.

1.1 Propriétés algébriques

L'étude de l'algorithme du pivot dans le module de calcul matriciel de [L1] a mis en lumière le rôle particulier joué par les matrices de transvection et de dilatation. Rappelons des notations de ce module : la *matrice élémentaire* dont le seul coefficient non nul est un 1 en position (i,j) est notée $E_{i,j}^{(n,n)} \in M_n(K)$; la *matrice de transvection* $I_n + \alpha E_{i,j}^{(n,n)}$ ($\alpha \in K$, $i \neq j$) est notée $T_{i,j}^{(n)}(\alpha)$; et la *matrice de dilatation* $I_n + (\lambda - 1)E_{i,i}^{(n,n)}$ ($\lambda \in K^*$) est notée $D_i^{(n)}(\lambda)$: c'est donc la matrice diagonale $\mathrm{Diag}(1, \ldots, \lambda, \ldots, 1)$, où λ est en $i^{ème}$ position. Nous abrégerons ces notations en $E_{i,j}$, $T_{i,j}(\alpha)$ et $D_i(\lambda)$, sous-entendant la dimension n. Voici la contrepartie géométrique de ces notions.

Définition 1. Soient H un hyperplan du K-espace vectoriel E de dimension finie $n \geqslant 2$ et f est une forme linéaire de noyau H. Une *transvection d'hyperplan* H est un endomorphisme de E de la forme $x \mapsto x + f(x)u$, où $u \in H$. La *transvection triviale* est l'identité (cas où $u = 0$). Une *dilatation d'hyperplan* H est un automorphisme de E de la forme $x \mapsto x + f(x)u$, où $u \notin H$.

Une transvection est nécessairement un automorphisme : sa restriction à H est l'identité, et tout élément du noyau est un multiple de u, donc dans H. Une transvection non triviale admet pour espace fixe (espace propre associé à la valeur propre 1) l'hyperplan H, et elle n'admet pas d'autres valeurs et vecteurs propres. Un argument similaire montre que, si $u \notin H$, l'endomorphisme $x \mapsto x + f(x)u$ est bijectif (donc une dilatation) si, et seulement si, $f(u) \neq -1$. Une telle dilatation admet pour espace fixe l'hyperplan H, et une unique droite propre supplémentaire de H, la droite $\mathrm{Vect}(u)$. la valeur propre correspondante est $1 + f(u)$ est appelée *rapport* de la dilatation.

Proposition 1. (i) Si la matrice de $\varphi \in \mathcal{L}(E)$ dans une base de E est une matrice de transvection, alors φ est une transvection. Réciproquement, si φ est une transvection autre que l'identité, il existe une base de E dans laquelle la matrice de φ est $T_{n-1,n}(1)$.
(ii) Si la matrice de $\varphi \in \mathcal{L}(E)$ dans une base de E est une matrice de dilatation $D_i(\lambda) \neq I_n$, alors φ est une dilatation de rapport λ. Réciproquement, si φ est une dilatation de rapport $\lambda \in K \setminus \{0,1\}$, il existe une base de E dans laquelle la matrice de φ est $D_n(\lambda)$.

Démonstration.
(i) Si la matrice de φ dans la base $(e_1, \ldots, e_n)$ est $T_{i,j}(\alpha)$, il suffit de prendre pour f la forme linéaire $j^{ème}$ coordonnée et pour u le vecteur αe_i. Réciproquement, si φ est

l'application $x \mapsto x + f(x)u$ (avec $u \neq 0$), on complète u en une base $e_1, \ldots, e_{n-1} := u$ de l'hyperplan $H := \mathrm{Ker}\, f$ et l'on choisit $e_n \in E$ tel que $f(e_n) = 1$.

(ii) Si la matrice de φ dans la base $(e_1, \ldots, e_n)$ est $D_i(\lambda)$, il suffit de prendre pour f la forme linéaire $i^{ème}$ coordonnée et pour u le vecteur $(\lambda - 1)e_i$. Réciproquement, si φ est la dilatation $x \mapsto x + f(x)u$, on complète une base $e_1, \ldots, e_{n-1}$ de l'hyperplan $H := \mathrm{Ker}\, f$ par $e_n := u$. La matrice de φ dans cette base est $D_n(\lambda)$ avec $\lambda = 1 + f(u)$. ∎

Corollaire 2.
Les transvections non triviales forment une classe de conjugaison (ou classe de similitude) dans $\mathcal{GL}(E)$. Pour tout $\lambda \in K \setminus \{0, 1\}$, les dilatations de rapport λ forment une classe de conjugaison dans $\mathcal{GL}(E)$ (voir aussi l'exercice I.6.1 de la page 307).

Lemme 3. Toute matrice de $GL_n(K)$ peut être transformée, par des transvections sur les colonnes en une matrice de la forme $\mathrm{Diag}(1, \ldots, 1, d)$. Le scalaire d est le déterminant de la matrice de départ.

Démonstration. Si $n = 1$, il n'y a rien à démontrer. Supposons donc $n \geqslant 2$. Pour les *opérations élémentaires* qui suivent, nous reprenons les notations du module de calcul matriciel de [L1], section 2. Puisque la matrice est inversible, un coefficient non nul au moins est présent sur la première ligne. Si c'est $a_{1,j}$, $j \neq 1$, l'opération $C_1 \leftarrow C_1 + \dfrac{1 - a_{1,1}}{a_{1,j}} C_j$ met un 1 en position $(1, 1)$; si seul $a_{1,1}$ est non nul, on commence par faire $C_2 \leftarrow C_2 + C_1$ et l'on est ramené au cas précédent. Partant maintenant d'une matrice telle que $a_{1,1} = 1$, les transvections $C_j \leftarrow C_j - a_{1,j}C_1$ (pour $2 \leqslant j \leqslant n$) nous donnent une matrice dont la première ligne est $(1, 0, \ldots, 0)$.

On applique le même procédé itérativement (selon le même principe que dans l'algorithme du pivot) et l'on arrive à une matrice triangulaire inférieure dont la diagonale est $\mathrm{Diag}(1, \ldots, 1, d)$ (le dernier coefficient correspond au cas $n = 1$, et l'on ne peut le transformer).

Une fois la forme triangulaire obtenue, des transvections successives $C_1 \leftarrow C_1 + \alpha C_j$ (pour j allant de 2 à n, *dans cet ordre*) permettent d'annuler toute la première colonne (sauf $a_{1,1} = 1$) sans modifier la diagonale ni la forme triangulaire inférieure. Puis on recommence en transformant la deuxième colonne, etc.

Toutes ces transformations conservent le déterminant, d'où la dernière assertion. ∎

Théorème 4.
(i) Le groupe spécial linéaire $SL_n(K)$ est engendré par les $T_{i,j}(\alpha)$. Le groupe spécial linéaire $\mathcal{SL}(E)$ est engendré par les transvections.
(ii) Le groupe linéaire $GL_n(K)$ est engendré par les $T_{i,j}(\alpha)$ et les $D_i(\lambda)$. Le groupe linéaire $\mathcal{GL}(E)$ est engendré par les transvections et les dilatations.

Démonstration.
Du lemme et de l'interprétation des opérations sur les lignes et les colonnes dans [L1] (module II.4, avant la proposition 17) découle que toute matrice $A \in GL_n(K)$ est de la forme $A = \mathrm{Diag}(1, \ldots, 1, \det A)B$, où B est un produit de matrices de transvection. Si $\det A = 1$, on en tire la première assertion. La deuxième est immédiate. ∎

Théorème 5. (i) Le centre de $GL_n(K)$ est le sous-groupe K^*I_n formé des matrices scalaires non nulles. Le centre de $\mathcal{GL}(E)$ est $Z\big(\mathcal{GL}(E)\big) = K^* \mathrm{Id}_E$.
(ii) Le centre de $SL_n(K)$ est le sous-groupe $\mu_n(K)I_n$ formé des matrices scalaires dont le coefficient est une racine $n^{ème}$ de l'unité.
Le centre de $\mathcal{SL}(E)$ est $Z\big(\mathcal{SL}(E)\big) = \mu_n(K)\,\mathrm{Id}_E$.

Démonstration. Rappelons que le *centre* d'un groupe G est le sous-groupe distingué $Z(G) := \{g \in G \mid \forall h \in G \ , \ gh = hg\}$. Si une matrice $M \in M_n(K)$ commute avec toutes les matrices de $SL_n(K)$, elle commute en particulier avec toutes les matrices $I_n + E_{i,j}$ telles que $i \neq j$, donc avec les $E_{i,j}$ (puisqu'elle commute avec I_n). Le calcul du centre de $M_n(K)$ (module de calcul matriciel de [L1], exercice 6) avait permis de montrer que la condition $\forall i \neq j \ , \ ME_{i,j} = E_{i,j}M$ implique que M est scalaire : $M = \lambda I_n$. Bien entendu, $\lambda I_n \in GL_n(K)$ équivaut à $\lambda \neq 0$ et $\lambda I_n \in SL_n(K)$ équivaut à $\lambda^n = 1$. ∎

Proposition et définition 6. On appelle *groupe projectif* le groupe quotient d'un groupe linéaire par son centre : $\mathbf{P}GL_n(K) := GL_n(K)/Z\big(GL_n(K)\big)$, et $\mathbf{P}\mathcal{GL}(E) := \mathcal{GL}(E)/Z\big(\mathcal{GL}(E)\big)$; et *groupe spécial projectif* le groupe quotient d'un groupe spécial linéaire par son centre : $\mathbf{P}SL_n(K) := SL_n(K)/Z\big(SL_n(K)\big)$ et $\mathbf{P}\mathcal{SL}(E) := \mathcal{SL}(E)/Z\big(\mathcal{SL}(E)\big)$ (cette terminologie sera expliquée à la section 1.2.4 de la page 270).
Le groupe spécial projectif s'identifie à un sous-groupe du groupe projectif.

Démonstration. C'est un cas particulier de l'injection de $S/(S \cap H)$ dans G/H lorsque S et H sont des sous-groupes de G (H étant de plus distingué), avec $G := GL_n(K)$, $S := SL_n(K)$, $H := Z\big(GL_n(K)\big)$ et $S \cap H = SL_n(K) \cap Z\big(GL_n(K)\big) = Z\big(SL_n(K)\big)$ (d'après le théorème). ∎

Corollaire 7. Lorsque le corps K est algébriquement clos, $\mathbf{P}GL_n(K) = \mathbf{P}SL_n(K)$.

Démonstration. L'image de $\mathbf{P}SL_n(K)$ dans $\mathbf{P}GL_n(K)$ est la projection dans $\mathbf{P}GL_n(K)$ du sous-groupe $K^*SL_n(K) \subset GL_n(K)$. Soient $M \in GL_n(K)$, $\lambda \in K^*$ tel que $\lambda^n = \det M$ et $N := \lambda^{-1}M$. Alors $N \in SL_n(K)$ et $M = \lambda N \in K^*SL_n(K)$, d'où l'égalité. ∎

Exercice 2.
Qu'en est-il lorsque $K := \mathbb{R}$?
Solution. Si n est impair, il existe, pour toute matrice $M \in GL_n(\mathbb{R})$ un scalaire $\lambda \in \mathbb{R}^*$ tel que $\lambda^n = \det M$, d'où $N := \lambda^{-1}N \in SL_n(\mathbb{R})$ et $M = \lambda N \in \mathbb{R}^* SL_n(\mathbb{R})$, de sorte que $\mathbf{P}SL_n(\mathbb{R}) = \mathbf{P}GL_n(\mathbb{R})$. Si n est pair, un calcul similaire montre que $M \in \mathbb{R}^* SL_n(\mathbb{R})$ si, et seulement si, $\det M > 0$. L'image de $\mathbf{P}SL_n(\mathbb{R})$ dans $\mathbf{P}GL_n(\mathbb{R})$ est donc, dans ce cas, le sous-groupe d'indice 2 formé des classes des matrices à déterminant positif.

Dans un groupe G, on appelle *commutateur* de deux éléments g et h l'élément :

$$[g,h] := ghg^{-1}h^{-1}.$$

Le sous-groupe de G engendré par les commutateurs est appelé *groupe dérivé de* G et noté $D(G)$. Ses propriétés sont étudiées dans l'exercice I.6.4 de la page 307.

Théorème 8. Soit $n \geqslant 3$. Alors $D\big(GL_n(K)\big) = D\big(SL_n(K)\big) = SL_n(K)$ (pour le cas $n := 2$, voir l'exercice I.6.5 de la page 307).

Démonstration. Comme le déterminant de $[g,h] = ghg^{-1}h^{-1}$ est 1, les inclusions :

$$D\big(GL_n(K)\big) \subset SL_n(K) \qquad \text{et} \qquad D\big(SL_n(K)\big) \subset SL_n(K)$$

sont évidentes.
Soient $i,j,k \in [\![1,n]\!]$ deux à deux distincts. Les matrices élémentaires se multiplient selon la règle : $E_{i,j}E_{k,l} = \delta_{j,k}E_{i,l}$. On en déduit d'abord, par des calculs simples, que $\big(T_{i,j}(\alpha)\big)^{-1} = T_{i,j}(-\alpha)$, puis que $[T_{i,j}(\alpha), T_{j,k}(\beta)] = T_{i,k}(\alpha\beta)$. Ainsi, toute transvection appartient à $D\big(SL_n(K)\big)$. D'après le théorème 4, on a donc $SL_n(K) \subset D\big(SL_n(K)\big)$ et, *a fortiori*, $SL_n(K) \subset D\big(GL_n(K)\big)$. ∎

1.2 Propriétés géométriques

Dans toute cette section, on fixe un K-espace vectoriel E de dimension $n \in \mathbb{N}^*$.

1.2.1 Actions sur des vecteurs et des sous-espaces

L'application $(\varphi, x) \mapsto \varphi(x)$ de $\mathcal{GL}(E) \times E$ dans E définit une action de $\mathcal{GL}(E)$. Comme $\varphi \in \mathcal{GL}(E)$ et $x \neq 0 \Rightarrow \varphi(x) \neq 0$, on voit que $E \setminus \{0\}$ est stable par cette action. Si φ est dans le noyau de l'action, $\varphi(x) = x$ pour tout $x \neq 0$, et il est clair que $\varphi = \mathrm{Id}_E$. L'action de $\mathcal{GL}(E)$ sur $E \setminus \{0\}$ est donc fidèle. Soient d'autre part x, y deux éléments de $E \setminus \{0\}$. Soient $\mathcal{B} := (e_1, \ldots, e_n)$ et $\mathcal{C} := (f_1, \ldots, f_n)$ deux bases de E ayant respectivement pour premier élément $e_1 := x$ et $f_1 := y$. L'application $e_i \mapsto f_i$ est un élément de $\mathcal{GL}(E)$ qui envoie x sur y. L'action est donc transitive.

Exercice 3.
Décrire matriciellement le stabilisateur d'un élément fixé.
Solution. Soit $x \in E \setminus \{0\}$ fixé. Prenons à nouveau une base $\mathcal{B} := (e_1, \ldots, e_n)$ de E telle que $e_1 := x$. Cette base fournit un isomorphisme de $\mathcal{GL}(E)$ sur $GL_n(K)$. L'image

par cet isomorphisme du stabilisateur de x est le sous-groupe de $GL_n(K)$ constitué des matrices de la forme $M_{X,A} := \begin{pmatrix} 1 & X \\ 0 & A \end{pmatrix}$, où $X \in M_{1,n-1}(K)$ et $A \in GL_{n-1}(K)$ (et $0 \in M_{n-1,1}(K)$ désigne un vecteur colonne). La multiplication dans ce groupe se fait selon la règle : $M_{X,A}M_{Y,B} = M_{Y+XB,AB}$. Notons que, l'action étant transitive, tous les stabilisateurs sont conjugués les uns des autres, ce que confirme la description ci-dessus.

Pour tout $r \in \mathbb{N}^*$, l'action de $\mathcal{GL}(E)$ sur E s'étend naturellement en une action sur l'ensemble E^r des familles de r vecteurs. Si de plus, $1 \leqslant r \leqslant n$, il est clair que le sous-ensemble $L_r(E) \subset E^r$ formé des familles *libres* est stable pour cette action.

Proposition 9. L'action de $\mathcal{GL}(E)$ sur $L_r(E)$ est fidèle est transitive.

Démonstration. Puisque tout $x \in E\setminus\{0\}$ peut-être prolongé en une famille libre de r éléments, tout φ du noyau de l'action fixe tous les vecteurs non nuls, donc est égal à l'identité : l'action est bien fidèle. Soient $(x_1, \dots, x_r)$ et $(y_1, \dots, y_r)$ deux familles libres. On les prolonge en des bases $\mathcal{B} := (x_1, \dots, x_n)$ et $\mathcal{C} := (y_1, \dots, y_n)$. L'application $x_i \mapsto y_i$ est un élément de $\mathcal{GL}(E)$ qui envoie la première famille sur la deuxième, d'où la transitivité. ∎

Exercice 4.
Décrire matriciellement les stabilisateurs.

Solution. Prolongeant $(x_1, \dots, x_r) \in L_r(E)$ en une base $(x_1, \dots, x_n)$, on voit que le stabilisateur de $(x_1, \dots, x_r)$ est isomorphe au sous-groupe de $GL_n(K)$ constitué des matrices de la forme $M_{X,A} := \begin{pmatrix} I_r & X \\ 0 & A \end{pmatrix}$, où $X \in M_{r,n-r}(K)$ et $A \in GL_{n-r}(K)$ (et $0 \in M_{n-r,r}(K)$ désigne un vecteur colonne et I_r la matrice identité de rang r). La multiplication dans ce groupe se fait encore selon la règle : $M_{X,A}M_{Y,B} = M_{Y+XB,AB}$.

On retrouve en corollaire le résultat suivant (remarque de la page 71) : le groupe $\mathcal{GL}(E)$ agit simplement transitivement sur l'ensemble $L_n(E)$ des bases de E. Ainsi, pour tout choix d'une base $\mathcal{B}_0$ particulière de E, on obtient une bijection $\varphi \mapsto \varphi(\mathcal{B}_0)$ de $\mathcal{GL}(E)$ sur $L_n(E)$. Une application au groupe linéaire d'un corps fini avait été proposée dans la proposition 24 de la page 85, on la complète dans l'exercice I.6.8 de la page 307.

Application : orientation des espaces vectoriels réels. On se place dans le cas où le corps de base est $\mathbb{R}$. Le sous-groupe $\mathcal{GL}^+(E)$ formé des automorphismes de déterminant positif est alors un sous-groupe d'indice 2 de $\mathcal{GL}(E)$. L'action de $\mathcal{GL}^+(E)$ est simple mais n'est plus transitive : il y a deux orbites. Pour que deux bases soient dans la même orbite, il faut, et il suffit, que la matrice de passage entre ces bases ait un déterminant positif. On appelle *orientation* de E le choix d'une de ces deux orbites. Choisir une orientation revient donc à déclarer que certaines bases sont *directes* (celles de l'orbite choisie) et d'autres *indirectes*. Par exemple, lorsque $E := \mathbb{R}^n$, l'orientation de la base canonique est appelée *orientation canonique* : les bases directes de $\mathbb{R}^n$ sont celles de déterminant positif.

L'image d'un sous–espace vectoriel F de E par un élément φ de $\mathcal{GL}(E)$ est un sous–espace vectoriel de même dimension. On obtient ainsi, pour $0 \leqslant r \leqslant n$, une action de $\mathcal{GL}(E)$ sur l'ensemble $S_r(E)$ des sous-espaces vectoriels de dimension r. Cette action est transitive : pour $r := 0$ et $r := n$, l'ensemble $S_r(E)$ n'a qu'un élément (respectivement, $\{0\}$ et E) et c'est trivial. Si $r \in [\![1, n-1]\!]$, soient $F, G \in S_r(E)$. Par choix de bases, en appliquant la proposition 9, on trouve $\varphi \in \mathcal{GL}(E)$ tel que $\varphi(F) = G$.

Exercice 5.

Quel est le noyau de cette action ?

Solution. Si φ est dans le noyau, il laisse stable tout sous–espace vectoriel de dimension r. Supposons d'abord que $r := 1$. On a déjà vu, dans le module d'algèbre linéaire de [L1] (preuve du théorème 41), que φ est une homothétie dans ce cas. Prenons maintenant $r \in [\![2, n-1]\!]$ quelconque. Toute droite est intersection des sous-espaces vectoriels de dimension r qui la contiennent (exercice amusant, laissé au lecteur). Ainsi, φ laisse fixe toute droite et l'on est ramené au cas précédent. On conclut donc que le noyau de l'action est $K^* \operatorname{Id}_E$.

Drapeaux

Soit $d := (d_1, \ldots, d_k)$, où $1 \leqslant d_1 < \cdots < d_k \leqslant n$. On appelle *$d$-drapeau* de E un k-uplet $(F_1, \ldots, F_k)$ de sous-espaces vectoriels de E tels que $F_1 \subset \cdots \subset F_k$ et que $\dim F_1 = d_1, \ldots, \dim F_k = d_k$. Cette définition étend celle de l'exercice I.2.0 de la page 88, où il seulement de *drapeaux maximaux*, *i.e.* tels que $d := (1, \ldots, n)$ (*cf.* l'exercice I.6.10 de la page 308). Puisque l'application d'un automorphisme préserve les dimensions et les relations d'inclusion, le groupe $\mathcal{GL}(E)$ agit sur l'ensemble $D_d(E)$ des d-drapeaux de E.

Proposition 10. L'action de $\mathcal{GL}(E)$ sur $D_d(E)$ est transitive. Son noyau est $K^* \operatorname{Id}_E$.

Démonstration. Appelons *base adaptée* au d-drapeau $D := (F_1, \ldots, F_k)$ une base $(e_1, \ldots, e_n)$ de E telle que, pour $1 \leqslant i \leqslant k$, la famille $(e_1, \ldots, e_{d_i})$ soit une base de F_i. Il est clair qu'une telle base adaptée existe (étendre successivement de F_{i-1} à F_i). De plus, si $(e'_1, \ldots, e'_n)$ est une base adaptée au d-drapeau $D' := (F'_1, \ldots, F'_k)$, l'application $e_i \mapsto e'_i$ est un élément de $\mathcal{GL}(E)$ qui envoie D sur D', d'où la transitivité de l'action. Il est clair que $K^* \operatorname{Id}_E$ agit trivialement sur les d-drapeaux, Comme tout sous-espace vectoriel de dimension d_1 peut être prolongé en un d-drapeau (exercice amusant, laissé au lecteur), on déduit de l'exercice ci-dessus que le noyau de l'action est bien $K^* \operatorname{Id}_E$. ∎

1.2.2 Représentations linéaires

Définition 2. Une *représentation linéaire* du groupe G dans le K-espace vectoriel de dimension finie E est une action de G sur E telle que, pour tout $g \in G$, l'application $x \mapsto g.x$ est un automorphisme de E.

Il revient au même de dire que le morphisme de groupes $\rho : G \to \mathfrak{S}(E)$ associé à cette action (proposition 1 de la page 58) a son image dans $\mathcal{GL}(E)$. On identifiera donc une représentation linéaire de G à un morphisme $\rho : G \to \mathcal{GL}(E)$. Un sous–espace vectoriel F de E est dit *stable* si c'est une partie stable pour cette action : par exemple E lui-même, le sous-espace $\{0\}$ ou toute

droite qui est propre pour tous les $\rho(g)$ à la fois. On appelle alors *représentation induite sur* F l'action de G sur F (qui est, bien entendu, une représentation linéaire). La représentation est dite *simple* si E n'admet pas d'autre sous-espace stable que $\{0\}$ et lui-même et *irréductible* si E n'est pas somme directe de sous-espaces stables non triviaux et *complètement réductible* si E est somme directe de sous-espaces stables tels que les représentations induites sont simples.

Exemples.
(i) Soit $\varphi \in \mathcal{L}(E)$ qui commute à l'action de G, autrement dit, $\forall g \in G\ ,\ \rho(g) \circ \varphi = \varphi \circ \rho(g)$. Alors $\operatorname{Ker}\varphi$ et $\operatorname{Im}\varphi$ sont stables, ainsi que tous les espaces propres de φ.
(ii) Le morphisme $a \mapsto \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$ du groupe additif K dans $GL_2(K)$ définit une représentation de K dans K^2. Le seul sous-espace stable (autre que K^2 et $\{0\}$) est la droite $K \times \{0\}$. Cette représentation est donc irréductible, mais pas simple (ni, évidemment, complètement réductible).
(iii) On suppose que la caractéristique de K n'est pas 2. Soit s une symétrie de E (c'est-à-dire un endomorphisme involutif, voir [L1], module II.3). L'application de $\mathbb{Z}/2\mathbb{Z}$ dans $\mathcal{GL}(E)$ qui, à 0 associe Id_E et à 1 associe s, est un morphisme de groupes, donc une représentation linéaire. Si $\dim E = 2$ et si $s \neq \pm \mathrm{Id}_E$, l'axe $\operatorname{Ker}(s - \mathrm{Id}_E)$ et la direction $\operatorname{Ker}(s + \mathrm{Id}_E)$ de s sont les seuls sous-espaces stables non triviaux. Comme ce sont des droites, les représentations associées sont simples. Comme $E = \operatorname{Ker}(s - \mathrm{Id}_E) \oplus \operatorname{Ker}(s + \mathrm{Id}_E)$, la représentation de départ est complètement réductible.

Pour généraliser ce dernier exemple, nous supposerons désormais que G est un groupe fini dont l'ordre $N := \operatorname{card} G$ n'est pas un multiple de la caractéristique du corps K : autrement dit, l'image de N dans K est inversible. Le procédé de *moyennisation* ci-dessous a déjà été rencontré (il est implicite dans l'exercice 9 de la page 73). Pour tout $\varphi \in \mathcal{L}(E)$, on définit :

$$\int_G \varphi := \frac{1}{N} \sum_{g \in G} \rho(g) \circ \varphi \circ \rho(g)^{-1}.$$

Nous allons voir que $\varphi' := \displaystyle\int_G \varphi$ commute à l'action de G. Fixons $g_0 \in G$. Alors $g \mapsto g_0 g$ est une bijection de G sur lui-même, ce qui justifie le calcul suivant :

$$\begin{aligned} \rho(g_0) \circ \varphi' \circ \rho(g_0)^{-1} &= \rho(g_0) \circ \Big(\frac{1}{N} \sum_{g \in G} \rho(g) \circ \varphi \circ \rho(g)^{-1} \Big) \circ \rho(g_0)^{-1} \\ &= \frac{1}{N} \sum_{g \in G} \big(\rho(g_0) \circ \rho(g) \circ \varphi \circ \rho(g)^{-1} \circ \rho(g_0)^{-1} \big) \\ &= \frac{1}{N} \sum_{g \in G} \rho(g_0 g) \circ \varphi \circ \rho(g_0 g)^{-1} = \varphi', \end{aligned}$$

d'où l'égalité $\rho(g_0) \circ \varphi' = \varphi' \circ \rho(g_0)$.

Soit maintenant p un projecteur d'image F stable. Nous allons voir que $p' := \displaystyle\int_G p$ est un projecteur d'image F. En notation allégée, on a : $p'(x) = \dfrac{1}{N} \sum\limits_{g \in G} g.p(g^{-1}.x)$. Puisque chaque $p(g^{-1}.x)$ appartient à F, qui est stable, chaque $g.p(g^{-1}.x)$ appartient à F, donc aussi $p'(x)$: on a donc $\operatorname{Im} p' \subset F$. Par ailleurs, si $x \in F$, alors $g^{-1}.x \in F \Rightarrow p(g^{-1}.x) = g^{-1}.x$, et l'on a $p'(x) = \dfrac{1}{N} \sum\limits_{g \in G} g.g^{-1}.x = x$. Ainsi, la restriction de p' à F est Id_F et l'on voit que p' est bien un projecteur d'image F.

Théorème 11 (Théorème de Maschke).
Soit G un groupe fini dont l'ordre $N := \operatorname{card} G$ n'est pas un multiple de la caractéristique du corps K. Toute représentation K-linéaire de G est complètement réductible.

Démonstration. Prouvons d'abord que, pour toute représentation de G dans un K-espace vectoriel E, tout sous-espace stable F de E admet un supplémentaire stable. Soit en effet un projecteur p quelconque sur F. Le projecteur $p' := \int_G p$ a pour image F et commute à tous les $\rho(g)$. Son noyau est donc un supplémentaire stable de F. Le théorème en découle alors par récurrence forte sur la dimension. Pour $\dim E := 0$ ou $\dim E := 1$, la représentation est simple donc complètement réductible. Supposons $\dim E \geqslant 2$. Si la représentation est simple, elle est complètement réductible ; sinon, soit F un sous-espace stable distinct de E et de $\{0\}$; soit G un supplémentaire stable de F ; les sous-espaces F et G sont de dimension strictement inférieure à $\dim E$, on peut donc leur appliquer l'hypothèse de récurrence et la conclusion est alors immédiate. ∎

L'hypothèse sur la caractéristique est d'ailleurs indispensable, comme le montre le deuxième exemple de la page 268 appliqué au cas du corps fini à deux éléments $K := \mathbf{F}_2$: le groupe G est alors le groupe additif K, c'est-à-dire $\mathbb{Z}/2\mathbb{Z}$.

1.2.3 Le groupe affine

Rappelons quelques acquis du module I.2 (section 1.2 de la page 62) sur les espaces affines. Soient E un K-espace vectoriel de dimension n et X un espace affine de direction E. Les transformations affines de X forment le groupe affine $\mathcal{G}A(X)$. L'application qui, à toute transformation affine $f \in \mathcal{G}A(X)$, associe sa partie linéaire $\vec{f} \in \mathcal{GL}(E)$ est un morphisme de groupes surjectif, que nous noterons ici $\pi : \mathcal{G}A(X) \to \mathcal{GL}(E)$. Le noyau de π est le sous-groupe distingué $\mathcal{T}(X)$. Il est formé des translations $t_{\vec{u}}$ de vecteur $\vec{u} \in E$, et l'on a un isomorphisme de groupes $\tau : E \to \mathcal{T}(X)$, $\vec{u} \mapsto t_{\vec{u}}$. Le groupe $\mathcal{G}A(X)$ agit fidèlement transitivement sur X. Pour tout $O \in X$, notons $\mathcal{G}A_O(X)$ le stabilisateur de O pour cette action : l'action étant transitive, ces stabilisateurs sont tous conjugués. Nous noterons I_f l'automorphisme intérieur $g \mapsto f \circ g \circ f^{-1}$ de $\mathcal{G}A(X)$.

Proposition 12. (i) Soit $O \in X$. La restriction de π à $\mathcal{G}A_O(X)$ est un isomorphisme de $\mathcal{G}A_O(X)$ sur $\mathcal{GL}(E)$. On notera σ_O l'isomorphisme réciproque.
(ii) Soit de plus $O' \in X$. Alors $\sigma_{O'} = I_{t_{\overrightarrow{OO'}}} \circ \sigma_O$.

Démonstration. (i) Si $f, g \in \mathcal{G}A_O(X)$ ont même image par π, la transformation $g^{-1} \circ f$ est une translation qui fixe O, donc l'identité : la restriction de π à $\mathcal{G}A_O(X)$ est donc injective. Si maintenant $\varphi \in \mathcal{GL}(E)$ et si $g \in \mathcal{G}A(X)$ est un antécédent quelconque de φ par π, alors $f := \left(t_{\overrightarrow{Og(O)}}\right)^{-1} \circ g$ est un antécédent de φ par π dans $\mathcal{G}A_O(X)$.

(ii) Soient $\varphi \in \mathcal{GL}(E)$ et $f := \sigma_O(\varphi)$. La transformation $f' := t_{\overrightarrow{OO'}} \circ f \circ \left(t_{\overrightarrow{OO'}}\right)^{-1}$ a même partie linéaire que f, et $f'(O') = O'$; c'est donc un antécédent de φ par π dans $\mathcal{G}A_{O'}(X)$, donc $f' = \sigma_{O'}(\varphi)$. On a bien $\sigma_{O'}(\varphi) = I_{t_{\overrightarrow{OO'}}}\big(\sigma_O(\varphi)\big)$. Il en découle que $\mathcal{G}A_{O'}(X) = I_{t_{\overrightarrow{OO'}}}\big(\mathcal{G}A_O(X)\big)$, et l'on retrouve le fait que ces groupes sont conjugués. ∎

On peut donc dire que, modulo le choix d'une « origine » $O \in X$, le groupe linéaire $\mathcal{GL}(E)$ s'identifie à un sous-groupe de $\mathcal{G}A(X)$: mais ce sous-groupe et cette identification *dépendent essentiellement du choix de l'origine* O. Une fois ce choix effectué, on obtient donc une bijection de $\mathcal{GL}(E) \times E$ sur $\mathcal{G}A(X)$ en associant à $(\varphi, \vec{u})$ l'unique transformation affine de partie linéaire φ et telle que $O \mapsto O+\vec{u}$, autrement dit, la transformation $t_{\vec{u}} \circ \sigma_O(\varphi)$. *Il ne s'agit pas d'un isomorphisme de groupes.* La loi de groupe sur $\mathcal{GL}(E) \times E$ obtenue en *transportant* la loi de groupe de $\mathcal{G}A(X)$ par cette bijection ci-dessus est la suivante :

$$(\varphi', \vec{u}') \star (\varphi, \vec{u}) = \big(\varphi' \circ \varphi, \vec{u}' + \varphi'(\vec{u})\big).$$

Il suffit, pour s'en assurer, de considérer la partie linéaire d'une part, l'effet sur O d'autre part. Le choix d'une origine $O \in X$ permet d'identifier X à E : l'application $\vec{u} \mapsto O + \vec{u}$ est un isomorphisme affine de E sur X. On peut alors identifier $\mathcal{G}A(X)$ à $\mathcal{GL}(E) \times E$ (muni de la loi décrite plus haut). Ensuite, le choix d'une base $\mathcal{B} := (\vec{u}_1, \dots, \vec{u}_n)$ de E permet d'identifier E à K^n : l'application $(x_1, \dots, x_n) \mapsto \sum x_i \vec{u}_i$ est un isomorphisme d'espaces vectoriels de K^n sur E. On peut alors identifier $\mathcal{GL}(E)$ à $GL_n(K)$. Le choix d'un *repère* $(O; \vec{u}_1, \dots, \vec{u}_n)$ formé d'une origine O de X et d'une base $\mathcal{B} := (\vec{u}_1, \dots, \vec{u}_n)$ de E permet donc d'identifier X à K^n : l'application $(x_1, \dots, x_n) \mapsto O + \sum x_i \vec{u}_i$ est un isomorphisme d'espaces affines de K^n sur X. On peut alors identifier $\mathcal{G}A(X)$ à $GL_n(K) \times K^n$, le couple (A, B) correspondant à la transformation affine $U \mapsto AU + B$ (avec la convention usuelle que les éléments B et U de K^n sont écrits comme des vecteurs colonnes). Le groupe affine $GL_n(K) \times K^n$ est muni de la loi suivante :

$$(A', B') \star (A, B) = (A'A, A'B + B').$$

Lorsque $E := K$, et que l'on prend $O := 0$, on retrouve la description du groupe affine $\mathcal{G}A(K)$ comme ensemble des applications $x \mapsto ax + b$ $(a, b) \in K^* \times K$ et la loi correspondante sur $K^* \times K$ est la suivante : $(a', b') \star (a, b) = (a'b', b' + a'b)$. Cette loi avait été trouvée dans le cas particulier $K := \mathbb{C}$ dans le module sur les complexes de [L1].

1.2.4 Le groupe projectif

L'application $x \mapsto \mathrm{Vect}(x)$ de $E \setminus \{0\}$ dans l'ensemble $S_1(E)$ des droites de E est surjective et deux éléments $x, y \in E \setminus \{0\}$ ont même image si, et seulement si, ils sont dans la même orbite $K^*x = K^*y$ pour l'action de K^* sur E (par homothéties). On peut donc identifier l'ensemble quotient $(E \setminus \{0\})/K^*$ à l'ensemble $S_1(E)$.

Définition 3. On appelle *espace projectif* associé au K-espace vectoriel E, et l'on note $\mathbf{P}(E)$, l'ensemble des droites de E ou, de manière équivalente, l'ensemble quotient $(E \setminus \{0\})/K^*$. L'espace projectif $\mathbf{P}(K^n)$ est noté $\mathbf{P}^{n-1}(K)$.

Puisque le noyau de l'action de $\mathcal{GL}(E)$ sur $\mathbf{P}(E) = S_1(E)$ est K^* (exercice 5 de la page 267), le groupe quotient $\mathbf{P}\mathcal{GL}(E) = \mathcal{GL}(E)/K^*$ agit fidèlement transitivement sur $\mathbf{P}(E)$; de même, le groupe quotient $\mathbf{P}GL_n(K) = GL_n(K)/K^*$ agit fidèlement transitivement sur $\mathbf{P}^{n-1}(K)$: c'est pourquoi ils ont été nommés « groupes projectifs ». On considère les éléments du groupe projectif comme des *transformations* (projectives) agissant sur les *points* de l'espace projectif.

On notera $[x_1 : \cdots : x_n]$ la classe dans $\mathbf{P}^{n-1}(K) = (K^n \setminus \{0\})/K^*$ d'un élément $x := (x_1, \ldots, x_n) \in K^n \setminus \{0\}$, ou bien, ce qui revient au même, la droite $\mathrm{Vect}(x) \in S_1(K^n)$. On dit aussi que les x_i sont des *coordonnées homogènes* du point correspondant de l'espace projectif $\mathbf{P}^{n-1}(K)$. Il faut prendre garde que la valeur de x_i en un point $[x_1 : \cdots : x_n]$ de l'espace projectif $\mathbf{P}^{n-1}(K)$ n'est pas définie, puisque tout $[\lambda x_1 : \cdots : \lambda x_n]$ ($\lambda \in K^*$) est encore une écriture de ce même point. Cependant, la condition $x_i = 0$ a un sens bien déterminé : en effet, sa vérification ne dépend pas du choix d'une écriture particulière $[\lambda x_1 : \cdots : \lambda x_n]$. Si l'on note H_i l'hyperplan $x_i = 0$ de K^n, le sous-ensemble $\overline{H_i}$ de $\mathbf{P}^{n-1}(K)$ défini par la condition $x_i = 0$ est l'image de H_i par la projection $K^n \setminus \{0\} \to \mathbf{P}^{n-1}(K)$. Comme H_i est stable sous l'action de K^* sur $K^n \setminus \{0\}$, on a $\overline{H_i} = H_i/K^* = \mathbf{P}(H_i)$. Le complémentaire U_i de $\overline{H_i}$ dans $\mathbf{P}^{n-1}(K)$ est défini par la condition $x_i \neq 0$. De l'égalité $[x_1 : \cdots : x_n] = \left[\frac{x_1}{x_i} : \cdots : \frac{x_n}{x_i}\right]$, on déduit que tout élément de U_i admet une unique représentation $[\xi_1 : \cdots : \xi_n]$ dans laquelle $\xi_i = 1$. Il y a donc une bijection p_i de U_i sur K^{n-1}, définie par $[x_1 : \cdots : x_n] \mapsto \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_n}{x_i}\right)$.

Exercice 6.

Expliciter l'application réciproque de K^{n-1} sur U_i.

Solution. Attention au décalage des indices ! L'image de $(y_1, \ldots, y_{n-1}) \in K^{n-1}$ par p_i^{-1} est $[\xi_1 : \cdots : \xi_n] \in U_i$, où $\xi_k := y_k$ pour $k < i$, $\xi_i := 1$ et $\xi_k := y_{k-1}$ pour $k > i$.

Puisque, pour tout $[x_1 : \cdots : x_n] \in \mathbf{P}^{n-1}(K)$, l'un au moins des x_i est non nul, les « cartes » U_i recouvrent $\mathbf{P}^{n-1}(K)$. Nous les surnommons « cartes »parce qu'elles peuvent servir à repérer des points : un point x de U_i est parfaitement déterminé par les composantes de $p_i(x) \in K^{n-1}$. Naturellement, lorsque l'on change de carte, on doit changer de coordonnées (exercice I.6.16 de la page 309).

Soit $A = (a_{i,j}) \in GL_n(K)$. Notons $[A]$ son image dans $\mathbf{P}GL_n(K)$. L'action de $[A]$ sur $[x_1 : \cdots : x_n]$ est décrite par la formule :

$$[A][x_1 : \cdots : x_n] = \left[\sum_{j=1}^{n} a_{1,j}x_j : \cdots : \sum_{j=1}^{n} a_{n,j}x_j\right].$$

La transformation correspondante de l'espace projectif $\mathbf{P}^{n-1}(K)$ est appelée *homographie de matrice* A. Supposons que la carte U_n contienne à la fois $x := [x_1 : \cdots : x_n]$ et $x' := [A][x_1 : \cdots : x_n]$. Si l'on note $(y_1, \ldots, y_{n-1}) := p_n(x) \in K^{n-1}$ (donc $y_i = x_i/x_n$) et $(y'_1, \ldots, y'_{n-1}) := p_n(x') \in K^{n-1}$ (donc $y'_i = x'_i/x'_n$), on a les relations :

$$\forall i \in [\![1, n-1]\!]\ ,\ y'_i = \frac{x'_i}{x'_n} = \frac{\sum_{j=1}^{n} a_{i,j}x_j}{\sum_{j=1}^{n} a_{n,j}x_j} = \frac{\sum_{j=1}^{n-1} a_{i,j}y_j + a_{i,n}}{\sum_{j=1}^{n-1} a_{n,j}y_j + a_{n,n}}.$$

Cette expression est également appelée une *homographie* : c'est donc une fraction rationnelle en les y_j dont numérateur et dénominateur ont pour degré total 0 ou 1. Le calcul de l'effet de $[A]$ dans les autres cartes est similaire mais compliqué par les décalages d'indices (exercice I.6.16 de la page 309). Nous poursuivrons cette étude dans le cas particulier de la *droite projective* $\mathbf{P}^1(K)$ à la section 3 de la page 296.

1.3 Propriétés topologiques et différentielles

Dans toute cette section, la lettre K désignera le corps $\mathbb{R}$ ou le corps $\mathbb{C}$. Rappelons que, selon la section 4 de la page 440 du module II.1, les espaces vectoriels de dimension finie sur K peuvent être munis de normes et que celles-ci sont toutes équivalentes entre elles, donc définissent les mêmes propriétés topologiques. Cela s'applique donc à $M_n(K)$, et à $\mathcal{L}(E)$ si E est lui-même un K-espace vectoriel de dimension finie. *Il sera commode de faire appel à une norme sur* $M_n(K)$ *telle que* $\|BC\| \leqslant \|B\|\,\|C\|$ (section 4 de la page 440).

Selon le module II.5, on peut parler de fonctions continues, ou de fonctions différentiables de classe $\mathcal{C}^\infty$ sur $M_n(K)$ et de leurs différentielles. Il est d'ailleurs équivalent de travailler sur $M_n(K)$ ou sur $\mathcal{L}(E)$: pour tout choix d'une base de E, l'isomorphisme associé de $\mathcal{L}(E)$ sur $M_n(K)$ est en effet continu et différentiable de classe $\mathcal{C}^\infty$, et il en est de même de l'isomorphisme réciproque. Nous travaillerons donc sur les matrices de $M_n(K)$ et traduirons ensuite certains de nos résultats pour $\mathcal{L}(E)$.

Proposition 13. (i) Le sous-ensemble $GL_n(K)$ de $M_n(K)$ est un ouvert dense. Le sous-ensemble $\mathcal{GL}(E)$ de $\mathcal{L}(E)$ est un ouvert dense.
(ii) Le sous-ensemble $SL_n(K)$ de $M_n(K)$ est un fermé d'intérieur vide. Le sous-ensemble $\mathcal{SL}(E)$ de $\mathcal{L}(E)$ est un fermé d'intérieur vide.

Démonstration. (i) Le sous-ensemble $GL_n(K)$ est l'image réciproque de l'ouvert K^* de K par l'application déterminant, qui est polynomiale donc continue : c'est donc un ouvert.

Pour toute matrice $A \in M_n(K)$, l'ensemble $\mathrm{sp}(A) \subset K$ des valeurs propres de A est fini. Il existe donc $n_0 \in \mathbb{N}^*$ tel que $\forall p \geqslant n_0\ ,\ \dfrac{1}{p} \notin \mathrm{sp}(A)$. Posons $A_p := A - \dfrac{1}{p} I_n$: pour $p \geqslant n_0$, on a donc $A_p \in GL_n(A)$, et, par ailleurs, $\lim\limits_{p\to+\infty} A_p = A$, qui appartient donc à l'adhérence de $GL_n(K)$: ce dernier est bien dense.

(ii) Le sous-ensemble $SL_n(K)$ est l'image réciproque du fermé $\{1\}$ de K par l'application déterminant, qui est polynomiale donc continue : c'est donc un fermé.

Supposons que $SL_n(K)$ contienne un ouvert non vide ; il contiendrait donc en particulier une boule ouverte $\|A - A_0\| < \varepsilon$ de centre $A_0 \in SL_n(A)$. Pour tout $\lambda \in K$ tel que $|\lambda - 1| < \dfrac{\varepsilon}{\|A_0\|}$, on aurait donc $\lambda A_0 \in SL_n(A)$, donc $\lambda^n = \dfrac{1}{\det A_0}$.

Mais il ne peut y avoir qu'un nombre fini de tels λ, d'où une contradiction. ∎

Exercice 7.

Démontrer que, pour deux matrices quelconques $A, B \in M_n(K)$, on a l'égalité $\chi_{AB} = \chi_{BA}$ des polynômes caractéristiques. On commencera par vérifier que, si B est inversible, AB et BA sont semblables. Puis, notant $c_k(M)$ le coefficient de T^k dans $\chi_M(T)$, on étudiera la fonction $A \mapsto c_k(AB) - c_k(BA)$ de $M_n(K)$ dans K (continuité, valeur sur $GL_n(K)$).

Solution. Si B est inversible, on a $BA = B(AB)B^{-1}$ et les matrices AB et BA étant semblables, elles ont bien même polynôme caractéristique. Dans le cas général, pour $k \in [\![0, n]\!]$, la fonction $A \mapsto c_k(AB) - c_k(BA)$ est continue car polynomiale. Elle est nulle sur $GL_n(K)$ d'après l'argument ci-dessus. Puisque $GL_n(K)$ est dense dans $M_n(K)$, cette fonction est identiquement nulle. Comme c'est vrai pour tout k, on a bien $\chi_{AB} = \chi_{BA}$.

Proposition 14. L'ensemble des matrices diagonalisables de $M_n(\mathbb{C})$ est dense.

Démonstration. Soit $A \in M_n(\mathbb{C})$. Il existe $P \in GL_n(K)$ telle que $T := P^{-1}AP$ est triangulaire (corollaire 30 de la page 239), de coefficients diagonaux $\lambda_1, \dots, \lambda_n$. Il existe des réels $a_1, \dots, a_n$ tels que, pour tout $p \in \mathbb{N}^*$, les $\lambda_i + \dfrac{a_i}{p}$ sont deux à deux distincts (exercice I.6.18 de la page 309).

La matrice $A_p := A + \dfrac{1}{p} P \operatorname{Diag}(a_1, \dots, a_n) P^{-1}$ a pour valeurs propres les $\lambda_i + \dfrac{a_i}{p}$, donc elle est diagonalisable ; et l'on a évidemment $\lim\limits_{p \to +\infty} A_p = A$. ∎

Exercice 8.

En déduire une nouvelle preuve du théorème de Cayley-Hamilton. (Commencer, de manière algébrique, par le cas d'une matrice diagonalisable ; puis appliquer la proposition.)

Solution. Si $P^{-1}AP = D := \operatorname{Diag}(\lambda_1, \dots, \lambda_n)$, on a $\chi_A = \chi_D$, donc :

$$\begin{aligned}\chi_A(A) = \chi_D(A) = \chi_D(PDP^{-1}) &= P\chi_D(D)P^{-1} \\ &= P \operatorname{Diag}\big(\chi_D(\lambda_1), \dots, \chi_D(\lambda_n)\big) P^{-1} = 0.\end{aligned}$$

Dans le cas général, notons, pour $(i, j) \in [\![1, n]\!]^2$, $d_{i,j}(A)$ le coefficient d'indices (i, j) de la matrice $\chi_A(A)$. Les fonctions $d_{i,j}$ de $M_n(\mathbb{C})$ sont continues (car polynomiales) et nulles sur le sous-ensemble dense des matrices diagonalisables, donc identiquement nulles. Cette preuve ne concerne que $M_n(\mathbb{C})$; pour une preuve reposant sur le même principe mais valable sur un corps arbitraire, voir l'exercice I.7.41 de la page 350.

Proposition 15. Les groupes $SL_n(\mathbb{R})$, $SL_n(\mathbb{C})$ et $GL_n(\mathbb{C})$ sont connexes par arcs. Le groupe $GL_n(\mathbb{R})$ a deux composantes connexes : $GL_n^+(\mathbb{R})$ (matrices à déterminant strictement positif) et $GL_n^-(\mathbb{R})$ (matrices à déterminant strictement négatif).

Démonstration. Soit $A \in SL_n(K)$. D'après le théorème 4 de la page 263, on peut écrire : $A := T_{i_1,j_1}(\alpha_1) \cdots T_{i_k,j_k}(\alpha_k)$. Posons $A(t) := T_{i_1,j_1}(t\alpha_1) \cdots T_{i_k,j_k}(t\alpha_k)$.

Alors $t \mapsto A(t)$ est un chemin continu dans $SL_n(K)$ tel que $A(0) = I_n$ et $A(1) = A$. Toutes les matrices de $SL_n(K)$ sont donc dans la même composante connexe par arcs que I_n.
Soit $B \in GL_n(K)$. On peut l'écrire $B = AD_n(\lambda)$, où $A \in SL_n(K)$. On reprend le chemin $A(t)$ ci-dessus. Si $K := \mathbb{C}$, il existe un chemin $\lambda(t)$ dans $\mathbb{C}^*$ tel que $\lambda(0) = 1$ et $\lambda(1) = \lambda$. Alors le chemin $B(t) := A(t)D_n\big(\lambda(t)\big)$ relie $B(0) = I_n$ à $B(1) := B$.
Si $K := \mathbb{R}$, il existe un chemin $\lambda(t)$ dans $\mathbb{R}^*$ tel que $\lambda(0) = +1$ (si $\lambda > 0$) ou -1 (si $\lambda < 0$) et $\lambda(1) = \lambda$. Alors le chemin $B(t) := A(t)D_n\big(\lambda(t)\big)$ relie, selon le cas, $B(0) = I_n$ ou $D_n(-1)$ à $B(1) := B$. Ainsi, $GL_n^+(\mathbb{R})$ et $GL_n^-(\mathbb{R})$ sont connexes par arcs. Si l'on pouvait relier I_n à $D_n(-1)$ par un chemin continu $B(t)$ dans $GL_n(\mathbb{R})$, la fonction continue $\det B(t)$ sur $[0,1]$ passerait de la valeur $+1$ à la valeur -1 sans s'annuler ! ∎

Exercice 9.

Les groupes $GL_n(K)$ et $SL_n(K)$ sont-ils compacts ?
On rappelle qu'en dimension finie, un compact est une partie fermée et bornée.
Solution. Le groupe $GL_n(K)$ est dense dans $M_n(K)$ mais n'est pas égal à $M_n(K)$ (il ne contient pas 0) : il n'est donc pas fermé, donc pas compact.
Le groupe $SL_1(K) = \{1\}$ est évidemment compact. Supposons donc $n \geqslant 2$.
Alors $SL_n(K)$ contient en particulier toutes les matrices $\mathrm{Diag}(\lambda, \lambda^{-1}, 1, \dots, 1)$ pour $\lambda \in K^*$: il n'est donc pas borné, donc pas compact.

Rappelons (module sur les fonctions à plusieurs variables de [L1]) qu'une application f définie sur un ouvert U d'un K-espace vectoriel V de dimension finie et à valeurs dans un K-espace vectoriel W de dimension finie est dite *différentiable* en $a \in U$ s'il existe une application linéaire $L : V \to W$ telle que, lorsque $h \to 0$: $f(a+h) = f(a)+L(h)+\mathrm{o}(h)$, où le terme dénoté $\mathrm{o}(h)$ est une fonction $\varphi(h)$ telle que $\lim\limits_{\substack{h\to 0\\ h\neq 0}} \dfrac{\|\varphi(h)\|}{\|h\|} = 0$. Avec les notations de *loc. cit.*, $\mathrm{o}(h) = \|h\|\,\varepsilon(h)$. L'application linéaire L est alors unique ; on la note D_af et on l'appelle *différentielle de f en a*. Lorsque l'application f est différentiable en tout point $a \in U$, on dit qu'elle est différentiable sur U ; si de plus l'application différentielle $a \mapsto D_af$ de U dans $\mathcal{L}(V,W)$, est continue sur U, on dit que f est de classe $\mathcal{C}^1$ (module II.5). Si l'on peut itérer ce processus, on dit que f est de classe $\mathcal{C}^\infty$. Par exemple, les applications polynomiales sur V sont de classe $\mathcal{C}^\infty$, ainsi que les applications rationnelles sur un ouvert U de V.

Exercice 10.

Étudier la différentiabilité et la différentielle de l'application $f : A \mapsto A^2$ de $M_n(K)$ dans $M_n(K)$. (Dans $f(A+H)-f(A)$, isoler le terme linéaire en H et le terme résiduel.)
Solution. On a $f(A+H)-f(A) = (A+H)^2 - A^2 = AH + HA + H^2$. Le terme linéaire en H est $L(H) := AH + HA$. Il reste à voir que $\varphi(H) := H^2$ est bien un $\mathrm{o}(H)$. Mais on a $\|H^2\| \leqslant \|H\|^2$, d'où $\dfrac{\|\varphi(H)\|}{\|H\|} \leqslant \|H\| \underset{H\to 0}{\longrightarrow} 0$, ce qui le confirme. Finalement, D_Af est l'application linéaire $H \mapsto AH + HA$ de $M_n(K)$ dans $M_n(K)$.

Proposition 16. (i) L'application $\det$ de $M_n(K)$ dans K^* est différentiable de classe $\mathcal{C}^\infty$. Sa différentielle en $A \in M_n(K)$ est donnée par la formule :

$$D_A \det(H) = \mathrm{Tr}\left({}^t\tilde{A}H\right),$$

où $\tilde{A}$ désigne la comatrice de A (module sur la dimension finie de [L1]).
(ii) L'application $I : A \mapsto A^{-1}$ de $GL_n(K)$ dans $M_n(K)$ est différentiable de classe $\mathcal{C}^\infty$. Sa différentielle en $A \in GL_n(K)$ est donnée par la formule :

$$D_A I(H) = -A^{-1}HA^{-1}.$$

Démonstration. (i) L'application $\det$ étant polynomiale, elle est bien différentiable de classe $\mathcal{C}^\infty$. Supposons d'abord que $A := I_n$. On utilise alors le polynôme caractéristique, évalué en -1, pour obtenir le développement limité : $\det(I_n + H) = 1 + \mathrm{Tr}(H) + \mathrm{o}(H)$. En effet, les termes manquant sont tous de degré supérieur ou égal à 2 en les coefficients de H. Pour une matrice inversible A quelconque, on en déduit :

$$\begin{aligned}\det(A+H) = \det A \det(I_n + A^{-1}H) &= \det A\big(1 + \mathrm{Tr}(A^{-1}H) + \mathrm{o}(H)\big)\\ &= \det A + \mathrm{Tr}\left({}^t\tilde{A}H\right) + \mathrm{o}(H),\end{aligned}$$

puisque $\det A \,\mathrm{Tr}(A^{-1}H) = \mathrm{Tr}\big((\det A)A^{-1}H\big)$ et que $(\det A)A^{-1} = {}^t\tilde{A}$. Ainsi, lorsque $A \in GL_n(K)$, on a bien $D_A \det(H) = \mathrm{Tr}\left({}^t\tilde{A}H\right)$. Mais l'application $A \mapsto D_A \det$ est continue et $GL_n(K)$ est dense dans $M_n(K)$: l'égalité est donc vraie pour tout A.
(ii) L'application I étant rationnelle sur l'ouvert $GL_n(K)$ (par exemple en vertu de la formule $A^{-1} = \dfrac{1}{\det A}{}^t\tilde{A}$ du module sur la dimension finie de [L1]), elle est bien différentiable de classe $\mathcal{C}^\infty$. On calcule alors :

$$(A+H)^{-1} - A^{-1} + A^{-1}HA^{-1} = (A+H)^{-1}(HA^{-1})^2.$$

Tout cela est bien défini lorsque H est proche de 0, puisque $GL_n(K)$ est ouvert. Pour vérifier l'égalité, il suffit de multiplier ses deux membres par $A + H$ à gauche. Posant $L(H) := -A^{-1}HA^{-1}$, il reste à voir que $\varphi(H) := (A+H)^{-1}(HA^{-1})^2$ est un $\mathrm{o}(H)$. On a : $\|\varphi(H)\| \leqslant \left\|(A+H)^{-1}\right\| \left\|(A^{-1})^2\right\| \|H\|^2$. Comme $\left\|(A+H)^{-1}\right\|$ tend vers $\left\|A^{-1}\right\|$ lorsque H tend vers 0, on a bien la conclusion voulue. ∎

Rappelons que, si $\gamma : [0,1] \to E$ est un chemin (autrement dit, une courbe paramétrée) de classe $\mathcal{C}^1$ d'origine $x := \gamma(0)$ dans le K-espace vectoriel de dimension finie E, le *vecteur tangent à γ en a* est simplement le vecteur dérivé à droite $\gamma'(0)$. Soit X une partie arbitraire de E. On appellera *vecteur tangent à X en $x \in X$* tout vecteur $\gamma'(0)$, où $\gamma : [0,1] \to E$ est un chemin de classe $\mathcal{C}^1$ tel que $\gamma(0) = x$ et que $\forall t \in [0,1]\ ,\ \gamma(t) \in X$.

Proposition et définition 17. L'ensemble de tous les vecteurs tangents à $SL_n(K)$ en A est le sous–espace vectoriel de $M_n(K)$ formé des matrices H telles que $\mathrm{Tr}(A^{-1}H) = 0$. On l'appelle *espace tangent à $SL_n(K)$ en A*. L'espace tangent à $SL_n(K)$ en I_n est noté $\mathrm{sl}_n(K)$. On a donc : $\mathrm{sl}_n(K) = \{H \in M_n(K) \mid \mathrm{Tr}\, H = 0\}$.

Démonstration. Soit H un vecteur tangent à $SL_n(K)$ en A. il existe donc un chemin $\gamma : [0,1] \to M_n(K)$, de classe $\mathcal{C}^1$, tel que $\gamma(0) = A$ et $\gamma'(0) = H$ et tel que, pour tout $t \in [0,1]$, on ait $\gamma(t) \in SL_n(K)$. La composée des fonctions $\gamma : [0,1] \to M_n(K)$ et $\det : M_n(K) \to K$ est de classe $\mathcal{C}^1$ et sa dérivée se calcule selon la « règle des chaînes » pour tout $t \in [0,1]$, $(\det\circ\gamma)'(t) = D_{\gamma(t)} \det\big(\gamma'(t)\big)$. En particulier, pour $t := 0$, on trouve :

$$(\det\circ\gamma)'(0) = D_A \det(H) = \mathrm{Tr}\left({}^t\tilde{A}H\right) = \det A\ \mathrm{Tr}(A^{-1}H),$$

d'après la proposition 16. Mais la fonction $\det\circ\gamma$ est constante de valeur 1 sur $[0,1]$, puisque $\gamma(t) \in SL_n(K)$ pour tout t. Sa dérivée est donc nulle, et l'on a bien $\mathrm{Tr}(A^{-1}H) = 0$.

Soit réciproquement $H \in M_n(K)$ tel que $\mathrm{Tr}(A^{-1}H) = 0$. Anticipant un peu sur le rappel des propriétés de l'exponentielle, posons $\gamma(t) := A\exp(tA^{-1}H)$. C'est un chemin de classe $\mathcal{C}^1$ et $\gamma(0) = A\exp(0) = AI_n = A$. De plus, on a, $\mathrm{Tr}(tA^{-1}H) = 0$ pour tout t, donc $\det\gamma(t) = \det A \exp \mathrm{Tr}(tA^{-1}H) = 1$: c'est bien un chemin dans $SL_n(K)$. Enfin, par la règle des chaînes : $\gamma'(0) = AD_0\exp(A^{-1}H) = AA^{-1}H = H$, ce qui implique que H est dans l'espace tangent. ∎

Les propriétés suivantes de l'exponentielle seront établies dans les modules II.1 (page 518), II.2 et II.5. La série :

$$\exp A := \sum_{k\geqslant 0} \frac{1}{k!}A^k = \lim_{m\to+\infty} E_m(A),$$

où $E_m(X) := \displaystyle\sum_{k=0}^{m} \frac{1}{k!}X^k \in \mathbb{Q}[X]$, converge pour tout $A \in M_n(K)$. Elle définit une fonction continue et différentiable de classe $\mathcal{C}^\infty$ et à valeurs dans $GL_n(K)$ (proposition 58 de la page 534). On peut compléter ces propriétés comme suit :

1. Sa différentielle en $0 \in M_n(K)$ est donnée par la formule : $D_0\exp(H) = H$ (le calcul de la différentielle en d'autres points est beaucoup plus compliqué).
2. Si A et B commutent, on a $\exp(A+B) = \exp(A)\exp(B)$.
3. Soient $A \in M_n(K)$ et $P \in GL_n(K)$. Alors $\exp\left(PAP^{-1}\right) = P\,(\exp A)\,P^{-1}$.
4. Si A est triangulaire supérieure (respectivement diagonale) de coefficients diagonaux $\lambda_1, \ldots, \lambda_n$, alors $\exp A$ est triangulaire supérieure (respectivement diagonale) de coefficients diagonaux $e^{\lambda_1}, \ldots, e^{\lambda_n}$. En conséquence, si A est trigonalisable (respectivement diagonalisable), $\exp A$ l'est également ; si $K = \mathbb{C}$ et si l'on nomme $\lambda_1, \ldots, \lambda_n$ les valeurs propres de A (éventuellement comptées plusieurs fois), alors $e^{\lambda_1}, \ldots, e^{\lambda_n}$ sont les valeurs propres de $\exp A$; enfin, sans condition, on a $\det(\exp A) = e^{\mathrm{Tr}\, A}$.

Exercice 11.

Soient $A := \begin{pmatrix} 0 & -\pi \\ 0 & 0 \end{pmatrix}$, $B := \begin{pmatrix} 0 & 0 \\ \pi & 0 \end{pmatrix}$ et $C := A + B$. Calculer $\exp A$, $\exp B$ et leurs produits. Montrer que $\exp C$ est diagonalisable et calculer ses valeurs propres. En déduire $\exp C$ sans calcul. Que remarque-t-on ?

Solution. Puisque $A^2 = B^2 = 0$, on a :

$$\exp A = I_2 + A = \begin{pmatrix} 1 & -\pi \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad \exp B = I_2 + B = \begin{pmatrix} 1 & 0 \\ \pi & 1 \end{pmatrix},$$

d'où :

$$\exp A \exp B = \begin{pmatrix} 1-\pi^2 & -\pi \\ \pi & 1 \end{pmatrix} \quad \text{et} \quad \exp B \exp A = \begin{pmatrix} 1 & -\pi \\ \pi & 1-\pi^2 \end{pmatrix}.$$

Les valeurs propres de A sont $\pm \mathrm{i}\pi$, elle est donc diagonalisable. Les valeurs propres de $\exp A$ sont donc $e^{\pm \mathrm{i}\pi} = -1$ (compté deux fois) et elle est diagonalisable, donc égale à $-I_2$. On a donc $\exp(A+B) \neq \exp A \exp B$ et $\exp(A+B) \neq \exp B \exp A$.

Le fait suivant mérite d'être cité à part. Pour toute matrice $A \in M_n(K)$, le sous–espace vectoriel de $M_n(K)$ engendré par les puissances de A est égal à l'espace vectoriel $K_{n-1}[A] = \mathrm{Vect}(A^0, \ldots, A^{n-1})$: cela découle du théorème de Cayley-Hamilton. Prenons $K := \mathbb{R}$ ou $\mathbb{C}$. Chacune des matrices $E_m(A)$ appartient à $K_{n-1}[A]$, qui est *fermé* dans $M_n(K)$. À la limite, on a donc $\exp A \in K_{n-1}[A]$, c'est-à-dire qu'il existe $F \in K_{n-1}[X]$ tel que $\exp A = F(A)$. On dira que « $\exp A$ est un polynôme en A ». Naturellement, *F dépend de A : la fonction* exp *n'est pas polynomiale en A* (le cas des matrices nilpotentes est à part, *cf.* la preuve de la proposition 19 de la page suivante).

Logarithme d'une matrice inversible complexe

Notre but est maintenant de définir, pour toute matrice $B \in GL_n(\mathbb{C})$, un « logarithme » $A := \log B$, c'est-à-dire une matrice $A \in M_n(\mathbb{C})$ telle que $\exp A = B$. Pour cela, nous ferons appel à des propriétés de l'exponentielle complexe rencontrées dans le module sur les complexes de [L1] : l'application exponentielle $\exp : \mathbb{C} \to \mathbb{C}^*$ est surjective et sa restriction à l'ensemble $E := \{\lambda \in \mathbb{C} \mid \mathrm{Im}\,\lambda \in]-\pi, \pi]\}$ est bijective. *Nous appellerons logarithme complexe la restriction de cette application* : ainsi, si $z \in \mathbb{C}^*$, $\log z$ est l'unique $w \in E$ tel que $e^w = z$.

Proposition 18. Notons D l'ensemble des matrices diagonalisables de $M_n(K)$ dont toutes les valeurs propres appartiennent à E et D^* l'ensemble des matrices diagonalisables de $GL_n(K)$. Alors l'application exp induit une bijection de D sur D^*. L'antécédent $\log B$ d'une matrice $B \in D^*$ se calcule ainsi : on écrit $B = P\,\mathrm{Diag}(\mu_1, \ldots, \mu_n)P^{-1}$, où les $\mu_i \in \mathbb{C}^*$; on pose alors $\log B = P\,\mathrm{Diag}(\log \mu_1, \ldots, \log \mu_n)P^{-1}$. Enfin, $\log B \in \mathbb{C}_{n-1}[B]$.

Démonstration. Si deux matrices diagonalisables A et B vérifient $\exp A = B$, alors elles commutent (car $\exp A \in \mathbb{C}[A]$), donc sont simultanément diagonalisables. On les écrit

$A := P\,\mathrm{Diag}(\lambda_1,\dots,\lambda_n)P^{-1}$, où les $\lambda_i \in E$ et $B = P\,\mathrm{Diag}(\mu_1,\dots,\mu_n)P^{-1}$, où les $\mu_i \in \mathbb{C}^*$. Puisque $\exp A = P\,\mathrm{Diag}(e^{\lambda_1},\dots,e^{\lambda_n})P^{-1}$, il y a une unique solution avec $A \in D$, qui est de prendre $\lambda_i = \log \mu_i$ et A comme dans l'énoncé.
Supposons maintenant A et B ainsi calculées et soit S l'ensemble des valeurs propres de B. Cet ensemble a au plus n éléments, et il existe donc un polynôme $F \in \mathbb{C}_{n-1}[X]$ tel que $\forall \mu \in S\ ,\ F(\mu) = \log \mu$. Alors :

$$F(B) = P\,\mathrm{Diag}\big(F(\mu_1),\dots,F(\mu_n)\big)P^{-1} = P\,\mathrm{Diag}(\log\mu_1,\dots,\log\mu_n)P^{-1} = A. \qquad \blacksquare$$

Comme dans le cas de l'exponentielle (page 277), le polynôme d'interpolation F dépend de B. Par exemple, en dimension $n := 1$, si $B := (\mu)$, F est constant de valeur $\log\mu$. Naturellement, $\log B = F(B) = (\log\mu)$. En dimension quelconque, μI_n (où $\mu \neq 0$) est diagonalisable et a pour unique valeur propre μ. Le polynôme F est le polynôme constant $\log\mu$ et $\log \mu I_n = (\log\mu) I_n$.
Les résultats suivants proviennent du module « Approximation » de [L1] : la fonction exponentielle réelle admet au voisinage de 0 un développement limité à tout ordre $m \in \mathbb{N}$: $e^x = \sum_{k=0}^{m} \frac{1}{k!}x^k + \mathrm{o}(x^m)$; sa réciproque, la fonction logarithme réel, admet au voisinage de 1 un développement limité à tout ordre $m \in \mathbb{N}$: $\ln(1+x) = L_m(x) + \mathrm{o}(x^m)$, où l'on a posé : $L_m := \sum_{k=1}^{m} \frac{(-1)^{k+1}}{k} X^k \in \mathbb{Q}[X]$ (il y a un lien entre ces deux approches du logarithme : il est établi, dans les modules II.2 et II.6, que la série $\sum_{k\geqslant 1} \frac{(-1)^{k+1}}{k} z^k$ converge pour $|z| < 1$ et a pour valeur le logarithme complexe $\log z$). D'après les règles de calcul sur les développements limités, la partie polynomiale du développement limité au voisinage de 0 à l'ordre m de $x = \ln e^x$ est la troncature au degré m (*i.e.* on élimine les termes à partir de X^{m+1}) du polynôme $L_m \circ (E_m - 1)$ (le polynôme E_m a été défini page 276) ; par unicité des développements limités, cette troncature est X, d'où la congruence : $L_m \circ (E_m - 1) \equiv X \pmod{X^{m+1}}$. De même, la troncature à l'ordre m de $E_m \circ L_m$ est le développement limité en 1 de la fonction $1 + x = e^{\ln(1+x)}$; par unicité des développements limités, cette troncature est $1+X$. On en déduit la congruence : $E_m \circ L_m \equiv 1 + X \pmod{X^{m+1}}$.
Ainsi, pour $m := 2$, on a $E_2 = 1 + X + \frac{1}{2}X^2$ et $L_2 = X - \frac{1}{2}X^2$, d'où :

$$E_2 \circ L_2 = 1 + L_2 + \frac{1}{2}L_2^2 = 1 + X - \frac{1}{2}X^2 + \frac{1}{2}X^2 - \frac{1}{2}X^3 + \frac{1}{8}X^4 \equiv 1 + X \pmod{X^3}$$

$$L_2 \circ (E_2 - 1) = E_2 - 1 - \frac{1}{2}(E_2 - 1)^2 = X + \frac{1}{2}X^2 - \frac{1}{2}X^2 - \frac{1}{2}X^3 + \frac{1}{8}X^4 \equiv X \pmod{X^3}.$$

Proposition 19. (i) Soit $N \in M_n(K)$ une matrice nilpotente. Alors $\exp N = E_{n-1}(N)$ est unipotente. Pour toute matrice unipotente $U \in M_n(K)$, posons $\log U := L_{n-1}(U - I_n)$. Alors $\log U$ est nilpotente.
(ii) Les applications $\exp$ et $\log$ sont des bijections réciproques l'une de l'autre entre l'ensemble des matrices nilpotentes et l'ensemble des matrices unipotentes de $M_n(K)$.

Démonstration. (i) Pour toute matrice nilpotente N, on a $N^n = 0$ (conséquences de la proposition 38 de la page 247) ; il ne reste donc que les n premiers termes $E_{n-1}(N)$ dans la série $\exp N$. De plus, $E_{n-1}(N) - I_n$ est une combinaison linéaire de $N,\dots,N^{n-1}$, donc nilpotente.

De même, $U-I_n$ est nilpotente et $\log U$ est une combinaison linéaire de $U-I_n,\dots,(U-I_n)^{n-1}$, donc nilpotente.

(ii) Les termes de $\log\exp N = L_{n-1}\big(E_{n-1}(N)-N\big)$ autres que N sont des multiples de N^n d'après les congruences ci-dessus, donc nuls. On a donc bien $\log\exp N = N$. Les termes de $\exp\log U = E_{n-1}\circ L_{n-1}(U-I_n)$ autres que U sont des multiples de $(U-I_n)^n$ d'après les congruences ci-dessus, donc nuls. On a donc bien $\exp\log U = U$.

Pour calculer $\exp(A)$ dans le cas général, on utilise la *décomposition de Dunford* (théorème 44 de la page 250) : *la matrice A admet une unique écriture de la forme $A=D+N$, où D est diagonalisable, N est nilpotente, et D et N commutent ; de plus, D et N sont des polynômes en A.* Pratiquement, si les valeurs propres de A sont $\lambda_1,\dots,\lambda_k$ avec les multiplicités $r_1,\dots,r_k\in\mathbb{N}^*$, on peut écrire :

$$A = P\ \mathrm{Diag}(\lambda_1 I_{r_1}+T_1,\dots,\lambda_k I_{r_k}+T_k)\,P^{-1}, \tag{1}$$

où $T_1,\dots,T_k$ sont triangulaires supérieures strictes. On a alors :

$$D = P\ \mathrm{Diag}(\lambda_1 I_{r_1},\dots,\lambda_k I_{r_k})\,P^{-1} \qquad \text{et} \qquad N = P\ \mathrm{Diag}(T_1,\dots,T_k)\,P^{-1},$$

$$\text{puis}\quad \exp(D) = P\ \mathrm{Diag}(e^{\lambda_1}I_{r_1},\dots,e^{\lambda_k}I_{r_k})\,P^{-1}\quad\text{(qui est diagonalisable inversible),}$$

$$\exp(N) = P\ \mathrm{Diag}(\exp T_1,\dots,\exp T_k)\,P^{-1}\quad\text{(qui est unipotente)}$$

$$\text{et}\quad \exp(A)=\exp(D)\exp(N)=\exp(N)\exp(D).$$

C'est une *décomposition de Dunford multiplicative* de $\exp(A)$. En fait, *toute matrice $A\in GL_n(\mathbb{C})$ admet une unique décomposition de Dunford multiplicative $A=DU=UD$* (D diagonalisable, U unipotente) : partant de la décomposition de Dunford $A=D+N$, on prend $U:=I_n+D^{-1}N$. De plus, D et U sont des polynômes en A (pour ce dernier point, voir l'exercice I.6.20 de la page 309).

Exercice 12.

En déduire toutes les matrices A telles que $\exp(A)=I_n$.

Solution. Avec les notations ci-dessus, $\exp(D)=I_n$ et $\exp(N)=I_n$. On a donc $N=0$, et A est une matrice diagonalisable telle que $\mathrm{sp}(A)\subset 2\mathrm{i}\pi\mathbb{Z}$.

Théorème et définition 20. L'application $\exp : M_n(\mathbb{C})\to GL_n(\mathbb{C})$ est surjective. Notons $\mathcal{E}_n$ (*étoile canonique*) l'ensemble des matrices de $M_n(\mathbb{C})$ dont toutes les valeurs propres sont dans E (ensemble défini page 277). La restriction $\exp : \mathcal{E}_n\to GL_n(\mathbb{C})$ est bijective. La réciproque de cette bijection est l'application *logarithme matriciel*, notée $\log$. Pour tout $B\in GL_n(\mathbb{C})$, on a $\log B\in\mathbb{C}_{n-1}[B]$.

Démonstration. Écrivons $B=DU$ la décomposition de Dunford multiplicative de B. D'après la propriété énoncée juste avant l'exercice, la seule matrice A qui puisse répondre aux conditions $\exp A=B$ et $A\in\mathcal{E}_n$ est donc $A:=\log D+\log U$. Or les matrices $\log D$ et $\log U$ commutent car ce sont respectivement des polynômes en D et U qui commutent (c'est là qu'intervient la propriété finale de chacune des deux propositions qui précèdent). On a donc bien :

$$\exp A=\exp(\log D+\log U)=\exp\log D\,\exp\log U=DU=B.$$

Enfin, $\log D$ est un polynôme en D qui est un polynôme en B, et il en est de même pour $\log U$ donc pour A. On a même $A\in\mathbb{C}_{n-1}[B]$ puisque $\mathbb{C}[B]=\mathbb{C}_{n-1}[B]$ (pour tout polynôme F, $F(B)=R(B)$ où $R=P\mod\chi_B$).

Encore une fois, il faut prendre garde que le logarithme de B est un polynôme en B, mais ce polynôme n'est pas le même pour toute matrice B : *la fonction logarithme n'est pas polynomiale*.

Pratiquement, partant d'une écriture : $B = P\ \mathrm{Diag}(\mu_1 I_{r_1} + T_1, \ldots, \mu_k I_{r_k} + T_k)\, P^{-1}$, on aura :

$$\log B = P\ \mathrm{Diag}(\log \mu_1 I_{r_1}, \ldots, \log \mu_k I_{r_k})\, P^{-1} + P\ \mathrm{Diag}(N_1, \ldots, N_k)\, P^{-1}, \qquad (2)$$

où, pour $1 \leqslant i \leqslant k$, $N_i := \log(I_{r_i} + \mu_i^{-1} T_i) = L_{r_i - 1}(\mu_i^{-1} T_i)$.

Exercice 13.

Démontrer que $-I_2$ est l'exponentielle d'une matrice réelle, mais pas $B := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

Solution. Pour la première affirmation, voir l'exercice 11 de la page 277. Pour la deuxième affirmation : on aurait $\det B = \exp(\mathrm{Tr}\, A) > 0$, ce qui n'est pas le cas.

Puisque $\det(\exp A) = \exp(\mathrm{Tr}\, A)$, on voit que l'exponentielle envoie $\mathrm{sl}_n(\mathbb{C})$ (proposition 17 de la page 276) dans $SL_n(\mathbb{C})$. *Mais cette restriction-corestriction n'est pas surjective* :

Exercice 14.

Démontrer que $\begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} \in SL_n(\mathbb{C})$ n'est pas l'exponentielle d'une matrice $A \in \mathrm{sl}_n(\mathbb{C})$.

Solution. La matrice A aurait des valeurs propres de la forme $(2k+1)\mathrm{i}\pi$ et de somme nulle, donc distinctes ; elles serait donc diagonalisable, et $\exp A$ aussi, ce qui n'est pas.

2 Groupe orthogonal, groupe unitaire

L'étude de ces groupes repose principalement sur les modules I.3 et I.4 où ils ont été introduits. On se cantonnera, pour l'essentiel, au cas des formes (quadratiques ou hermitiennes) *définies positives*, donc, sur $\mathbb{R}$, au cas du groupe orthogonal $\mathcal{O}(E)$ et du groupe spécial orthogonal $\mathcal{SO}(E)$ (où E désigne un espace vectoriel euclidien ; voir la définition 33 de la page 167) et, sur $\mathbb{C}$, au cas du groupe unitaire $\mathcal{U}(E)$ et du groupe spécial unitaire $\mathcal{SU}(E)$ (où E désigne un espace vectoriel hermitien ; voir la définition 99 de la page 200).
Pratiquement, nous étudierons le plus souvent les groupes de matrices $O_n(\mathbb{R})$, $SO_n(\mathbb{R})$, $U_n(\mathbb{C})$ et $SU_n(\mathbb{C})$ (définitions 35 de la page 168 et 101 de la page 201), puis nous transposerons (ou laisserons au lecteur le soin de transposer ...) nos résultats sous forme plus géométrique. Le théorème de réduction des endomorphismes normaux (théorème 113 de la page 204) rend d'ailleurs les propriétés du groupe unitaire essentiellement plus simples que celles du groupe orthogonal : elles seront donc bien souvent traitées dans les exercices à la fin de ce module. Par ailleurs, le cas de la dimension $n := 1$ étant trivial, nous supposerons toujours $n \geqslant 2$.

Nous dirons qu'une matrice A est *orthogonalement semblable* (resp. *unitairement semblable*) à une matrice B si elle lui est semblable via une matrice de passage orthogonale (resp. unitaire), c'est-à-dire si $B = PAP^{-1}$ avec $P \in O_n(\mathbb{R})$ (resp. $P \in U_n(\mathbb{C})$).

2.1 Propriétés algébriques

2.1.1 Formes normales

Notre premier résultat est à rapprocher de l'exercice 17 de la page 205 sur les matrices antisymétriques réelles et repose sur la réduction des endomorphismes normaux. Pour le formuler, nous introduirons la notation $R(\theta) := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

Théorème 21. (i) Toute matrice orthogonale $A \in O_n(\mathbb{R})$ est orthogonalement semblable à une unique matrice diagonale par blocs de la forme $\mathrm{Diag}(I_r, -I_s, B_1, \ldots, B_p)$, les matrices $B_k \in M_2(\mathbb{R})$ étant de la forme $B_k := R(\theta_k)$, où $0 < \theta_1 \leqslant \cdots \leqslant \theta_p < \pi$. Dans le cas où $r = 0$, resp. $s = 0$, resp. $p = 0$, on convient que la notation de l'énoncé représente $\mathrm{Diag}(-I_s, B_1, \ldots, B_p)$, resp. $\mathrm{Diag}(I_r, B_1, \ldots, B_p)$, resp. $\mathrm{Diag}(I_r, -I_s)$.
(ii) Deux matrices orthogonales sont semblables si, et seulement si, elles ont mêmes invariants $(r, s, \theta_1, \ldots, \theta_p)$.
(iii) Si $A \in SO_n(\mathbb{R})$, l'entier s est pair.

Démonstration. (i) La démarche est similaire à celle de l'exercice 17 de la page 205. Considérée comme matrice complexe, A est unitaire ; si $\lambda \in \mathbf{U}$ en est valeur propre non réelle, $\overline{\lambda} = \lambda^{-1}$ en est aussi valeur propre, de même multiplicité (car χ_A est à coefficients réels). On écrit une décomposition orthogonale :

$$\mathbb{C}^n = \mathrm{Ker}(A - I_n) \oplus \mathrm{Ker}(A + I_n) \oplus (E_1 \oplus F_1) \oplus \cdots \oplus (E_p \oplus F_p),$$

où $E_k = \mathrm{Ker}(A - \lambda_k I_n)$ et $F_k = \mathrm{Ker}(A - \overline{\lambda_k} I_n)$, $\lambda_k := \cos\theta_k + \mathrm{i}\sin\theta_k$, $\theta_k \in]0, \pi[$. Fixons temporairement $k \in [\![1, p]\!]$. Soit $V_1, \ldots, V_h$ une base orthonormée de E_k ; alors, posant (pour $1 \leqslant j \leqslant h$) $W_j = \dfrac{1}{\sqrt{2}}(V_j + \overline{V_j})$ et $W'_j = \dfrac{1}{\mathrm{i}\sqrt{2}}(V_j - \overline{V_j})$, on voit que $(W_1, W'_1, \ldots, W_h, W'_h)$ est une base orthonormée de $E_k \oplus F_k$ formée de vecteurs de $\mathbb{R}^n$ et telle que $A(W_j) = \cos\theta_j W_j - \sin\theta_j W'_j$ et $A(W'_j) = \sin\theta_j W_j + \cos\theta_j W'_j$.
On réunit alors une base de $\mathrm{Ker}(A - I_n)$, une base de $\mathrm{Ker}(A + I_n)$ (toutes deux formées de vecteurs de $\mathbb{R}^n$) et les p bases de la forme $(W_1, W'_1, \ldots, W_h, W'_h)$ ainsi construites : à la numérotation près, la base orthonormée obtenue fournit les colonnes de P.
L'unicité de r et de s vient de leur caractérisation comme dimension des espaces propres $E_{\pm 1}(A)$; les θ_i sont les arguments modulo π des valeurs propres, rangés par ordre croissant. L'assertion (ii) est donc une reformulation de (i).
L'assertion (iii) en découle, puisque le déterminant de la matrice écrite est $(-1)^s$. ∎

Exemples.

1. Rappelons qu'une réflexion est une symétrie orthogonale dont l'axe est un hyperplan. La matrice d'une réflexion dans une base orthonormée est orthogonale de déterminant -1 ; avec les notations du théorème ci-dessus, $r = n - 1$, $s = 1$ et $p = 0$.
2. Appelons *renversement* une symétrie orthogonale dont l'axe est de codimension 2 (et donc la direction de dimension 2). La matrice d'une réflexion dans une base orthonormée est orthogonale de déterminant 1 ; avec les notations du théorème ci-dessus, $r = n - 2$, $s = 2$ et $p = 0$.
3. Tout élément de $SO_2(\mathbb{R})$ est orthogonalement semblable à une matrice $R(\theta)$ (le cas de $\pm I_2$ revient à prendre $\theta = 0$ ou π). Le seul renversement de $SO_2(\mathbb{R})$ est I_2. Tout élément de $O_2(\mathbb{R}) \setminus SO_2(\mathbb{R})$ est orthogonalement semblable à la matrice $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Ce sont tous des réflexions.

4. Tout élément de $O_3(\mathbb{R})$ autre que l'identité est orthogonalement semblable à une matrice de la forme $\mathrm{Diag}\big(1, R(\theta)\big)$, où $\theta \in]0, \pi]$. C'est donc une rotation. C'est un renversement si, et seulement si, $\theta = \pi$. Tout élément de $O_3(\mathbb{R}) \setminus SO_3(\mathbb{R})$ autre que $-I_3$ est orthogonalement semblable à une matrice de la forme $\mathrm{Diag}\big(-1, R(\theta)\big)$, où $\theta \in [0, \pi[$. C'est une réflexion si, et seulement si, $\theta = 0$.

2.1.2 Générateurs

Il a été démontré (corollaire 48 de la page 175) que tout élément de $O_n(\mathbb{R})$ est produit d'au plus n réflexions. Pour un élément de $SO_n(\mathbb{R})$, le nombre de réflexions qui apparaissent dans le produit est pair, puisque le déterminant d'une réflexion est -1.

Théorème 22. On suppose $n \geqslant 3$. Alors, tout élément de $SO_n(\mathbb{R})$ est produit d'au plus n renversements.

Démonstration. Nous raisonnerons géométriquement, dans un espace euclidien E en identifiant $\mathcal{SO}(E)$ à $SO_n(\mathbb{R})$. Il suffit d'établir que tout produit ss' de deux réflexions s et s' est un produit rr' de deux renversements r et r'. Soient u et u' des vecteurs directeurs des directions de s et s'. S'ils sont colinéaires, $s = s'$, donc $ss' = \mathrm{Id}_E$ et la conclusion est trivial. On suppose donc que $\mathrm{Vect}(u, u')$ est un plan. Ce plan est stable par s et par s' (exercice amusant de géométrie euclidienne, laissé au lecteur). Soit v un vecteur non nul orthogonal à u et u' (c'est possible puisque $n \geqslant 3$). Le sous-espace $F := \mathrm{Vect}(u, u', v)$ est stable par s et s'. On définit r comme $-s$ sur F et comme l'identité sur l'orthogonal $G := F^\perp$; c'est un renversement dont la direction est l'orthogonal de u dans F. On définit similairement r'. La restriction de rr' à F est égale à $(-s)(-s') = ss'$, sa restriction à G est triviale (comme celle de ss') et l'on a bien $rr' = ss'$. ∎

2.1.3 Sous-groupes

Rappelons que le groupe $SO_2(\mathbb{R})$ a été étudié dans le module I.4 (page 169) : ce groupe est formé des rotations $R(\theta)$ et il est isomorphe au groupe $\mathbb{R}/2\pi\mathbb{Z}$, donc commutatif. Le groupe $O_2(\mathbb{R})$ admet $SO_2(\mathbb{R})$ comme sous-groupe distingué d'indice 2 : c'est le noyau du morphisme déterminant, qui a pour image $\{+1, -1\}$.

Exercice 15.

Quels éléments de $\mathcal{L}(E)$ commutent avec toute réflexion ? avec tout renversement ?

Solution. Les endomorphismes qui commutent avec une symétrie sont ceux qui laissent stable son axe et sa direction (cela se déduit des propositions 40 et 43 du module II.3 de [L1]). Les éléments de $\mathcal{L}(E)$ qui commutent avec toute réflexion laissent donc stable toute droite, ce sont donc des homothéties. Ceux qui commutent avec tout renversement laissent stables tous les plans. Si $\dim E := 2$, on ne peut rien conclure ; mais, si $\dim E \geqslant 3$, toute droite étant intersection de deux plans, on retrouve les homothéties.

Théorème 23. Le centre de $O_n(\mathbb{R})$ est $Z\big(O_n(\mathbb{R})\big) = \{I_n, -I_n\}$. Si $n \geqslant 3$, le centre de $SO_n(\mathbb{R})$ est $Z\big(SO_n(\mathbb{R})\big) = \{I_n, -I_n\}$ si n est pair, $\{I_n\}$ si n est impair.

Démonstration. Un élément de $M_n(\mathbb{R})$ qui commute à tout automorphisme orthogonal $\big($resp. à tout élément de$SO_n(\mathbb{R})\big)$ commute en particulier à toute réflexion (resp. à tout renversement). D'après l'exercice, il reste à déterminer les homothéties de $O_n(\mathbb{R})$ et de $SO_n(\mathbb{R})$, ce qui est facile et laissé en exercice au lecteur. ∎

2.1.4 Décompositions

Il existe de nombreux « théorèmes de décomposition » dans le groupe linéaire, faisant intervenir des structures euclidiennes et hermitiennes : citons les factorisation de Cholesky réelle et complexe et la factorisation QR réelle (théorème 47 de la page 173), qui ont été démontrées dans les modules I.3 et I.4 (pour la factorisation QR complexe, voir l'exercice I.6.23 de la page 309). Un corollaire de la factorisation QR est la décomposition d'Iwasawa (exercice I.6.24 de la page 309). La *décomposition polaire* ci-dessous est analogue aux « coordonnées polaires » : la matrice S représente un module, la matrice O un angle (en dimension 2, penser à une rotation). Pour le cas complexe, voir l'exercice I.6.25 de la page 309.

Théorème 24 (Décomposition polaire réelle). Toute matrice A de $GL_n(\mathbb{R})$ s'écrit de manière unique sous la forme $A = OS$, où $O \in O_n(\mathbb{R})$ et $S \in S_n^{++}(\mathbb{R})$.

Démonstration. Nous invoquerons le résultat établi dans le module I.4 (proposition 60 de la page 186) : l'application $S \mapsto S^2$ de $S_n^{++}(\mathbb{R})$ dans lui-même est bijective. De l'écriture $A := OS$, on tire ${}^tAA = S^2$; comme ${}^tAA \in S_n^{++}(\mathbb{R})$ (car A est inversible), la matrice $S \in S_n^{++}(\mathbb{R})$ est unique ainsi que $O := AS^{-1}$. Réciproquement, en prenant pour S la « racine carrée symétrique définie positive » de ${}^tAA$, puis en posant $O := AS^{-1}$, on vérifie que ${}^tOO = {}^tS^{-1}{}^tAAS^{-1} = S^{-1}S^2S^{-1} = I_n$: on a donc bien la décomposition cherchée. ∎

2.2 Propriétés géométriques

2.2.1 Familles et bases orthonormées

On a vu, dans le module I.2 (remarque de la page 71) que l'action du groupe orthogonal $\mathcal{O}(E)$ sur l'ensemble des bases orthonormées de l'espace vectoriel euclidien E est simplement transitive. Le même argument permet de voir que l'action du groupe *unitaire* $\mathcal{U}(E)$ sur l'ensemble des bases orthonormées de l'espace vectoriel *hermitien* E est également simplement transitive. La possibilité de compléter toute famille orthonormée en une base orthonormée montre que, pour r fixé, $\mathcal{O}(E)$ et $\mathcal{U}(E)$ agissent transitivement sur l'ensemble des familles orthonormées de r éléments (même argument que pour la proposition 9 de la page 266). Dans le cas d'un espace euclidien *orienté* (page 266), on voit de plus que $\mathcal{SO}(E)$ agit simplement transitivement sur l'ensemble des bases orthonormées *directes*.

Exercice 16.
Si $1 \leqslant r < \dim E$, montrer que $\mathcal{SO}(E)$ (resp. $\mathcal{SU}(E)$) agit transitivement sur l'ensemble des familles orthonormées de r éléments.

Solution. Soient $(u_1, \dots, u_r)$ et $(u'_1, \dots, u'_r)$ deux familles orthonormées. On les complète en des bases orthonormées $(u_1, \dots, u_n)$ et $(u'_1, \dots, u'_n)$: la matrice de passage est dans $\mathcal{O}(E)$ (resp. $\mathcal{U}(E)$). Soit λ son déterminant : donc $|\lambda| = 1$; quitte à changer u_n en λu_n (ce qui n'affecte pas la famille de départ, puisque $r < n$), on est ramené à une matrice de passage dans $\mathcal{SO}(E)$ (resp. $\mathcal{SU}(E)$).

2.2.2 Sous-espaces, drapeaux

Le théorème d'orthonormalisation de Gram-Schmidt (corollaire 19 de la page 156) peut être interprété comme suit : *tout drapeau maximal admet une base adaptée orthonormée* (dans le cas euclidien comme dans le cas hermitien). Il en découle que $\mathcal{O}(E)$ (resp. $\mathcal{U}(E)$) agit transitivement sur les drapeaux maximaux. Comme on peut multiplier le dernier vecteur d'une base orthonormée par λ tel que $|\lambda| = 1$ (sans changer le drapeau), on voit même que $\mathcal{SO}(E)$ (resp. $\mathcal{SU}(E)$) agit transitivement sur les drapeaux maximaux. Comme tout drapeau s'étend en un drapeau maximal, on en déduit facilement que, pour tout $d := (d_1, \dots, d_k)$, $\mathcal{SO}(E)$ (resp. $\mathcal{SU}(E)$) agit transitivement sur les d-drapeaux. En particulier, pour tout r, ils agissent transitivement sur l'ensemble $S_r(E)$ des sous-espaces vectoriels de dimension r.

Exercice 17.
Décrire matriciellement les stabilisateurs de l'action de $\mathcal{SO}(E)$ (resp. $\mathcal{SU}(E)$) sur l'ensemble des drapeaux maximaux.

Solution. Une fois choisis un drapeau maximal et une base adaptée orthonormée, on voit tout d'abord qu'une matrice du stabilisateur est triangulaire, puis (du fait qu'elle est orthogonale, resp. unitaire) que c'est une matrice diagonale. Le stabilisateur est donc le groupe des $\mathrm{Diag}(\lambda_1, \dots, \lambda_n)$ tels que $|\lambda_1| = \cdots = |\lambda_n| = 1$ et $\lambda_1 \cdots \lambda_n = 1$.

2.2.3 Sphères

La *sphère unité* d'un espace euclidien E est l'ensemble des vecteurs de norme 1. En particulier, on note S^{n-1} la sphère unité de $\mathbb{R}^n$ (muni de sa structure euclidienne canonique). Par exemple, le cercle $S^1 = \{(\cos\theta, \sin\theta) \mid \theta \in \mathbb{R}\}$ est de « dimension » 1, c'est à dire qu'un seul paramètre suffit à le décrire (c'est une courbe). De la même manière, la sphère $S^2 = \{(\cos\theta\cos\varphi, \sin\theta\cos\varphi, \sin\varphi) \mid \theta, \varphi \in \mathbb{R}\}$ est de « dimension » 2 (deux paramètres sont nécessaires et suffisants pour la décrire, c'est une surface), etc.

Soit E un espace euclidien. Le groupe orthogonal $\mathcal{O}(E)$ agit transitivement sur la sphère unité S de E. Pour décrire les stabilisateurs, fixons un élément $u \in S$. Tout élément du stabilisateur $\mathcal{O}_u(E)$ de u laisse stable l'hyperplan $E' := \mathrm{Vect}(u)^\perp$. De plus, par restriction à E', on obtient un isomorphisme : $\mathcal{O}_u(E) \simeq \mathcal{O}(E')$. Du module I.2 (théorème 14 de la page 72) on déduit alors que l'application $u \mapsto g.u$ de $\mathcal{O}(E)$ sur S donne, par passage au quotient, une bijection : $\mathcal{O}(E)/\mathcal{O}(E') \to S$. Dans le cas particulier où $E := \mathbb{R}^n$, on

trouve ainsi une bijection : $O_n(\mathbb{R})/O_{n-1}(\mathbb{R}) \to S^{n-1}$. Pour la réaliser, on peut choisir pour u le « pôle nord » $(0,\dots,0,1)$ de S^{n-1}. L'application de $O_n(\mathbb{R})$ sur S^{n-1} qui, à une matrice A, associe sa dernière colonne, est surjective ; et les matrices A, B ont même image si, et seulement si, $A^{-1}B = \begin{pmatrix} C & 0 \\ 0 & 1 \end{pmatrix}$, avec $C \in O_{n-1}(\mathbb{R})$.

Exercice 18.
L'action de $SO_n(\mathbb{R})$ sur S^{n-1} est-elle transitive ? Stabilisateurs ?
Solution. L'action est transitive. Le stabilisateur du pôle nord s'identifie à $SO_{n-1}(\mathbb{R})$.

2.2.4 Représentations orthogonales, représentations unitaires

On dit que la représentation linéaire $\rho : G \to \mathcal{GL}(E)$ est *orthogonale* (resp. *unitaire*) si E est un espace euclidien (resp. hermitien) et $\operatorname{Im}\rho \subset \mathcal{O}(E)$ (resp. $\operatorname{Im}\rho \subset \mathcal{U}(E)$). Nous démontrerons les résultats qui suivent dans le cas réel ; le cas complexe est en tous points semblable. On obtient d'abord une nouvelle démonstration du théorème de Maschke dans le cas où le corps de base est $\mathbb{R}$ ou $\mathbb{C}$.

Proposition 25. (i) Soit E un espace vectoriel de dimension finie sur $\mathbb{R}$ (resp. sur $\mathbb{C}$) et soit $\rho : G \to \mathcal{GL}(E)$ une représentation linéaire d'un groupe fini G. Il existe alors un produit scalaire sur E qui fait de ρ une représentation orthogonale (resp. unitaire).
(ii) Toute représentation orthogonale (resp. unitaire) est complètement réductible.

Démonstration. (i) La preuve repose sur un argument de moyennisation. On introduit un produit scalaire arbitraire $[x,y]$ sur E, puis l'on pose : $(x|y) := \sum_{g\in G} [gx, gy]$. Il est immédiat que l'on obtient ainsi une forme bilinéaire symétrique ; elle est de plus définie positive, puisque, si $x \neq 0$, tous les $[gx, gx]$ sont strictement positifs, donc également $(x|x)$. Enfin, fixant $g_0 \in G$, on calcule : $(g_0x|g_0y) = \sum_{g\in G} [gg_0x, gg_0y] = \sum_{g\in G} [gx, gy] = (x|y)$, puisque l'application $g \mapsto gg_0$ est bijective de G dans lui-même. L'égalité $(g_0x|g_0y) = (x|y)$ dit que $\rho(g_0) \in \mathcal{O}(E)$, d'où $\operatorname{Im}\rho \subset \mathcal{O}(E)$.
(ii) Comme dans la preuve du théorème de Maschke (théorème 11 de la page 269), il suffit de démontrer que tout sous-espace stable F de E admet un supplémentaire stable : mais c'est le cas de $F^\perp$ (si $f \in \mathcal{O}(E)$ et si F est stable par F, alors $F^\perp$ l'est aussi). ∎

Proposition 26. Soit E un espace euclidien (resp. hermitien). Tout sous-groupe fini de $\mathcal{GL}(E)$ est conjugué à un sous-groupe de $\mathcal{O}(E)$ (resp. de $\mathcal{U}(E)$). Par conséquent, tout sous-groupe fini de $GL_n(\mathbb{R})$ (resp. de $GL_n(\mathbb{C})$) est conjugué à un sous-groupe fini de $O_n(\mathbb{R})$ (resp. de $U_n(\mathbb{C})$).

Démonstration. Notons $[x,y]$ le produit scalaire qui donne sa structure euclidienne à E. Soit G un sous-groupe fini de $\mathcal{GL}(E)$. L'inclusion G dans $\mathcal{GL}(E)$ est une représentation. Soit $(x|y)$ un produit scalaire sur E pour lequel elle est orthogonale. Soient $\mathcal{B}$ une base orthonormée pour le produit scalaire $[x,y]$ et $\mathcal{C}$ une base orthonormée pour le produit scalaire $(x|y)$. Soit enfin $\varphi \in \mathcal{GL}(E)$ qui transforme $\mathcal{B}$ en $\mathcal{C}$. On a identiquement : $[x,y] = \big(\varphi(x)|\varphi(y)\big)$ (c'est vrai pour les éléments de $\mathcal{B}$ et cela s'étend par bilinéarité). On en déduit, pour tout $g \in G$:

$$\big[\varphi^{-1}g\varphi(x), \varphi^{-1}g\varphi(y)\big] = \big(g\varphi(x), g\varphi(y)\big) = \big(\varphi(x)|\varphi(y)\big) = [x,y],$$

donc $\varphi^{-1}g\varphi \in \mathcal{O}(E)$, d'où l'on conclut que $\varphi^{-1}G\varphi \subset \mathcal{O}(E)$. ∎

2.2.5 Le groupe des isométries d'un espace affine euclidien

Un *espace affine euclidien* est un espace affine X sur $\mathbb{R}$ dont la direction E est munie d'un produit scalaire, qui en fait donc un espace euclidien. Une *isométrie* de X est alors une transformation affine de X dont la partie linéaire est dans $\mathcal{O}(E)$. Le morphisme surjectif $\pi : \mathcal{G}A(X) \to \mathcal{GL}(E)$ permet de définir le *groupe des isométries affines de* X comme $\mathrm{Is}(X) := \pi^{-1}\big(\mathcal{O}(E)\big)$ et le *sous-groupe des isométries affines positives, ou directes de* X (encore appelées *déplacements*) comme $\mathrm{Is}^+(X) := \pi^{-1}\big(\mathcal{SO}(E)\big)$. Ainsi, π induit un morphisme surjectif $\mathrm{Is}(X) \to \mathcal{O}(E)$, de noyau $\mathcal{T}(X)$; et morphisme surjectif $\mathrm{Is}^+(X) \to \mathcal{SO}(E)$, également de noyau $\mathcal{T}(X)$. Les mêmes raisonnements qu'à la section 1.2.3 de la page 269 permettent d'établir que, pour tout choix d'une origine $O \in X$, le stabilisateur $\mathrm{Is}_O(X)$ de O dans $\mathrm{Is}(X)$ est isomorphe (via π) à $\mathcal{O}(E)$. De même, le stabilisateur $\mathrm{Is}^+_O(X)$ de O dans $\mathrm{Is}^+(X)$ est isomorphe (via π) à $\mathcal{SO}(E)$. Le groupe $\mathrm{Is}^+(X)$ est donc un sous-groupe distingué d'indice 2 de $\mathrm{Is}(X)$.

Groupe des symétries d'une figure. Soit Y un sous-ensemble de X (une « figure »). L'ensemble des isométries (resp. des déplacements) qui laissent Y globalement invariante est un sous-groupe G de $\mathrm{Is}(X)$, appelé *groupe des symétries* de cette figure. Le sous-groupe distingué $G^+ := G \cap \mathrm{Is}^+(X)$ est le *groupe des symétries directes*. C'est un sous-groupe d'indice 2 de G si G contient des isométries non directes, sinon $G^+ = G$. L'exemple du cube de $\mathbb{R}^3$ a été étudié dans le module I.2 (exemple 2 de la page 76), d'autres exemples le seront à la section 2.4.2 et dans [MPA1], chapitre sur les groupes. Puisque G agit sur Y, on a un morphisme de groupes de G dans $\mathfrak{S}(Y)$. Le noyau de ce morphisme est formé des isométries qui laissent invariants tous les points de Y : c'est donc le groupe $\mathrm{Is}_Y(X) := \bigcap_{a \in Y} \mathrm{Is}_a(X)$.

Supposons que Y engendre affinement X (exercice I.6.14 de la page 308). Alors une transformation affine est déterminée par son effet sur Y. Par conséquent, le morphisme de G dans $\mathfrak{S}(Y)$ est injectif.

Si l'on suppose de plus que Y est fini, il en découle que le groupe G est fini. Dans ce cas, il a été démontré dans le module I.2 (exercice 9 de la page 73) que G fixe au moins un point $O \in X$, donc que G est un sous-groupe de $\mathrm{Is}_O(X)$. Par le choix de l'origine O, on peut donc considérer G comme un sous-groupe de $\mathcal{O}(E)$ et *linéariser* le problème.

2.3 Propriétés topologiques et différentielles

Nous reprenons ici les conventions de 1.3 et ne donnons que des énoncés matriciels : les traductions pour $\mathcal{O}(E)$, $\mathcal{SO}(E)$, $\mathcal{U}(E)$ et $\mathcal{SU}(E)$ sont laissées au plaisir du lecteur.

Proposition 27. (i) Les groupes $O_n(\mathbb{R})$, $SO_n(\mathbb{R})$, $U_n(\mathbb{C})$ et $SU_n(\mathbb{C})$ sont compacts.
(ii) Les groupes $SO_n(\mathbb{R})$, $U_n(\mathbb{C})$ et $SU_n(\mathbb{C})$ sont connexes par arcs. Le groupe $O_n(\mathbb{R})$ a deux composantes connexes, $O_n^+(\mathbb{R}) = SO_n(\mathbb{R})$ et $O_n^-(\mathbb{R})$.

Démonstration. (i) Les groupes $SO_n(\mathbb{R})$ et $SU_n(\mathbb{C})$ sont fermés respectivement dans $O_n(\mathbb{R})$ et $U_n(\mathbb{C})$ car définis par l'équation $\det A = 1$; il suffit donc de montrer

que $O_n(\mathbb{R})$ et $U_n(\mathbb{C})$ sont compacts. Ces derniers sont respectivement définis par les équations ${}^tAA = I_n$ et $A^*A = I_n$, donc fermés dans $M_n(\mathbb{R})$ et $M_n(\mathbb{C})$. Enfin, ils sont bornés (chaque colonne d'une matrice orthogonale ou unitaire est de norme euclidienne 1).
(ii) Une matrice $A \in SO_n(\mathbb{R})$ est de la forme $P\,\mathrm{Diag}\big(I_r, -I_s, R(\theta_1), \ldots, R(\theta_p)\big)P^{-1}$, la matrice P étant de plus orthogonale et s étant pair (théorème 21 de la page 281). Écrivons $-I_s$ sous la forme d'une diagonale de $s/2$ blocs $R(\pi) = -I_2$. Pour tout $t \in [0,1]$, on pose : $A(t) := P\,\mathrm{Diag}\big(I_r, R(t\pi), \ldots, R(t\pi), R(t\theta_1), \ldots, R(t\theta_p)\big)P^{-1}$ (avec $s/2$ blocs $R(t\pi)$), et l'on a $A(0) = I_n$, $A(1) = A$ et, pour tout $t \in [0,1]$, $A(t) \in SO_n(\mathbb{R})$. Il en découle que $O_n^-(\mathbb{R}) = \mathrm{Diag}(-1,1,\ldots,1)SO_n(\mathbb{R})$ est également connexe par arcs. Mais $O_n(\mathbb{R})$ ne peut l'être puisque le déterminant n'y a pas un signe constant.

Une matrice $A \in U_n(\mathbb{C})$ se met sous la forme $A = P\,\mathrm{Diag}(e^{\mathrm{i}\theta_1}, \ldots, e^{\mathrm{i}\theta_n})P^{-1}$, où P est unitaire et les θ_i réels (théorème 113 de la page 204). On pose alors :

$$A(t) := P\,\mathrm{Diag}(e^{\mathrm{i}t\theta_1}, \ldots, e^{\mathrm{i}t\theta_n})P^{-1},$$

d'où $A(0) = I_n$, $A(1) = A$ et, pour tout $t \in [0,1]$, $A(t) \in U_n(\mathbb{C})$. Si $A \in SU_n(\mathbb{C})$, on peut de plus supposer que $\theta_1 + \cdots + \theta_p = 0$. Avec le même choix, on voit que pour tout $t \in [0,1]$, $A(t) \in SU_n(\mathbb{C})$. ∎

Proposition 28.
L'application $(O,S) \mapsto OS$ est un homéomorphisme de $O_n(\mathbb{R}) \times S_n^{++}(\mathbb{R})$ sur $GL_n(\mathbb{R})$.

Démonstration. D'après le théorème 24 de la page 283, c'est une bijection. Les coefficients de OS étant calculés à partir de ceux de O et de S par additions et multiplications, l'application $(O,S) \mapsto OS$ est continue. La preuve, plus délicate, que l'application réciproque est continue fait l'objet de l'exercice I.6.28 de la page 310. ∎

Comme dans le cas $SL_n(\mathbb{R})$ et $SL_n(\mathbb{C})$ (proposition 17 de la page 276), nous appellerons *espace tangent en* I_n, pour chacun des groupes $O_n(\mathbb{R})$, $U_n(\mathbb{C})$ et $SU_n(\mathbb{C})$, l'ensemble des vecteurs tangents à ce groupe en I_n. Ces espaces tangents seront respectivement notés $\mathrm{o}_n(\mathbb{R})$, $\mathrm{u}_n(\mathbb{C})$ et $\mathrm{su}_n(\mathbb{C})$. Nous n'avons pas besoin de notation pour l'espace tangent à $SO_n(\mathbb{R})$, car il est égal à $\mathrm{o}_n(\mathbb{R})$: en effet, tout chemin dans $O_n(\mathbb{R})$ qui part de I_n est entièrement contenu dans la composante connexe de I_n dans $O_n(\mathbb{R})$, et cette dernière est $SO_n(\mathbb{R})$ (proposition 27 de la page précédente).

Proposition 29. (i) L'espace $\mathrm{o}_n(\mathbb{R})$ est égal à l'espace $\mathrm{o}'_n(\mathbb{R})$ des matrices antisymétriques de $M_n(\mathbb{R})$. L'espace $\mathrm{u}_n(\mathbb{C})$ est égal à l'espace $\mathrm{u}'_n(\mathbb{C})$ des matrices antihermitiennes de $M_n(\mathbb{C})$. L'espace $\mathrm{su}_n(\mathbb{C})$ est égal à l'espace $\mathrm{su}'_n(\mathbb{C})$ des matrices antihermitiennes de trace nulle de $M_n(\mathbb{C})$.
(ii) L'exponentielle envoie chaque espace tangent dans le groupe correspondant.

Démonstration.
Soit $A(t)$ un chemin de classe $\mathcal{C}^1$ dans $O_n(\mathbb{R})$ et tel que $A(0) = I_n$. En dérivant l'iden-

tité ${}^tA(t)A(t) = I_n$ et en évaluant en $t := 0$, on trouve ${}^tA'(0) + A'(0) = 0$. L'espace $\mathrm{o}_n(\mathbb{R})$ est donc inclus dans l'espace $\mathrm{o}'_n(\mathbb{R})$ des matrices antisymétriques de $M_n(\mathbb{R})$. L'inclusion $\mathrm{u}_n(\mathbb{C}) \subset \mathrm{u}'_n(\mathbb{C})$ se prouve de la même manière (mais en tenant compte de la conjugaison). Si de plus le chemin est dans $SU_n(\mathbb{C})$, donc dans $SL_n(\mathbb{C})$, on a déjà vu que $\operatorname{Tr} A'(0) = 0$ (proposition 17 de la page 276). L'espace $\mathrm{su}_n(\mathbb{R})$ est donc inclus dans $\mathrm{su}'_n(\mathbb{R})$.
Avant de démontrer les inclusions réciproques, étudions l'effet de l'exponentielle. De l'égalité ${}^tA = -A$ on tire, en posant $B := \exp A$, l'égalité ${}^tB = B^{-1}$. L'exponentielle envoie donc $\mathrm{o}'_n(\mathbb{R})$ dans $O_n(\mathbb{R})$. On prouve de même que l'exponentielle envoie $\mathrm{u}'_n(\mathbb{C})$ dans $U_n(\mathbb{C})$. Comme $\mathrm{su}'_n(\mathbb{C}) = \mathrm{u}'_n(\mathbb{C}) \cap \mathrm{sl}_n(\mathbb{C})$, en déduit que l'exponentielle envoie $\mathrm{su}'_n(\mathbb{C})$ dans $SU_n(\mathbb{C})$. Cela entraîne *a fortiori* l'assertion (ii).
Soit enfin $A \in \mathrm{o}'_n(\mathbb{R})$. En posant $B(t) := \exp tA$, on définit un chemin de classe $\mathcal{C}^1$ dans $O_n(\mathbb{R})$ tel que $B(0) = I_n$ et $B'(0) = A$: donc A appartient à l'espace tangent $\mathrm{o}_n(\mathbb{R})$. Les autres inclusions se montrent de la même manière. ∎

Exemple. Dans le cas où $n := 2$, l'espace $\mathrm{o}_n(\mathbb{R})$ est formé des matrices de la forme $A(a) := \begin{pmatrix} 0 & -a \\ a & 0 \end{pmatrix}$, avec $a \in \mathbb{R}$. Le groupe $SO_2(\mathbb{R})$ est formé des matrices $R(\theta)$ (module I.4, page 169). L'exponentielle de $A(a)$ est $R(a)$ (exercice I.6.29 de la page 310).

Retour sur le repère de Serret-Frenet. Le repère (ou trièdre) de Serret-Frenet a été défini dans le module « Courbes paramétrées » de [L1]. On peut le considérer comme une fonction $X : I \to M_3(\mathbb{R})$ qui, à tout t de l'intervalle ouvert $I \subset \mathbb{R}$, associe la matrice (spéciale orthogonale, puisque le repère est orthonormé direct) des vecteurs de ce repère dans la base canonique de $\mathbb{R}^3$. En accord avec l'origine cinématique du problème, nous étudierons une fonction $X(t)$ de classe $\mathcal{C}^1$ et à valeurs dans $SO_n(\mathbb{R})$. Alors la fonction matricielle $t \mapsto A(t) := X'(t)\big(X(t)\big)^{-1} = X'(t){}^tX(t)$ est continue de I dans $M_n(\mathbb{R})$ et l'on a une équation différentielle $X' = AX$. En dérivant l'identité $X(t){}^tX(t) = I_n$, on obtient $X'(t){}^tX(t) + X(t){}^tX'(t) = 0$. Reportant la relation $X' = AX$, on trouve : $A(t)X(t){}^tX(t) + X(t){}^tX(t){}^tA(t) = 0$, d'où enfin (puisque $X(t){}^tX(t) = I_n$) :

$$\forall t \in I\ ,\ A(t) + {}^tA(t) = 0.$$

Ainsi, $A(t)$ est continue de I dans l'espace $\mathrm{o}_n(\mathbb{R})$ des matrices antisymétriques de $M_n(\mathbb{R})$.
Supposons réciproquement que A soit une telle fonction et que X soit une solution de classe $\mathcal{C}^1$ de l'équation différentielle $X' = AX$. Nous imposerons de plus la *condition initiale* $X(t_0) = I_n$. Alors, en dérivant la fonction $t \mapsto {}^tX(t)X(t)$, on trouve :

$$\begin{aligned} {}^tX'(t)X(t) + {}^tX(t)X'(t) &= {}^tX(t){}^tA(t)X(t) + {}^tX(t)A(t)X(t) \\ &= {}^tX(t)\big({}^tA(t) + A(t)\big)X(t) = 0, \end{aligned}$$

puisque $A(t)$ est antisymétrique. Ainsi, ${}^tX(t)X(t)$ est constant, donc égal à I_n (à cause de la condition initiale). On a donc un chemin dans $O_n(\mathbb{R})$, qui passe par I_n, donc dans la

composante connexe par arcs $SO_n(\mathbb{R})$.
Soit maintenant $Y(t)$ une solution quelconque de l'équation différentielle (on n'impose plus de condition initiale). On pose $Z := X^{-1}Y = {}^tXY$ et l'on vérifie par un petit calcul que $Z' = {}^tX\big({}^tA + A\big)Y = 0$, donc que Z est constant. Comme $Z(t_0) = Y(t_0)$, on a en fin de compte : $Y(t) = X(t)Y(t_0)$. À partir de ce résultat on peut retrouver le « théorème fondamental » de *loc. cit.* : *une courbe de l'espace birégulière est essentiellement caractérisée par sa courbure et sa torsion*. Le cas où la matrice A est constante correspond à $X(t) = \exp(tA)$.

2.4 Les groupes $O_3(\mathbb{R})$ et $SO_3(\mathbb{R})$

2.4.1 Rotations de $\mathbb{R}^2$ et $\mathbb{R}^3$

Nous utiliserons ici les résultats du module I.4 ainsi que les exemples de la page 281. Soit Π un plan euclidien orienté. Le groupe $\mathcal{SO}(\Pi)$ agit simplement transitivement sur l'ensemble de ses bases orthonormées directes. Comme ce groupe est commutatif (il est isomorphe au groupe $SO_2(\mathbb{R})$, qui a été décrit dans le module I.4), la matrice d'un élément $\varphi \in \mathcal{SO}(\Pi)$ est la même dans l'une quelconque de ses bases. On obtient donc un isomorphisme $\mathcal{SO}(\Pi) \simeq SO_2(\mathbb{R})$ qui ne dépend que du choix de l'orientation de Π (et non du choix d'une base orthonormée directe particulière). Nous dirons que φ est une *rotation d'angle* θ si l'élément de $SO_2(\mathbb{R})$ qui lui est ainsi associé est $R(\theta)$. L'angle θ est défini modulo 2π (si l'on change d'orientation, l'angle est transformé en son opposé). Par ailleurs, il a été démontré dans *loc. cit.* que tout élément de $\mathcal{O}(\Pi) \setminus \mathcal{SO}(\Pi)$ est une symétrie orthogonale par rapport à une droite, c'est-à-dire une réflexion.

Pour l'étude en dimension 3, nous nous cantonnerons au cas de l'espace euclidien orienté $\mathbb{R}^3$. Toute matrice $A \in SO_3(\mathbb{R}) \setminus \{I_3\}$ est orthogonalement semblable à une unique matrice $\mathrm{Diag}\big(1, R(\theta)\big)$ avec $0 < \theta \leqslant \pi$; et c'est un renversement si, et seulement si, $\theta = \pi$. Il revient au même de dire que A admet une unique droite Δ de points fixes et que sa restriction au plan $\Pi := \Delta^\perp$ est une rotation d'angle θ pour une orientation convenable du plan Π. Cependant, la matrice A^{-1} admet également Δ pour droite de points fixes et sa restriction au plan Π est une rotation d'angle θ *pour l'orientation opposée*.
Pour éviter cette ambiguïté, on se restreint aux éléments de $SO_3(\mathbb{R})$ qui ne sont pas des renversements ; on convient de munir le plan Π de l'unique orientation telle que la restriction de A à Π est la rotation d'angle θ (c'est possible parce que $\theta \not\equiv -\theta \mod 2\pi$). Il y a alors sur Δ une unique orientation telle que, si u est une base directe de Δ et si (v, w) est une base directe de Π, alors (u, v, w) est une base directe de $\mathbb{R}^3$ (on peut prendre pour u le produit vectoriel $v \wedge w$). Notant $\vec{\Delta}$ la droite orientée correspondante, nous dirons que A est une rotation d'axe (orienté) $\vec{\Delta}$ et d'angle θ, que l'on notera $R(\vec{\Delta}, \theta)$. On obtient ainsi une bijection entre les couples $(\vec{\Delta}, \theta)$ et les éléments de $SO_3(\mathbb{R})$ qui ne sont pas des renversements. Naturellement, on peut encore définir un renversement $R(\vec{\Delta}, \pi)$, mais il est est impossible de le distinguer du renversement $R(\vec{\Delta}', \pi)$, d'axe $\vec{\Delta}'$ opposé à $\vec{\Delta}$.

Exercice 19.
(i) Comment calculer l'angle de $A \in SO_3(\mathbb{R})$ et reconnaître si c'est un renversement ?
(ii) On suppose que A n'est ni un renversement ni l'identité. Calculer son axe à partir de sa partie antisymétrique $B := \dfrac{A - {}^tA}{2}$.
Commencer par le cas d'une matrice $\mathrm{Diag}\big(1, R(\theta)\big)$.

Solution.
(i) On a $A = PA'P^{-1}$, avec $A' := \mathrm{Diag}\big(1, R(\theta)\big)$ et $P \in O_3(\mathbb{R})$, d'où $\mathrm{Tr}\, A = 1+2\cos\theta$, ce qui fournit θ, et caractérise les renversements par l'égalité $\mathrm{Tr}\, A = -1$.
(ii) La partie antisymétrique de A' est $B' := \mathrm{Diag}\big(0, C\big)$, où $C := \begin{pmatrix} 0 & -\sin\theta \\ \sin\theta & 0 \end{pmatrix}$ et, par hypothèse, $0 < \theta < \pi$. L'axe de A' est $\mathrm{Vect}(1,0,0)$, qui est égal au noyau de B'. Comme la conjugaison par une matrice orthogonale conserve la propriété d'antisymétrie, on a $B = PB'P^{-1}$. Le noyau de B (resp. l'axe de A) est l'image par P du noyau de B' (resp. de l'axe de A'). *L'axe de A est donc le noyau de sa partie antisymétrique.*

Toute matrice $A \in O_3(\mathbb{R}) \setminus SO_3(\mathbb{R})$ autre que $-I_3$ est orthogonalement semblable à une unique matrice de la forme $\mathrm{Diag}\big(-1, R(\theta)\big)$, où $\theta \in [0, \pi[$ (et c'est une réflexion si, et seulement si, $\theta = 0$). C'est le produit de deux matrices qui commutent : $\mathrm{Diag}\big(-1, R(\theta)\big)$ et $\mathrm{Diag}(-1,1,1)$. La matrice A admet donc une unique décomposition comme produit d'une matrice de rotation $R(\vec{\Delta}, \theta)$ et de la matrice de réflexion d'axe $\Pi := \Delta^\perp$, et ces deux matrices commutent. On notera $S(\vec{\Delta}, \theta)$ cette matrice. On obtient ainsi une bijection entre les couples $(\vec{\Delta}, \theta)$ et les éléments de $O_3(\mathbb{R}) \setminus SO_3(\mathbb{R})$ autres que $-I_3$. On peut encore définir $S(\vec{\Delta}, \pi)$, mais c'est $-I_3$ quel que soit l'axe orienté $\vec{\Delta}$. De cette étude découle :

Proposition 30. Deux éléments A_1 et A_2 du groupe $O_3(\mathbb{R})$ sont conjugués (ou, ce qui revient au même, orthogonalement semblables) si, et seulement si, l'un des cas suivants se produit : $A_1 = A_2 = I_3$ ou $A_1 = R(\vec{\Delta}_1, \theta)$ et $A_2 = R(\vec{\Delta}_2, \theta)$ (et ces cas prennent en compte tous les éléments de $SO_3(\mathbb{R})$) ; $A_1 = A_2 = -I_3$ ou $A_1 = S(\vec{\Delta}_1, \theta)$ et $A_2 = S(\vec{\Delta}_2, \theta)$ (et ces cas prennent en compte tous les éléments de $O_3(\mathbb{R}) \setminus SO_3(\mathbb{R})$).

Corollaire 31. Soit E un espace euclidien orienté de dimension 3.
Les rotations $R(\vec{\Delta}_1, \theta)$ et $R(\vec{\Delta}_2, \theta)$ sont conjuguées dans $\mathcal{SO}(E)$.

Démonstration. Notons d'abord que $\mathcal{SO}(E)$ opère transitivement sur l'ensemble des axes orientés : en effet, leurs bases directes sont simplement les éléments de la sphère unité de E. Si maintenant $\varphi \in \mathcal{SO}(E)$ transforme $\vec{\Delta}_1$ en $\vec{\Delta}_2$, les rotations $\varphi \circ R(\vec{\Delta}_1, \theta) \circ \varphi^{-1}$ et $R(\vec{\Delta}_2, \theta)$: ont même axe orienté $\vec{\Delta}_2 = \varphi\big(\vec{\Delta}_1\big)$ et même angle, donc sont égales. ∎

2.4.2 Sous-groupes de $O_3(\mathbb{R})$

Nous allons d'abord étudier un sous-groupe fini de $O_3(\mathbb{R})$ qui est un groupe de symétrie au sens de la page 286. Cette étude, amorcée dans le module I.2, sera achevée dans [MPA1], chapitre sur les groupes.

Le groupe du tétraèdre

Créons d'abord un *tétraèdre régulier*, c'est-à-dire quatre points A, B, C, D de $\mathbb{R}^3$ dont les six distances sont égales. On choisit d'abord $A, B, C \in \mathbb{R}^2 \times \{0\}$ formant un triangle équilatéral centré en $(0,0,0)$: par exemple, s'inspirant de l'étude des racines de l'unité dans $\mathbb{C}$, on peut prendre $A := (1,0,0)$, $B := (-1/2, \sqrt{3}/2, 0)$ et $C := (-1/2, -\sqrt{3}/2, 0)$. Leurs distances sont $AB = BC = CA = \sqrt{3}$. Il reste à prendre D sur la $\{(0,0)\} \times \mathbb{R}$ à distance $\sqrt{3}$ de A, B et C : par exemple, $D := (0,0,\sqrt{2})$. À partir de ce point, nous utiliserons uniquement les propriétés suivantes des points A, B, C, D : les six longueurs AB, AC, AD, BC, BD, CD sont égales à un réel $\ell > 0$; l'isobarycentre $O := \frac{1}{4}A + \frac{1}{4}B + \frac{1}{4}C + \frac{1}{4}F$ est l'origine $(0,0,0)$ de $\mathbb{R}^3$.

Le sous-ensemble $Y := \{A, B, C, D\}$ de l'espace affine euclidien $X := \mathbb{R}^3$ réalise toutes les conditions de l'étude de la page 286. Le groupe des symétries de Y est donc fini, il se plonge dans le groupe symétrique $\mathfrak{S}(Y)$, qui a vingt-quatre éléments. Comme ce groupe d'isométries est fini, on peut linéariser le problème en prenant pour origine l'isobarycentre O de A, B, C, D (toute transformation affine qui laisse stable Y fixe O). Ici, $O = (0,0,0)$ et l'on obtient directement G comme sous-groupe de $\mathcal{O}(\mathbb{R}^3) = O_3(\mathbb{R})$. Nous allons vérifier que l'injection de G dans $\mathfrak{S}(Y)$ (que nous noterons $g \mapsto \sigma_g$) est en fait bijective.

Comme le groupe symétrique est engendré par les transpositions, il suffit de vérifier que chaque transposition est induite par une isométrie de Y. Les quatre points jouant des rôles symétriques, considérons la transposition $\tau_{A,B}$ de A et B. Le lecteur prouvera que le plan médiateur de AB (plan orthogonal à AB en son milieu) passe par C et D. La réflexion $r_{A,B}$ d'axe ce plan est donc un élément de $O_3(\mathbb{R})$ qui échange A et B et laisse fixe C et D. C'est donc un élément de G dont l'image dans $\mathfrak{S}(Y)$ est $\tau_{A,B}$.

Il découle de cette construction que $G \simeq \mathfrak{S}(Y)$ et que les six réflexions $r_{A,B}, r_{A,C}, \ldots$ engendrent G. Comparons les deux morphismes $g \mapsto \det g$ et $g \mapsto \varepsilon(\sigma_g)$ (signature) de G dans $\mathbb{R}^*$: puisque le déterminant d'une réflexion et la signature d'une transposition valent -1, ces deux morphismes coïncident sur la famille génératrice $r_{A,B}, r_{A,C}, \ldots$, donc ils sont égaux : pour tout $g \in G$, le déterminant de g est égal à la signature de σ_g.

Il en résulte que le sous-groupe $G^+ := G \cap SO_3(\mathbb{R})$ des symétries directes de Y est isomorphe à $\mathfrak{A}^4$. On peut le vérifier en décrivant la réalisation géométrique des permutations paires de Y. La permutation $\tau_{A,B} \circ \tau_{C,F}$ est réalisée comme le renversement dont l'axe est la droite qui joint les milieux de AB et de CD. Les deux 3-cycles de support $\{A, B, C\}$ sont réalisés comme les rotations d'angle $2\pi/3$ dont l'axe (non orienté) est la droite qui joint D au centre de gravité du triangle ABC.

La simplicité de $SO_3(\mathbb{R})$

Puisque $-I_3 \in O_3(\mathbb{R}) \setminus SO_3(\mathbb{R})$, l'application $(A,B) \mapsto AB$ de $SO_3(\mathbb{R}) \times \{+I_3, -I_3\}$ sur $O_3(\mathbb{R})$ est un isomorphisme. Cela permet de ramener en partie l'étude des sous-groupes de $O_3(\mathbb{R})$ à celle des sous-groupes de $SO_3(\mathbb{R})$. En ce qui concerne ce dernier, le phénomène le plus notable est qu'il n'admet aucun sous-groupe distingué autre que lui-même et le sous-groupe trivial : on dit que c'est un groupe *simple*. De nombreux exemples de groupes simples sont décrits dans [MPA1] (chapitre sur les groupes), liés aux groupes symétriques et aux groupes classiques. Soient u, v deux éléments de la sphère unité de E. Il existe alors un unique $\theta \in [0;\pi]$ tel que $\cos\theta = (u|v)$: on dira que θ est *l'angle (non orienté) de* u *et de* v.

Théorème 32. Le groupe $SO_3(\mathbb{R})$ est simple.

Démonstration. Soit G un sous-groupe distingué non trivial de $SO_3(\mathbb{R})$. Il s'agit de prouver que $G = SO_3(\mathbb{R})$. Puisque les renversements engendrent $SO_3(\mathbb{R})$, il suffit de prouver que G contient tous les renversements. Pour cela, il suffit de prouver qu'il en contient un : en effet, étant distingué, il contiendra tous ses conjugués ; or on a vu que les renversements sont tous conjugués entre eux.

Soit donc $R(\vec{\Delta},\theta)$ un élément non trivial de G : donc $0 < \theta \leqslant \pi$. Si $\theta = \pi$, c'est un renversement et l'on a terminé. Supposons donc $0 < \theta < \pi$. Grâce au corollaire 31 de la page 290, on peut de plus supposer que son axe est l'axe vertical de $\mathbb{R}^3$ orienté vers le haut. On a donc $A := \mathrm{Diag}\big(R(\theta),1\big) \in G$. Le point $U := (1,0,0)$ de l'équateur $S^2 \cap (\mathbb{R}^2 \times \{0\})$ est transformé en $V := (\cos\theta, \sin\theta, 0)$. Notons $W := (0,0,1)$ le pôle nord (qui est invariant par la rotation A). Pour $0 \leqslant \psi \leqslant \pi/2$, notons $U(\psi) := \cos\psi U + \sin\psi W$ et $V(\psi) := \cos\psi V + \sin\psi W$ les points de latitude ψ des méridiens de U et V. Ainsi, $AU(\psi) = V(\psi)$. Par ailleurs, $\big(U(\psi)|V(\psi)\big) = \cos^2\psi\cos\theta + \sin^2\psi$, qui prend toutes les valeurs de $\cos\theta$ à 1 lorsque ψ varie de 0 à $\pi/2$. L'angle de $U(\psi)$ et $V(\psi)$ prend donc toutes les valeurs comprises entre 0 et θ. En particulier, en choisissant n tel que $0 < \pi/n < \theta$, on conclut qu'il existe $u, v \in S^2$ d'angle π/n et tels que $Au = v$.

Soit $B \in SO_3(\mathbb{R})$ la rotation telle que $Bu = U_0 := U$ et $Bv = U_1 := (\cos\pi/n, \sin\pi/n, 0)$ (le lecteur vérifiera qu'il en existe une et une seule). Alors $C_1 := BAB^{-1} \in G$ (car celui-ci est distingué) et $C_1U_0 = U_1$ (petit calcul facile). Soit maintenant $B' \in SO_3(\mathbb{R})$ la rotation telle que $B'u = U_1$ et $B'v = U_2 := (\cos 2\pi/n, \sin 2\pi/n, 0)$. Alors $C_2 := B'AB'^{-1} \in G$ et $C_2U_1 = U_2$. En itérant ce procédé, on arrive à $C_n \in G$ telle que $C_nU_{n-1} = U_n := (\cos\pi, \sin\pi, 0) = -U$. La rotation $C := C_n \cdots C_1 \in G$ transforme U en $-U$. C'est donc un renversement. ∎

2.4.3 Quaternions

Rappelons que l'on a défini, dans l'exercice II.5.4 du module II.5, sur les nombres complexes de [L1], le corps non commutatif $\mathbf{H} := \{a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \mid a,b,c,d \in \mathbb{R}\}$ des quaternions. Nous utiliserons librement les propriétés établies dans cet exercice. Nous noterons ici $\overline{z} := a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}$ le *conjugué* du quaternion $z := a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ et $N(z) := z\overline{z} = \overline{z}z = a^2 + b^2 + c^2 + d^2$ la *norme algébrique* de z. Enfin, nous appellerons *partie réelle* de z le réel $\mathrm{Re}\, z := \dfrac{1}{2}(z + \overline{z}) = a$. Les *quaternions purs* sont ceux dont la partie réelle est 0, donc le sous–espace vectoriel $E := \{b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \mid b,c,d \in \mathbb{R}\}$ de

dimension 3 de H. Les faits élémentaires suivants sont alors faciles à prouver.

La norme algébrique est visiblement une forme quadratique définie positive sur $\mathbf{H}$ pour laquelle $(1, \mathbf{i}, \mathbf{j}, \mathbf{k})$ est une base orthonormée. Le produit scalaire correspondant est donné par la formule : $(z|z') := \mathrm{Re}(\overline{z}z')$ (il suffit de le vérifier pour la base indiquée).

Le groupe multiplicatif $\mathbf{H}^*$ de $\mathbf{H}$ opère sur $\mathbf{H}$ par conjugaison : $g.z := gzg^{-1}$. De l'égalité $N(gzg^{-1}) = N(g)N(z)N(g)^{-1}$, on déduit que $\mathbf{H}^*$ opère par automorphismes orthogonaux. Le noyau de cette action est $\mathbb{R}^*$. Le sous-espace $\mathbb{R}$ est fixe, son orthogonal E est donc invariant. Le sous-groupe $G := \{z \in \mathbf{H}^* \mid N(z) = 1\}$ des quaternions de norme 1 est distingué, et l'on a $\mathbf{H}^* = \mathbb{R}_+^* G$ (car tout quaternion s'écrit $z = \sqrt{N(z)}g$ avec $g \in G$). On a de plus $G \cap \mathbb{R}^* = \{+1, -1\}$.

Quaternions et $SO_3(\mathbb{R})$

Nous allons nous intéresser à l'action de G sur E. De ce qui précède, on déduit un morphisme $\rho : G \to \mathcal{O}(E)$ (donc une représentation orthogonale) de noyau $\{+1, -1\}$: pour tout $g \in G$, $\rho(g)$ est l'application $z \mapsto gzg^{-1}$. Le choix de la base $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ de E fait de celui-ci un espace vectoriel euclidien orienté de dimension 3 et permet d'identifier $\mathcal{O}(E)$ à $O_3(\mathbb{R})$.

Lemme 33. (i) Le groupe G est compact et connexe par arcs.
(ii) L'application ρ de G dans $\mathcal{O}(E)$ est continue.

Démonstration. (i) Remarquons d'abord que $\mathbf{H}$ est un $\mathbb{R}$espace vectoriel de dimension 4, donc muni d'une topologie définie par n'importe laquelle de ses normes (section 4). Dans la base $(1, \mathbf{i}, \mathbf{j}, \mathbf{k})$, G est défini par l'équation $a^2 + b^2 + c^2 + d^2 = 1$: c'est donc la sphère unité S^3 de $\mathbb{R}^4$, qui est évidemment fermée (définie par une équation polynomiale) et bornée (par définition) donc compacte. On peut la paramétrer : c'est l'image de l'application $(\varphi, \psi, \gamma) \mapsto (\cos\varphi\cos\psi, \cos\varphi\sin\psi, \sin\varphi\cos\gamma, \sin\varphi\sin\gamma)$ de $\mathbb{R}^3$ dans $\mathbb{R}^4$. Elle est donc connexe par arcs.
(ii) Les neuf coefficients de la matrice de $\rho(g)$ dans la base $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ de E sont les coordonnées dans cette base de $g\mathbf{i}g^{-1} = g\mathbf{i}\overline{g}$, $g\mathbf{j}g^{-1} = g\mathbf{i}\overline{g}$ et $g\mathbf{i}g^{-1} = g\mathbf{i}\overline{g}$. Ce sont des fonctions polynomiales, donc continues, des coordonnées de g dans la base $(1, \mathbf{i}, \mathbf{j}, \mathbf{k})$ de $\mathbf{H}$. ∎

Une conséquence de ce lemme est que l'image de G par ρ est incluse dans la composante connexe par arcs de $\mathcal{O}(E) \simeq O_3(\mathbb{R})$ qui contient $\rho(1) = \mathrm{Id}_E$, donc dans $\mathcal{SO}(E) \simeq SO_3(\mathbb{R})$.

Théorème 34.
Le morphisme $\rho : G \to \mathcal{SO}(E)$ est surjectif et son noyau est $\{+1, -1\}$.

Démonstration. On a déjà trouvé son noyau et prouvé que $\mathrm{Im}\,\rho \subset \mathcal{SO}(E)$. Soit $D := \mathbb{R}u$ une droite de E ; on peut supposer u normé. On a donc $u\overline{u} = 1$ (car u est normé) et

$\overline{u} = -u$ (car $u \in E$), donc $u^2 = -1$. Notons $s_u := \rho(u)$ (cela a un sens car $u \in G$). Alors $s_u(u) = uuu^{-1} = u$ et s_u fixe D. Comme $u^2 = -1 \Rightarrow s_u^2 = -\mathrm{Id}_E$ (car ρ est un morphisme), s_u est le renversement d'axe D. On voit donc que $\mathrm{Im}\,\rho$ contient tous les renversements, il est donc égal à $\mathcal{SO}(E)$.

Puisque $O_3(\mathbb{R}) \simeq SO_3(\mathbb{R}) \times \{+1, -1\}$, on aurait pu conjecturer que $G \simeq O_3(\mathbb{R})$: l'exercice qui suit montre qu'il n'en est rien !

Exercice 20.
Montrer que G n'admet pas de sous-groupe H tel que la restriction de ρ à H soit un isomorphisme sur $SO_3(\mathbb{R})$.

Solution. Si un tel H existait, on aurait, pour tout $g \in G$, $g \in H$ ou $-g \in H$ (voyez-vous pourquoi ?). Mais tout $u \in G \cap E$ vérifie $u^2 = (-u)^2 = -1$: on aurait donc $-1 \in H$. Mais $-1 \in \mathrm{Ker}\,\rho$ et l'on n'aurait pas un isomorphisme de H sur $SO_3(\mathbb{R})$.

Quaternions et $SU_2(\mathbb{C})$

En identifiant $\mathrm{i} \in \mathbb{C}$ à $\mathbf{i} \in \mathbf{H}$, on fait de $\mathbb{C}$ un sous-corps de $\mathbf{H}$, donc de ce dernier un $\mathbb{C}$-espace vectoriel de dimension 2. On vérifie que $\mathbb{C}.1 = \mathbb{R} + \mathbf{i}$, que $\mathbb{C}.\mathbf{j} = \mathbb{R}\mathbf{j} + \mathbb{R}\mathbf{k}$, et enfin que $\mathbb{C}.1 \oplus \mathbb{C}.\mathbf{j} = \mathbb{R} + \mathbf{i} + \mathbb{R}\mathbf{j} + \mathbb{R}\mathbf{k} = \mathbf{H}$. Ainsi $(1, \mathbf{j})$ est une base du $\mathbb{C}$-espace vectoriel $\mathbf{H}$. Identifiant un quaternion $z \in \mathbf{H}$ au couple (z_1, z_2) de complexes tel que $z = z_1 + z_2\mathbf{j}$ (ses coordonnées dans la $\mathbb{C}$-base $(1, \mathbf{j})$), on trouve la formule : $N(z) = |z_1|^2 + |z_2|^2$. La forme hermitienne $N(z)$ sur le $\mathbb{C}$-espace vectoriel $\mathbf{H}$ correspond ainsi à la forme hermitienne canonique sur $\mathbb{C}^2$, autrement dit, la $\mathbb{C}$-base $(1, \mathbf{j})$ est orthonormée. Pour tout $g \in G$ et pour tout $z \in H$, posons $r_g(z) := zg^{-1}$. De l'égalité $(\lambda z + \lambda' z')g^{-1} = \lambda z g^{-1} + \lambda' z' g^{-1}$, on déduit que r_g est $\mathbb{C}$-linéaire. L'application $z \mapsto gz$, elle, n'est pas $\mathbb{C}$-linéaire, car $g \in G$ ne commute pas en général avec $\lambda \in \mathbb{C}$!

Proposition 35. On définit ainsi une représentation unitaire de G dans l'espace hermitien $\mathbf{H}$. Son noyau est trivial. Son image est $\mathcal{SU}(\mathbf{H})$.

Démonstration. On a $r_g \circ r_h(z) = (zh^{-1})g^{-1} = z(gh)^{-1} = r_{gh}(z)$, et l'application $g \mapsto r_g$ est bien un morphisme de G dans $\mathcal{GL}(\mathbf{H})$. Comme $N(zg^{-1}) = N(z)$ (pour $g \in G$), cette représentation est unitaire. Un élément g du noyau vérifie $zg^{-1} = z$ pour tout z, c'est donc nécessairement $g := 1$. Les éléments de G sont les $z_1 + z_2\mathbf{j}$, où $z_1, z_2 \in \mathbb{C}$ sont tels que $|z_1|^2 + |z_2|^2 = 1$.
Par un petit calcul, le lecteur vérifiera que la matrice de r_g dans la base orthonormée $(1, \mathbf{j})$ est $\begin{pmatrix} \overline{z_1} & \overline{z_2} \\ -z_2 & z_1 \end{pmatrix}$. Ainsi, l'image de notre représentation dans $\mathcal{GL}(\mathbf{H}) \simeq GL_2(\mathbb{C})$ est le groupe $\left\{ \begin{pmatrix} \overline{z_1} & \overline{z_2} \\ -z_2 & z_1 \end{pmatrix} \mid z_1, z_2 \in \mathbb{C} \text{ et } |z_1|^2 + |z_2|^2 = 1 \right\}$, c'est-à-dire $SU_2(\mathbb{C})$.

Les groupes $SO_3(\mathbb{R})$ ***et*** $SU_2(\mathbb{C})$

On a établi un lien entre G et $SO_3(\mathbb{R})$ et un lien entre G et $SU_2(\mathbb{C})$. On peut en déduire l'existence d'un morphisme surjectif de $SU_2(\mathbb{C})$ sur $SO_3(\mathbb{R})$, de noyau $\{I_2, -I_2\}$. Nous allons construire directement ce morphisme.
Soit F le $\mathbb{R}$-espace vectoriel des matrices hermitiennes de trace nulle de $M_2(\mathbb{C})$. Ces matrices s'écrivent $\begin{pmatrix} a & b+\mathrm{i}c \\ b-\mathrm{i}c & -a \end{pmatrix}$, avec $a, b, c \in \mathbb{R}$. On a donc $\dim F = 3$, et une base particulière formée de $U := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, $V := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ et $W := \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}$.
Sur F, l'application $M \mapsto -\det M$ est une forme quadratique définie positive pour laquelle (U, V, W) est une base orthonormée : en effet, $\det(aU+bV+cW) = -a^2-b^2-c^2$. On fait ainsi de F un espace euclidien de dimension 3. On fait agir $SU_2(\mathbb{C})$ sur F en posant $u.M := uMu^{-1}$.

Proposition 36. On obtient ainsi une représentation orthogonale de $SU_2(\mathbb{C})$, de noyau $\{I_2, -I_2\}$ et d'image $\mathcal{SO}(F) \simeq SO_3(\mathbb{R})$.

Démonstration. Si M est hermitienne de trace nulle et si u est unitaire, on vérifie sans peine que uMu^{-1} est hermitienne de trace nulle. On a donc une représentation de $SU_2(\mathbb{C})$ dans $\mathcal{GL}(F)$ (il faut vérifier que $u \mapsto (M \mapsto uMu^{-1})$ est bien un morphisme de groupes). Puisque $\det(uMu^{-1}) = \det M$, c'est bien une représentation orthogonale. Un élément du noyau est une matrice unitaire qui commute avec U, V, W ; un rapide calcul montre qu'elle est scalaire et la condition sur le déterminant que c'est $\pm I_2$. Puisque $SU_2(\mathbb{C})$ est connexe par arcs, le même argument que précédemment montre que l'image de la représentation est dans $\mathcal{SO}(F) \simeq SO_3(\mathbb{R})$. Soit enfin D une droite de F et soit $v \in D$ de norme 1. Alors $u := \mathrm{i}v$ vérifie $u^* = -u$ (car v est hermitienne), $\operatorname{Tr} u = 0$ (car $\operatorname{Tr} v = 0$) et $\det u = 1$ (car $-\det v = 1$). Du théorème de Cayley-Hamilton (très élémentaire en dimension 2 !), on déduit que $u^2 = -I_2$, donc que $u^* = u^{-1}$, donc que u est unitaire. Comme $uvu^{-1} = v$ et $u^2 = -1$, on voit que l'image de u est le renversement d'axe D. Ainsi, l'image de la représentation contient tous les renversements, c'est bien $\mathcal{SO}(F) \simeq SO_3(\mathbb{R})$. ∎

2.4.4 Le groupe de Lorentz

On considère un $\mathbb{R}$-espace vectoriel orienté E de dimension 4, muni d'une forme quadratique q de signature $(3, 1)$: en particulier, la forme q et la forme bilinéaire symétrique φ qui lui est associée sont *non dégénérées*. Dans certaines bases privilégiées, les formes q et φ admettent donc pour matrice $J := \operatorname{Diag}(1, 1, 1, -1)$ (corollaire 45 de la page 132). Nous appellerons (abusivement) *base orthonormée* de telles bases. Si les vecteurs u, u' ont respectivement pour coordonnées (x, y, z, t) et (x', y', z', t') dans une base orthonormée et si l'on note X, X' les vecteurs colonnes associés, on a donc : $q(u) = x^2 + y^2 + z^2 - t^2 = {}^tXJX$ et $\varphi(u, u') = xx' + yy' + zz' - tt' = {}^tXJX'$. Nous noterons encore $q(x, y, z, t) := x^2 + y^2 + z^2 - t^2$ la forme quadratique correspondante sur $\mathbb{R}^4$. Exactement comme dans le cas des espaces euclidiens, on en déduit que le groupe orthogonal de la forme quadratique : $\mathrm{O}(q) := \{f \in \mathcal{GL}(E) \mid q \circ f = q\}$ peut être « réalisé » sous la forme matri-

cielle : $\mathrm{O}(q) \simeq O_{3,1}(\mathbb{R}) := \{A \in GL_4(\mathbb{R}) \mid {}^tAJA = J\}$. Ainsi, une matrice est q-orthogonale si, et seulement si, ses vecteurs colonnes forment une base q-orthonormée de $\mathbb{R}^4$. On retrouve les propriétés usuelles du groupe orthogonal $O_4(\mathbb{R})$: une matrice de passage entre bases orthonormées est orthogonale ; le groupe orthogonal agit simplement transitivement sur l'ensemble des bases orthonormées.
Notons également que la relation ${}^tAJA = J$ entraîne $\det A = \pm 1$, et l'on peut encore définir un sous-groupe spécial, appelé *groupe de Lorentz* : $\mathcal{SO}(q) := \mathcal{SL}(E) \cap \mathrm{O}(q)$. La matrice J étant orthogonale de déterminant -1, on voit que $\mathcal{SO}(q)$ est un sous-groupe d'indice 2 de $\mathrm{O}(q)$. Il agit simplement transitivement sur l'ensemble des bases orthonormées directes.

Définition 4. Le groupe $\mathcal{SO}(q)$ est appelé *groupe de Lorentz*. On appelle *repère inertiel* une base orthonormée directe. Un vecteur $u \in E$ est dit *de genre espace* si $q(u) > 0$; *de genre temps* si $q(u) < 0$; et *isotrope* si $q(u) = 0$. Les vecteurs isotropes forment le *cône de lumière*.

Proposition 37. Le groupe $\mathcal{SO}(q)$ agit transitivement sur $\{u \in E \mid q(u) = -1\}$.
Tout vecteur du genre temps a pour coordonnées $(0, 0, 0, t)$ dans un repère convenable.

Démonstration. Remarquons d'abord que, si $u \in E$ est du genre temps, la restriction de q à $\mathrm{Vect}(u)^\perp$ est définie positive. Soit en effet $u' \in \mathrm{Vect}(u)^\perp$ non nul. Après choix d'une base orthonormée, on peut écrire : $t^2 > x^2 + y^2 + z^2$ et $tt' = xx' + yy' + zz'$. On ne peut avoir $(x', y', z') = (0, 0, 0)$, car $t \neq 0$ entraînerait $t' = 0$, or $u' \neq 0$. On applique l'inégalité de Cauchy-Schwarz :

$$t^2 t'^2 = (xx' + yy' + zz')^2 \leqslant (x^2 + y^2 + z^2)(x'^2 + y'^2 + z'^2) < t^2(x'^2 + y'^2 + z'^2),$$

d'où $q(u') > 0$. Soient alors u_1, u_2 tels que $q(u_1) = q(u_2) = -1$. On munit respectivement $\mathrm{Vect}(u_1)^\perp$ et $\mathrm{Vect}(u_2)^\perp$ de bases orthonormées *au sens euclidien* : c'est possible d'après le théorème 4 de la page 145. On obtient ainsi deux bases orthonormées (i_1, j_1, k_1, u_1) et (i_2, j_2, k_2, u_2) de E, que l'on peut choisir de même orientation (quitte à permuter deux vecteurs). L'automorphisme f qui amène une base sur l'autre est dans $\mathcal{SO}(q)$ et $f(u_1) = u_2$.
Enfin, un vecteur quelconque du genre temps s'écrit tu, où $t > 0$ et $q(u) = -1$. ∎

Soit A la matrice de passage entre deux repères inertiels (i, j, k, u) et (i', j', k', u'). Si $u' = \pm u$, cette matrice est de la forme $A := \pm \mathrm{Diag}(B, 1)$, où $B \in SO_3(\mathbb{R})$. Sinon, l'intersection $\mathrm{Vect}(i, j, k) \cap \mathrm{Vect}(i', j', k')$ est un plan, et l'on peut, par des transformations spatiales, changer (i, j, k) et (i', j', k') en des familles orthonormées spatiales (i_1, j_1, k_1) et (i_1, j_1, k_2). On en conclut que A est de la forme $A := \pm \mathrm{Diag}(B, 1)\,\mathrm{Diag}(I_2, C)\,\mathrm{Diag}(B', 1)$, où $B, B' \in SO_3(\mathbb{R})$ et où $C \in SO_{1,1}(\mathbb{R})$. Ainsi, après choix d'un bon repère spatial dans chaque repère inertiel, la transformation de coordonnées opère sur une seule coordonnée spatiale. Les conséquences sont détaillées à l'exercice I.6.32 de la page 310.

3 Homographies

3.1 Le groupe projectif $\mathrm{P}GL_2(K)$ et la droite projective $\mathrm{P}^1(K)$

La carte $U := \{[x : y] \in \mathbf{P}^1(K) \mid y \neq 0\}$ de $\mathbf{P}^1(K)$ est mise en bijection avec la droite affine K par l'application $[x : y] \mapsto x/y$. L'unique élément de $\mathbf{P}^1(K) \setminus U$ est $[1 : 0]$, que

Un peu d'Histoire

Signification physique du groupe de Lorentz

Selon la Théorie de la Relativité restreinte, les événements physiques sont situés dans un *espace-temps* E qui est un espace affine réel de dimension 4. Chaque observateur décrit les lois de la physique relativement à un repère affine. Il lui faut, pour cela, choisir un événement origine $O \in E$; se servir d'instrument géométriques pour repérer la position des autres événements, donc choisir une base orthonormée (i, j, k) de l'espace vis-à-vis de laquelle il est immobile ; et d'une horloge, instrument de physique universel (c'est-à-dire comparable à ceux des autres observateurs) pour mesurer l'écoulement du temps. Tout cela est relatif à son *référentiel* (ville, train, planète, vaisseau spatial, galaxie…) La théorie postule que les lois de la physique s'expriment de la même manière dans tous les *référentiels inertiels*, lesquels se déplacent les uns vis-à-vis des autres selon un mouvement uniforme. L'une de ces lois physiques est la constance de la vitesse de la lumière. Nous convenons que tous les observateurs ont choisi le même événement comme origine. Ce n'est pas innocent, car l'invariance par translation permet, en physique, de démontrer la conservation de l'énergie et de l'impulsion. Le véritable groupe de la relativité restreinte est le *groupe de Poincaré*, qui est engendré par le groupe de Lorentz et les translations. Les physiciens déduisent ensuite de l'invariance de la vitesse c de la lumière que les coordonnées (x, y, z, t) et (x', y', z', t') d'un même événement dans deux repères inertiels sont reliés par l'égalité :

$$x^2 + y^2 + z^2 - c^2t^2 = x'^2 + y'^2 + z'^2 - c^2t'^2.$$

Nous supposons ici les unités choisies telles que $c := 1$. Un changement de coordonnées entre deux tels repères a donc la forme $X = AX'$, où $A \in SO_{3,1}(\mathbb{R})$. L'origine spatiale du second repère est formée de tous les points de la trajectoire $x' := 0, y' := 0, z' := 0, t' \in \mathbb{R}$. Ses coordonnées dans le premier repère sont donc de la forme $x(t'), y(t'), z(t'), t(t')$. Les trois premières coordonnées décrivent un mouvement, la quatrième le fait que l'écoulement du temps dépend du repère. Un intervalle de genre espace sépare deux événements qui peuvent être rendus simultanés par le choix d'un repère. Un intervalle isotrope relie deux événements qui peuvent être reliés par un rayon lumineux ; un intervalle de genre temps sépare deux événements distincts qui ont mêmes coordonnées spatiales dans un repère adéquat (cette affirmation est justifiée par la proposition 37). Le lecteur pourra se référer aux ouvrages [Efimov] et [Landau] pour comprendre comment les conceptions d'Albert Einstein sur le temps et l'espace ont été ainsi *géométrisées* par le mathématicien Hermann Minkowski.

nous nommerons *point à l'infini* et noterons ∞. On peut donc *identifier* la droite projective $\mathbf{P}^1(K)$ à la réunion $K \cup \{\infty\}$ de la droite affine K et d'un point à l'infini.

Ce processus est distinct de celui qui consiste à ajouter à $\mathbb{R}$ *deux* points à l'infini $+\infty$ et $-\infty$, pour construire la *droite achevée* $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$.

Pour étudier l'action du groupe projectif $\mathbf{P}GL_2(K)$ sur la droite projective $\mathbf{P}^1(K)$, on notera dorénavant $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ l'image dans $\mathbf{P}GL_2(K)$ de la matrice $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2(K)$. Puisque $\mathbf{P}GL_2(K)$ agit fidèlement sur $\mathbf{P}^1(K)$ (voir page 270), on pourra même identifier $g := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ à l'application de $\mathbf{P}^1(K)$ dans lui-même définie par :

$$g.[x : y] := [ax + by : cx + dy].$$

Pour décrire g comme homographie de $K \cup \{\infty\}$ on applique les formules de la page 271 :

$$\forall x \in K\ ,\ g.x := \begin{cases} \dfrac{ax+b}{cx+d} & \text{si} \quad cx + d \neq 0 \\ \infty & \text{si} \quad cx + d = 0 \end{cases} \quad \text{et} \quad g.\infty := \begin{cases} \dfrac{a}{c} & \text{si} \quad c \neq 0 \\ \infty & \text{si} \quad c = 0. \end{cases}$$

Nous conviendrons d'écrire $g.x = \dfrac{ax+b}{cx+d}$, avec des règles de calcul spéciales pour ∞.

Exercice 21.

Quel est le stabilisateur de ∞ dans $\mathbf{P}GL_2(K)$? Quel est, dans le groupe obtenu, le stabilisateur de 0 ? Quel est, dans ce nouveau groupe, le stabilisateur de 1 ?

Solution. Avec les notations précédentes, on a $g.\infty = \infty$ si, et seulement si, $c = 0$. Le stabilisateur de ∞ est donc le sous-groupe de $\mathbf{P}GL_2(K)$ formé des $\begin{bmatrix} a & b \\ 0 & d \end{bmatrix}$, autrement dit, des homographies de la forme $\dfrac{a}{d}x + \dfrac{b}{d}$, avec $ad \neq 0$. On reconnaît le groupe affine $\mathcal{G}A(K)$ (étendu naturellement à $K \cup \{\infty\}$ par la convention $\infty \mapsto \infty$).

On sait que le stabilisateur de 0 dans $\mathcal{G}A(K)$ est le sous-groupe $\mathcal{GL}(K) \simeq K^*$ formé des homographies de la forme $x \mapsto \lambda x$; et le stabilisateur de 1 dans $\mathcal{GL}(K) \simeq K^*$ est évidemment le groupe trivial $\{\mathrm{Id}\} \simeq \{1\}$.

Théorème 38. L'action induite de $\mathbf{P}GL_2(K)$ sur :

$$\{(\alpha, \beta, \gamma) \in \mathbf{P}^1(K)^3 \mid \alpha, \beta, \gamma \text{ sont deux à deux distincts}\}$$

est simplement transitive : on dit l'action de $\mathbf{P}GL_2(K)$ sur $\mathbf{P}^1(K)$ est *simplement* 3-*transitive*.

Démonstration. Il s'agit de démontrer que, pour deux triplets (α, β, γ) et $(\alpha', \beta', \gamma')$ quelconques de la forme ci-dessus, il existe une unique homographie g telle que $g.\alpha = \alpha'$, $g.\beta = \beta'$, $g.\gamma = \gamma'$; de manière équivalente, pour tout tel triplet (α, β, γ), il existe une unique homographie g telle que $g.0 = \alpha$, $g.1 = \beta$, $g.\infty = \gamma$. L'unicité

découle de l'exercice ci-dessus : si g et g' conviennent, $g^{-1}g'$ fixe 0, 1 et ∞ donc $g = g'$. L'existence se prouve en résolvant le système : $\dfrac{b}{d} = \alpha, \dfrac{a+b}{c+d} = \beta, \dfrac{a}{c} = \gamma$. Dans le cas où $\alpha, \beta, \gamma \in K$ (donc différents de ∞), on trouve $g := \begin{bmatrix} \gamma(\beta-\alpha) & \alpha(\gamma-\beta) \\ \beta-\alpha & \gamma-\beta \end{bmatrix}$, qui est bien la classe d'une matrice de $GL_2(K)$. Nous laissons au soin du lecteur le cas où α, β ou γ est ∞.

Corollaire 39. Le groupe $\mathbf{P}GL_2(K)$ est engendré par $\mathcal{G}A(K)$ et l'homographie $\dfrac{1}{x}$.

Démonstration. Soit $g \in \mathbf{P}GL_2(K)$ et soit $(\alpha, \beta, \gamma) := (g.0, g.1, g.\infty)$. Si $\gamma = \infty$, on sait que $g \in \mathcal{G}A(K)$. Si $\gamma \in K$, on peut transformer (α, β, γ) en un triplet de la forme $(\alpha', \beta', \infty)$ par l'homographie $\dfrac{1}{x-\gamma}$ qui est le composé de $h : x \mapsto x - \gamma$ et de $i : x \mapsto 1/x$; puis $(\alpha', \beta', \infty)$ en $(0, 1, \infty)$ par un élément j de $\mathcal{G}A(K)$. L'homographie $j \circ i \circ h \circ g$ fixe $(0, 1, \infty)$, elle est donc triviale et $g = (j \circ i \circ h)^{-1}$.

On peut également invoquer l'égalité $\dfrac{ax+b}{cx+d} = \dfrac{a}{c} + \dfrac{cb-ad}{c(cx+d)}$: il suffit donc de composer $x \mapsto c(cx+d)$, $x \mapsto 1/x$ et $x \mapsto (cb-ad)x + a/c$ pour obtenir notre homographie (le cas où $c = 0$ correspond à un élément de $\mathcal{G}A(K)$).

Corollaire 40. Le stabilisateur du sous-ensemble $\{0, 1, \infty\}$ de $\mathbf{P}^1(K)$ est un sous-groupe de $\mathbf{P}GL_2(K)$ isomorphe à $\mathfrak{S}_3$.

Démonstration.

Pour toute permutation σ de $\{0, 1, \infty\}$, il existe un unique $g_\sigma \in \mathbf{P}GL_2(K)$ dont la restriction à $\{0, 1, \infty\}$ coïncide avec σ. L'application $\sigma \mapsto g_\sigma$ est un isomorphisme de $\mathfrak{S}_3$ sur le stabilisateur de $\{0, 1, \infty\}$ dans $\mathbf{P}GL_2(K)$.

Exercice 22.

Décrire les homographies correspondant aux six permutations de $\{0, 1, \infty\}$.

Solution.

La permutation identité correspond à l'homographie $x \mapsto x$.
La transposition $0 \leftrightarrow \infty$ correspond à l'homographie $x \mapsto 1/x$.
La transposition $0 \leftrightarrow 1$ correspond à l'homographie $x \mapsto -x + 1$.
La transposition $1 \leftrightarrow \infty$ correspond à l'homographie $x \mapsto x/(x-1)$.
Le 3-cycle $0 \to 1 \to \infty \to 0$ correspond à l'homographie $x \mapsto 1/(1-x)$.
Le 3-cycle $0 \to \infty \to 1 \to 0$ correspond à l'homographie $x \mapsto 1 - 1/x$.

Le birapport

Il s'agit du principal invariant de la géométrie projective. On peut l'étudier pratiquement sans calcul.

Proposition et définition 41. Soient $\alpha, \beta, \gamma, \delta$ des points de $\mathbf{P}^1(K)$ tels que α, β, γ sont deux à deux distincts. L'orbite de $(\alpha, \beta, \gamma, \delta)$ sous l'action du groupe projectif $\mathbf{P}GL_2(K)$ contient un unique élément de la forme $(\infty, 0, 1, x)$. L'élément $x \in \mathbf{P}^1(K)$ est appelé *birapport* des points $\alpha, \beta, \gamma, \delta$ et noté $[\alpha, \beta, \gamma, \delta]$.

Démonstration. Il existe une unique homographie g telle que $g(\alpha) = \infty$, $g(\beta) = 0$ et $g(\gamma) = 1$. Le seul point de la forme $(\infty, 0, 1, x)$ dans l'orbite de $(\alpha, \beta, \gamma, \delta)$ est donc $\big(\infty, 0, 1, g(\delta)\big)$, et l'on a nécessairement $x := g(\delta)$. ∎

Nous allons en déduire trois séries de formules. D'abord, le calcul explicite du birapport :

$$[\alpha, \beta, \gamma, \alpha] = \infty, \quad [\alpha, \beta, \gamma, \beta] = 0, \quad [\alpha, \beta, \gamma, \gamma] = 1,$$
$$\text{et, si } \delta \notin \{\alpha, \beta, \gamma\} : [\alpha, \beta, \gamma, \delta] = \frac{\delta - \beta}{\delta - \alpha} \div \frac{\gamma - \beta}{\gamma - \alpha}.$$

Si $\delta = \alpha$, β ou γ, ces formules découlent tautologiquement de la définition. En général, l'homographie $h(x) := \dfrac{x - \beta}{x - \alpha}$ envoie α en ∞ et β en 0. L'homographie g est donc nécessairement donnée par $g(x) := h(x) \div h(\gamma)$.

Dans les cas particuliers suivants, on a les formules :

$$[\infty, 0, 1, \delta] = \delta, \qquad [0, \infty, 1, \delta] = 1/\delta, \qquad [\infty, 1, 0, \delta] = 1 - \delta,$$
$$[1, 0, \infty, \delta] = \delta/(\delta - 1), \qquad [0, 1, \infty, \delta] = 1/(1 - \delta), \qquad [1, \infty, 0, \delta] = 1 - 1/\delta.$$

Si $\{\alpha, \beta, \gamma\} = \{0, 1, \infty\}$ (sans précision sur l'ordre), l'homographie g de la proposition est l'une de celles déterminées dans l'exercice 22 de la page précédente, que l'on applique à δ.

Voici l'effet de permutations des arguments :

$$[\beta, \alpha, \gamma, \delta] = 1/\,[\alpha, \beta, \gamma, \delta]\,, \quad [\alpha, \gamma, \beta, \delta] = 1 - [\alpha, \beta, \gamma, \delta]$$
$$\text{et} \quad [\alpha, \beta, \delta, \gamma] = 1/\,[\alpha, \beta, \gamma, \delta]\,.$$

Les homographies qui interviennent sont ici $1/g$ et $1 - g$. Comme ces trois transpositions engendrent $\mathfrak{S}_4$, on obtient, pour chacune des vingt-quatre permutations des arguments, une formule où intervient l'une des six homographies de l'exercice 22 de la page précédente (voir l'exercice I.6.37 de la page 311).

Théorème 42. Une transformation bijective h de la droite projective est une homographie si, et seulement si, elle conserve le birapport, autrement dit, si :

$$\forall \alpha, \beta, \gamma, \delta \in \mathbf{P}^1(K) \text{ deux à deux distincts }, \ [h(\alpha), h(\beta), h(\gamma), h(\delta)] = [\alpha, \beta, \gamma, \delta]\,.$$

Démonstration. Si $(\infty, 0, 1, x)$ appartient à l'orbite de $(\alpha, \beta, \gamma, \delta)$, il appartient également à l'orbite de $\big(h(\alpha), h(\beta), h(\gamma), h(\delta)\big)$, d'où l'assertion directe. Soit réciproquement h une

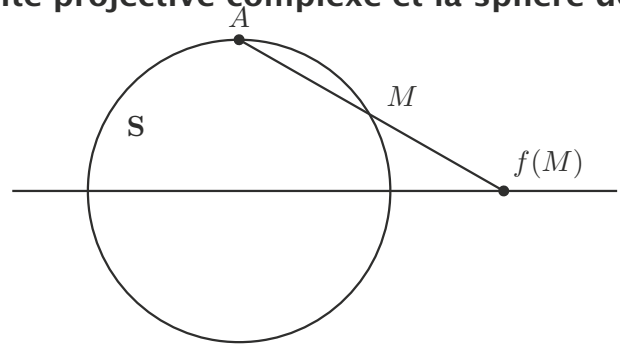
transformation bijective de $\mathbf{P}^1(K)$ qui conserve le birapport. Les formules :

$$\begin{aligned} h(x) &= [\infty, 0, 1, h(x)] \\ &= \left[h^{-1}(\infty), h^{-1}(0), h^{-1}(1), x\right] \\ &= \frac{x - h^{-1}(0)}{x - h^{-1}(\infty)} \div \frac{h^{-1}(1) - h^{-1}(0)}{h^{-1}(1) - h^{-1}(\infty)} \end{aligned}$$

montrent que h est une homographie. ∎

3.2 Le groupe spécial projectif réel et le demi-plan de Poincaré

3.2.1 La droite projective complexe et la sphère de Riemann



Un modèle géométrique simple permet de visualiser la droite projective $\mathbf{P}^1(\mathbb{C})$. Pour éviter toute confusion, nous noterons $z := x + \mathrm{i}y$ les points de $\mathbb{C}$ et $M := (X, Y, Z)$ ceux de $\mathbb{R}^3$. Notons $\mathbf{S} := S^2$ la sphère unité $\{(X, Y, Z) \in \mathbb{R}^3 \mid X^2 + Y^2 + Z^2 = 1\}$ de $\mathbb{R}^3$. Soit $A := (0, 0, 1)$ son pôle nord. À tout point $M \in \mathbf{S}$ distinct de A, on associe l'unique intersection $f(M)$ de la droite AM avec le plan équatorial $Z = 0$, que l'on identifie à $\mathbb{R}^2$, donc à $\mathbb{C}$. La figure représente une coupe de cette situation (par le plan vertical passant par A, M et $f(M)$). Cette application est donnée par la formule :

$$(X, Y, Z) \mapsto \frac{X}{1 - Z} + \mathrm{i}\frac{Y}{1 - Z}.$$

C'est une bijection de $\mathbf{S} \setminus \{A\}$ sur $\mathbb{C}$, dont la réciproque est donnée par la formule :

$$x + \mathrm{i}y \mapsto \left(\frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1}\right).$$

Au point A de $\mathbf{S}$, on associe le point à l'infini de $\mathbf{P}^1(\mathbb{C})$ (cela semble d'ailleurs naturel au vu des formules). On obtient ainsi une bijection de $\mathbf{S}$ sur $\mathbf{P}^1(\mathbb{C})$ que l'on baptise parfois, pour cette raison, *sphère de Riemann*. Les formules qui précèdent étant rationnelles, il est clair que l'on a un homéomorphisme de $\mathbf{S} \setminus \{A\}$ sur $\mathbb{C}$. On peut même dire que l'on a un homéomorphisme de $\mathbf{S}$ sur $\mathbf{P}^1(\mathbb{C})$, en étendant comme suit la notion de limite pour

prendre en compte le point à l'infini :

1. La bijection $f : \mathbf{S} \setminus \{A\} \to \mathbb{C}$ vérifie : $\lim_{M \to A} f(M) = \infty$, dans le sens suivant :

$$\forall R > 0\ ,\ \exists \varepsilon > 0\ :\ d(A, M) < \varepsilon \Longrightarrow |f(M)| > R,$$

où $d(A, M)$ désigne la distance euclidienne dans $\mathbb{R}^3$ et $|f(M)|$ le module dans $\mathbb{C}$.

2. Sa réciproque $g : \mathbb{C} \to \mathbf{S} \setminus \{A\}$ vérifie : $\lim_{z \to \infty} g(z) = A$, dans le sens suivant :

$$\forall \varepsilon > 0\ ,\ \exists\, R > 0\ :\ |z| > R \Longrightarrow d(A, g(z)) < \varepsilon.$$

Ces définitions apparemment *ad hoc* seront justifiées plus tard par la définition d'une « topologie » naturelle sur $\mathbf{P}^1(\mathbb{C})$.

Puisque $\mathbf{S}$ et $\mathbf{P}^1(\mathbb{C})$ sont homéomorphes. on doit pouvoir reconnaître les parties de l'un dans l'autre. Ainsi, l'équateur de $\mathbf{S}$ correspond au cercle unité $\mathbf{U}$ de $\mathbb{C}$; l'hémisphère nord de $\mathbf{S}$ (composante connexe du pôle nord dans $\mathbf{S}$ privée de l'équateur) correspond à l'extérieur de U (composante connexe de $\mathbf{P}^1(\mathbb{C}) \setminus \mathbf{U}$ qui contient le point à l'infini) ; et l'hémisphère sud de $\mathbf{S}$ correspond à l'intérieur de U, *i.e.* le disque unité ouvert $\overset{\circ}{D}(0, 1)$. Réciproquement, la droite projective réelle $\mathbf{P}^1(\mathbb{R}) = \mathbb{R} \cup \{\infty\} \subset \mathbf{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$ correspond au grand cercle de $\mathbf{S}$ défini par $y = 0$; le *demi-plan de Poincaré* $\mathcal{H} := \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$ correspond à l'hémisphère défini par $y > 0$ et le demi-plan opposé à l'hémisphère $y < 0$.

Exercice 23.
Quelles homographies de $\mathbf{P}GL_2(\mathbb{C})$ laissent invariant $\mathbf{P}^1(\mathbb{R}) \subset \mathbf{P}^1(\mathbb{C})$? Quelles homographies laissent invariant $\mathcal{H}$?

Solution. Puisqu'une homographie est déterminée par l'image de $0, 1, \infty \in \mathbf{P}^1(\mathbb{R})$, le groupe des homographies qui laissent invariant $\mathbf{P}^1(\mathbb{R})$ est $\mathbf{P}GL_2(\mathbb{R})$. Pour une telle homographie $g := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, on a vu (exercice I.2.5 de la page 89) que $\operatorname{Im} gz = \dfrac{ad - bc}{|cz + d|^2}$.
Le groupe des homographies qui laissent invariant $\mathcal{H}$ est $\mathbf{P}SL_2(\mathbb{R})$ (exercice 2 de la page 265).

Dans l'exercice II.5.14 du module sur les complexes de [L1], on a vu que l'homographie $g_0 : z \mapsto \dfrac{z - \mathrm{i}}{z + \mathrm{i}}$ est bijective de $\mathcal{H}$ sur $\overset{\circ}{D}(0, 1)$. On en déduit que le groupe des homographies qui laissent invariant $\overset{\circ}{D}(0, 1)$ est $g_0 \mathbf{P}SL_2(\mathbb{R}) g_0^{-1}$ (pour une description plus explicite, voir l'exercice I.6.40 de la page 311).

3.2.2 L'action de $\mathrm{P}SL_2(\mathbb{R})$ sur $\mathcal{H}$

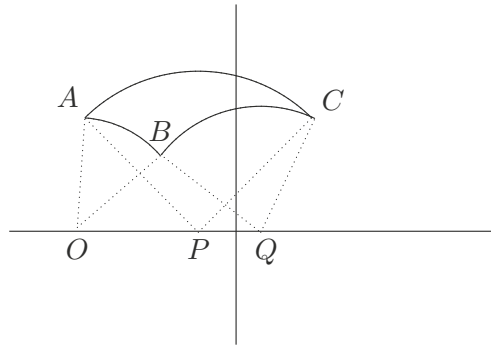
Le groupe $\mathbf{P}SL_2(\mathbb{R})$ est formé de toutes les homographies $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ de déterminant $ad - bc > 0$. Il contient en particulier le groupe $\mathcal{G}A_2^+(\mathbb{R})$ des applications affines $z \mapsto az + b$ telles que $a, b \in \mathbb{R}$ et $a > 0$ (prendre $b := 0$ et $d := 1$).

Théorème 43. (i) L'action de $\mathbf{P}SL_2(\mathbb{R})$ sur $\mathcal{H}$ est fidèle et transitive.
(ii) Le groupe $\mathbf{P}SL_2(\mathbb{R})$ est engendré par $\mathcal{G}A_2^+(\mathbb{R})$ et l'homographie $z \mapsto -1/z$.

Démonstration. (i) Tout élément z de $\mathcal{H}$ est l'image de i par une application affine du type ci-dessus, avec $b := \operatorname{Re} z$ et $a := \operatorname{Im} z$, d'où la transitivité. La fidélité découle du fait qu'une homographie qui fixe trois points est triviale.
(ii) On invoque la formule $\dfrac{az+b}{cz+d} = \dfrac{a}{c} + \dfrac{cb-ad}{c(cz+d)}$ de la preuve du corollaire 39 de la page 299. ∎

Notons $\mathbf{P}SO_2(\mathbb{R})$ l'image du groupe $SO_2(\mathbb{R}) \subset SL_2(\mathbb{R})$ dans $\mathbf{P}SL_2(\mathbb{R})$: c'est donc l'ensemble des homographies de la forme $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ $(\theta \in \mathbb{R})$.

Exercice 24.
Démontrer que les homographies de $\mathbf{P}SO_2(\mathbb{R})$ ont un point fixe commun dans $\mathcal{H}$.
Solution.
En écrivant la condition $gz = z$, on trouve que les points fixes de $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ dans $\mathbb{C}$ sont les racines de l'équation $\sin\theta(z^2+1) = 0$, donc, dans tous les cas, $\pm\mathrm{i}$ (si $\theta \in \pi\mathbb{Z}$, cette homographie est l'identité). Leur point fixe commun dans $\mathcal{H}$ est donc i. Il en découle que $\mathbf{P}SO_2(\mathbb{R})$ est inclus dans le stabilisateur de i. En fait, il y a même égalité (exercice I.2.5 de la page 89).



Nous appellerons *droite hyperbolique* du demi-plan de Poincaré d'une part, les demi-droites verticales ouvertes de la forme $D_\alpha := \alpha + \mathbb{R}_+^*\mathrm{i}$ $(\alpha \in \mathbb{R})$, d'autre part les demi-cercles ouverts centrés sur l'axe réel $C_{\alpha,r} := \{z \in \mathcal{H} \mid |z-\alpha| = r\}$. Cette terminologie sera justifiée à la section suivante, où nous verrons qu'il s'agit des « trajets de plus court chemin » (ou *géodésiques*) pour une certaine façon de calculer les distances. Sur la figure, on voit un *triangle hyperbolique*, dont les côtés AB, BC et CA sont des « segments » respectivement inclus dans des droites hyperboliques $C_{O,r}$, $C_{Q,r'}$ et $C_{P,r''}$.

Disons d'autre part que les *pieds* (sous-entendu : sur la frontière $\mathbf{P}^1(\mathbb{R})$) de D_α sont α et ∞, et que ceux de $C_{\alpha,r}$ sont $\alpha + r$ et $\alpha - r$.

Théorème 44. L'action de $\mathbf{P}SL_2(\mathbb{R})$ permute les droites hyperboliques et elle est transitive sur l'ensemble de ces droites.

Démonstration. L'élément $z \mapsto az+b$ de $\mathcal{G}A_2^+(\mathbb{R})$ transforme D_α en $D_{a\alpha+b}$ et $C_{\alpha,r}$ en $C_{a\alpha+b,ar}$. L'action de $\mathcal{G}A_2^+(\mathbb{R})$ permute donc transitivement les D_α d'une part, les $C_{\alpha,r}$ d'autre part. En vertu de l'assertion (ii) du théorème 43, il reste à voir l'effet de l'homographie $z \mapsto -1/z$. Celle-ci transforme D_0 en elle-même et toute autre D_α en un $C_{\beta,s}$ dont les pieds sont 0 et $-1/\alpha$; et elle transforme un $C_{\beta,s}$ dont les pieds sont 0 et α en $D_{-1/\alpha}$ et un $C_{\alpha,r}$ dont les pieds sont $\beta, \gamma \neq 0$ en un $C_{\alpha',r'}$ dont les pieds sont $-1/\beta, -1/\gamma \neq 0$. Il est alors évident que $\mathbf{P}SL_2(\mathbb{R})$ permute les droites hyperboliques. Il est également facile de voir que cette action est transitive. ∎

3.2.3 Géométrie hyperbolique

Nous allons maintenant définir la *géométrie hyperbolique* dans le demi-plan de Poincaré en *postulant* que la longueur d'un chemin $\gamma : [a,b] \to \mathcal{H}$ de classe $\mathcal{C}^1$ est donnée par la formule :

$$L(\gamma) := \int_a^b \frac{|\gamma'(t)|}{\operatorname{Im}\gamma(t)}\,\mathrm{d}t = \int_a^b \frac{\sqrt{x'^2(t)+y'^2(t)}}{y(t)}\,\mathrm{d}t,$$

où l'on a écrit sous forme cartésienne $\gamma(t) = x(t) + \mathrm{i}y(t)$. Notre but est de déterminer les *géodésiques* de $\mathcal{H}$, c'est-à-dire les chemins de longueur minimale joignant deux points donnés. Pour comprendre comment l'étude de la géométrie non euclidienne a conduit Poincaré à une telle définition, le lecteur ne peut rêver meilleur guide que le livre [Berger] (voir aussi [Efimov]). Le caractère géométrique de cette définition vient de ce qu'elle est invariante par changement de paramétrage, *i.e.* une application de classe $\mathcal{C}^1$ bijective $\varphi : [c,d] \to [a,b]$ telle que φ^{-1} est de classe $\mathcal{C}^1$:

Proposition 45. Avec ces notations, on a $L(\gamma \circ \varphi) = L(\gamma)$.

Démonstration. Les hypothèses sur φ entraînent que φ' ne s'annule pas, donc garde un signe constant. Si $\varphi' > 0$, on calcule :

$$L(\gamma\circ\varphi) = \int_c^d \frac{|(\gamma\circ\varphi)'(t)|}{\operatorname{Im}(\gamma\circ\varphi)(t)}\,\mathrm{d}t = \int_c^d \frac{\big|\gamma'\big((t)\big)\big|}{\operatorname{Im}(\gamma\circ\varphi)(t)}\varphi'(t)\,\mathrm{d}t = \int_a^b \frac{|\gamma'(t)|}{\operatorname{Im}\gamma(t)}\,\mathrm{d}t = L(\gamma),$$

par la formule de changement de variable dans une intégrale.
Si $\varphi' < 0$, le calcul est presque le même, à cela près que $|\varphi'(t)| = -\varphi'(t)$ et que les bornes sont permutées : $\varphi(c) = b$ et $\varphi(d) = b$. ∎

Exemples. Examinons le cas de segments de droites hyperboliques : Soient $\gamma_1(t) := \alpha + \mathrm{i}t$ sur $[a,b]$ (avec $0 < a_1 \leqslant b_1$) et $\gamma_2(t) := \alpha + r\cos t + r\mathrm{i}\sin t$ sur $[a_2,b_2]$ (avec $0 < a_2 \leqslant b_2 < \pi$). Alors :

$$L(\gamma_1) = \int_{a_1}^{b_1} \frac{1}{t}\,\mathrm{d}t = \log\frac{b_1}{a_1} \quad \text{et} \quad L(\gamma_2) = \int_{a_2}^{b_2} \frac{r}{r\sin t}\,\mathrm{d}t = \log\frac{\tan b_2/2}{\tan a_2/2}.$$

Théorème 46. Le groupe $\mathbf{P}SL_2(\mathbb{R})$ agit isométriquement sur $\mathcal{H}$ au sens de la géométrie hyperbolique ; autrement dit, pour tout chemin γ de classe $\mathcal{C}^1$ et pour toute homographie $g \in \mathbf{P}SL_2(\mathbb{R})$, on a : $L(\gamma) = L(g\circ\gamma)$.

Démonstration. En vertu de la proposition 45, on peut aussi bien supposer que l'intervalle de

définition de γ est $[0,1]$. En vertu de l'assertion (ii) du théorème 43 de la page 303, il suffit de vérifier l'égalité pour les homographies $z \mapsto az+b$ ($a,b \in \mathbb{R}$, $a>0$) et $z \mapsto -1/z$. Dans le premier cas :

$$L(g\circ\gamma) = \int_0^1 \frac{|(a\gamma+b)'(t)|}{\operatorname{Im}(a\gamma+b)(t)}\,\mathrm{d}t = \int_0^1 \frac{|a\gamma'(t)|}{a\operatorname{Im}\gamma(t)}\,\mathrm{d}t = L(\gamma).$$

Dans le second cas, grâce à l'identité $\operatorname{Im}(-1/z) = \operatorname{Im} z/|z|^2$:

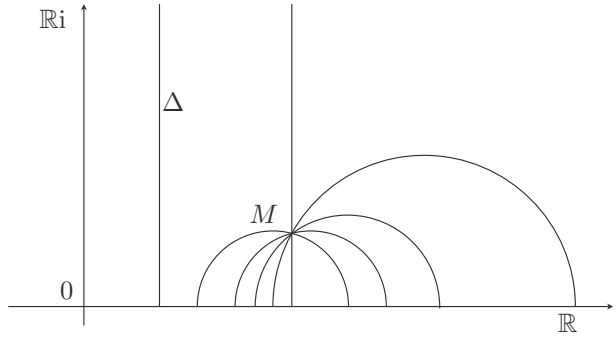
$$L(g\circ\gamma) = \int_0^1 \frac{|(-1/\gamma)'(t)|}{\operatorname{Im}(-1/\gamma)(t)}\,\mathrm{d}t = \int_0^1 \frac{\left|(\gamma'/\gamma^2)(t)\right|}{(\operatorname{Im}\gamma/|\gamma|^2)(t)}\,\mathrm{d}t = L(\gamma).$$ ■

Lemme 47. Toute orbite de $\mathbf{P}SO_2(\mathbb{R})$ dans $\mathcal{H}$ rencontre la droite $D_0 = \mathbb{R}_+^*\mathrm{i}$.

Démonstration. Pour tout $z \in \mathcal{H}$, la fonction continue $\theta \mapsto \operatorname{Re}\dfrac{\cos\theta z - \sin\theta}{\sin\theta z + \cos\theta}$ prend les valeurs de signes opposés $\operatorname{Re} z$ (pour $\theta := 0$) et $\operatorname{Re}(-1/z) = -(\operatorname{Re} z)/|z|^2$ (pour $\theta := \pi/2$) ; elle s'annule donc au moins pour une valeur de θ, définissant un point de l'orbite de z sur D_0. ■

Théorème 48. (i) Les géodésiques sont des segments de droites hyperboliques.
(ii) Par toute paire de points $z \neq z'$ il passe une unique droite hyperbolique.
(iii) La longueur du plus court chemin joignant z et z' est $|\log[\alpha,z,z',\beta]|$, où α et β sont les pieds de la droite hyperbolique joignant z et z'.

Démonstration. Chacune de ces affirmations est invariante par l'action de $\mathbf{P}SL_2(\mathbb{R})$. Or, combinant l'exercice 24 de la page 303 et le lemme 47, on voit que l'on peut amener (z,z') sur un couple $(\mathrm{i},\lambda\mathrm{i})$; et, dans ce cas, l'affirmation découle du premier exemple de la page 304. ■



Puisqu'une droite est traditionnellement définie comme « le plus court chemin entre deux quelconques de ses points » , nous étions fondés en définissant les droites hyperboliques. On peut aller plus loin et déclarer *parallèles* deux droites hyperboliques qui ne se rencontrent pas. Il est alors facile de vérifier qu'une droite hyperbolique Δ étant donnée, par tout point hors de Δ passent une infinité de parallèles à Δ (exercice I.6.43 de la page 311). Par exemple, sur la figure, il passe par le point M une infinité de parallèles à la droite Δ : une seule est de la forme D_α, toutes les autres sont de la forme $C_{\alpha,r}$. La géométrie hyperbolique est donc une géométrie *non euclidienne*, dans laquelle le célèbre *cinquième postulat d'Euclide* est transgressé.

Exercice type corrigé

I.6.0 (i) Nous allons décrire le groupe de Lorentz par une méthode proche de celle utilisée page 295 pour étudier $SO_3(\mathbb{R})$. Soit E le $\mathbb{R}$-espace vectoriel des matrices hermitiennes $M \in M_2(\mathbb{C})$. Montrer que $q(M) := -\det M$ est une forme quadratique non dégénérée de signature $(3,1)$ sur E. On a donc là un modèle de l'espace-temps relativiste décrit au paragraphe 2.4.4 de la page 295.
(ii) Montrer qu'en posant $g.M := gMg^*$, on définit une représentation ρ de $SL_2(\mathbb{C})$ dans E. Quel est son noyau ? Quel est l'effet de $\rho(g)$ sur q ? Démontrer que $\operatorname{Im}\rho$ est inclus dans le groupe de Lorentz.
(iii) Préciser le stabilisateur G_u de u et son action sur $\operatorname{Vect}(i,j,k)$. En déduire que son image par ρ est formée des matrices $\operatorname{Diag}(B,1)$, où $B \in SO_3(\mathbb{R})$. Qu'en déduire pour l'image et le noyau de ρ ?
(iv) Vérifier que $SL_2(\mathbb{C})$ agit transitivement sur les vecteurs normés de genre temps.
(v) Démontrer que le groupe de Lorentz est isomorphe à $\mathbf{P}SL_2(\mathbb{C})$.

Solution. (i) Toute matrice M de E s'écrit de manière unique : $M = xi + yj + zk + tu$ avec $x,y,z,t \in \mathbb{R}$, où $i := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $j := \begin{pmatrix} 0 & \mathrm{i} \\ -\mathrm{i} & 0 \end{pmatrix}$, $k := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ et $u := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. On a alors $q(M) = x^2 + y^2 + z^2 - t^2$ (c'est la forme quadratique du paragraphe 2.4.4 de la page 295), d'où une base orthonormée formée de i,j,k (genre espace) et de u (genre temps).
(ii) Si M est hermitienne, gMg^* l'est également, donc $\rho(g) \in \mathcal{L}(E)$. L'égalité $\rho(gh) = \rho(g)\rho(h)$ découle de $(gh)^* = h^*g^*$. En écrivant que $\rho(u)$ fixe les matrices i,j,k, on trouve qu'elle est scalaire. Le noyau de ρ est donc $\pm I_2 \subset SL_2(\mathbb{C})$. Pour $g \in SL_2(\mathbb{C})$, on a $\det gMg^* = \det M$. Ainsi, $\rho(g)$ préserve q et ρ est à valeurs dans $\mathcal{O}(q)$. Comme $SL_2(\mathbb{C})$ est connexe par arcs, la continuité du déterminant implique que ρ est à valeurs dans le groupe de Lorentz $\mathcal{SO}(q)$.
(iii) Les éléments de G_u sont les $g \in SL_2(\mathbb{C})$ tels que $gg^* = I_2$; on a donc $G_u = SU_2(\mathbb{C})$. Ce groupe laisse stable $\operatorname{Vect}(u)^\perp = \operatorname{Vect}(i,j,k)$ et agit sur cet espace *euclidien* par $M \mapsto gMg^{-1}$; il découle de la proposition 36 de la page 295 que le noyau de la restriction de ρ à G_u est $\{\pm I_2\}$ et que son image l'ensemble des $\operatorname{Diag}(B,1)$, où $B \in SO_3(\mathbb{R})$. On en déduit que $\operatorname{Im}\rho$ contient toutes ces matrices et que $\operatorname{Ker}\rho = \{\pm I_2\}$.
(iv) Les vecteurs normés de genre temps sont les $M \in E$ de déterminant 1. Du théorème 113 de la page 204, on déduit que, pour toute matrice hermitienne M de déterminant 1, il existe une matrice unitaire P telle que $PMP^* = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$, avec $\lambda \in \mathbb{R}^*$. Quitte à remplacer P par αP, avec $|\alpha| = 1$, on peut même prendre $P \in SU_2(\mathbb{C})$. D'autre part, l'action de $\begin{pmatrix} |\lambda|^{-1/2} & 0 \\ 0 & |\lambda|^{1/2} \end{pmatrix} \in SL_2(\mathbb{C})$ transforme $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$ en $\pm I_2 = \pm u$. Enfin, $g := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in SL_2(\mathbb{C})$ transforme $-u$ en $g(-u)g^* = u$.
(v) Il découle de la discussion qui suit la proposition 37 de la page 296 que $\operatorname{Im}\rho$ contient des générateurs du groupe de Lorentz. Comme $\operatorname{Ker}\rho = \{I_2, -I_2\}$, on a l'isomorphisme voulu.

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.6.1 $*$ (i) Soit H un hyperplan d'un K-espace vectoriel de dimension finie E.
Vérifier que $\Gamma(E,H) := \{f \in \mathcal{GL}(E) \mid f|_H = \mathrm{Id}_H\}$ est un sous-groupe de $\mathcal{GL}(E)$.
Pour tout $f \in \Gamma(E,H)$, on note $\overline{f}$ l'endomorphisme de E/H obtenu par passage au quotient (module I.1). Vérifier que $f \mapsto \overline{f}$ est un morphisme de groupes de $\Gamma(E,H)$ sur $\mathcal{GL}(E/H)$. Quel est le groupe $\mathcal{GL}(E/H)$?
(ii) Démontrer que le noyau $\Theta(E,H)$ de ce morphisme est formé des transvections d'hyperplan H.
Démontrer que le groupe $\Theta(E,H)$ est isomorphe au groupe additif H.
(iii) Montrer que les éléments de $\Gamma(E,H) \setminus \Theta(E,H)$ sont les dilatations d'hyperplan H.

I.6.2 Montrer que les transvections $T_{i,j}(\alpha)$ telles que $i < j$ engendrent le groupe $\mathbf{Tsu_n}(K)$ des matrices triangulaires supérieures unipotentes (*i.e.* dont la diagonale ne comporte que des 1).
Qu'en est-il du groupe $\mathbf{Tsr_n}(K)$ des matrices triangulaires supérieures régulières ?
Déterminer les centres des groupes $\mathbf{Tsr_n}(K)$ et $\mathbf{Tsu_n}(K)$.

I.6.3 $*$ Démontrer que les centres des groupes $\mathbf{P}GL_n(K)$ et $\mathbf{P}SL_n(K)$ sont triviaux.

I.6.4 $*$ (i) Démontrer que le groupe dérivé $D(G)$ est le plus petit sous-groupe distingué de G tel que le quotient soit abélien. En particulier, tout morphisme de G vers un groupe commutatif se factorise à travers le quotient $G/D(G)$.
(ii) À l'aide du théorème 8 de la page 265, en déduire que, pour $n \geqslant 3$, tout morphisme de $GL_n(K)$ dans K^* est de la forme $M \mapsto \varphi(\det M)$, où $\varphi : K^* \to K^*$ est un morphisme de groupes.

I.6.5 $*$ Montrer que les conclusions du théorème 8 de la page 265 demeurent si $n = 2$ et $K \neq \mathbf{F}_2$, $K \neq \mathbf{F}_3$. Lorsque $K = \mathbf{F}_2$, vérifier que $GL_n = SL_n$. Calculer le groupe dérivé de $GL_2(\mathbf{F}_2)$ et vérifier que ce n'est pas $SL_2(\mathbf{F}_2)$.

I.6.6 Vérifier que $GL_n(K) \times GL_p(K)$ agit sur $M_{n,p}(K)$ par $(P,Q).A := PAQ^{-1}$; deux matrices sont dans la même orbite si, et seulement si, elles sont équivalentes. Déduire du module II.4 de [L1] (section 2) que les orbites sous cette action sont les ensembles $\{A \in M_{n,p}(K) \mid \operatorname{rang} A = r\}$ pour $r \in [\![0, \min(n,p)]\!]$.

I.6.7 L'action de $\mathcal{SL}(E)$ sur $L_r(E)$ est-elle transitive ?

I.6.8 $*$ On suppose $\operatorname{card} K = q$. Calculer $\operatorname{card} L_r(E)$ par une méthode analogue à celle de la proposition 24 de la page 85 pour $\operatorname{card} GL_n(K)$. À l'aide de l'exercice 4 de la page 266, calculer le cardinal d'un stabilisateur. Déduire du théorème 14 de la page 72 que ce cardinal est égal à $\operatorname{card} \mathcal{GL}(E)/\operatorname{card} L_r(E)$. Vérifier la cohérence de ces résultats.

I.6.9 Décrire matriciellement (loi comprise) le stabilisateur d'un élément F de $S_r(E)$ sous l'action de $\mathcal{GL}(E)$. Décrire l'action de ce stabilisateur sur les espaces F, E/F et $F \times E/F$.
Quels sont les noyaux de ces actions ?

I.6.10 (i) À l'aide d'une base adaptée, montrer que le stabilisateur G d'un élément $\underline{F} := (F_1, \dots, F_k)$ de $D_d(E)$ est isomorphe au groupe des matrices de $GL_n(K)$ qui sont triangulaires supérieures par blocs, avec des blocs diagonaux inversibles de rangs $d_1, d_2 - d_1, \dots, d_k - d_{k-1}, n - d_k$. Décrire l'action de G sur le K-espace vectoriel « gradué » $\mathrm{gr}\underline{F} := F_1 \oplus (F_2/F_1) \oplus \cdots \oplus (F_k/F_{k-1}) \oplus (E/F_k)$. Quel est le noyau de cette action ?
(ii) Reprendre la question (i) dans le cas d'un drapeau maximal. Vérifier en particulier que les stabilisateurs s'identifient au groupe $\mathbf{Tsr_n}(K)$ des matrices triangulaires supérieures inversibles.

I.6.11 $*$
(i) Vérifier que, pour tout n non nul, $GL_n(\mathbf{F}_2) = \mathbf{P}GL_n(\mathbf{F}_2) = SL_n(\mathbf{F}_2) = \mathbf{P}SL_n(\mathbf{F}_2)$.
(ii) Vérifier que l'espace vectoriel $\mathbf{F}_2^2$ a 3 droites. Démontrer que le morphisme de groupes de $GL_2(\mathbf{F}_2)$ dans $\mathfrak{S}_3$ induit par l'action de $GL_2(\mathbf{F}_2)$ sur $\mathbf{P}^1(\mathbf{F}_2)$ est un isomorphisme.
(iii) Comparer le groupe dérivé de $GL_2(\mathbf{F}_2)$ à $SL_2(\mathbf{F}_2)$.

I.6.12
(i) Pour toute représentation ρ de G dans E, montrer qu'en posant $\rho'(g)(\varphi) := \varphi \circ \rho(g)^{-1}$, on définit une représentation de G dans le dual E^* de E (*représentation contragrédiente*).
(ii) Pour toute représentation de G dans E, montrer que $E^G := \{x \in E \mid \forall g \in G\ ,\ g.x = x\}$ est un sous-espace stable de E. Sous les hypothèses du théorème 11 de la page 269, montrer que $\dfrac{1}{N} \sum\limits_{g \in G} \rho(g)$ est un projecteur d'image E^G. En déduire, en caractéristique nulle, la formule : $\dim E^G = \dfrac{1}{N} \sum\limits_{g \in G} \operatorname{Tr} \rho(g)$.

I.6.13 $*$ Pour toute représentation ρ de G dans E, supposée *simple*, montrer que tout $\varphi \in \mathcal{L}(E)$ qui commute à l'action de G est nul ou inversible. Si l'on suppose de plus le corps de base algébriquement clos, montrer que φ est une homothétie (lemme de Schur).

I.6.14 (i) Soient X un espace affine de direction E et Y une partie de X. On dit que Y *engendre affinement* X si le plus petit sous-espace affine de X contenant Y est X lui-même. Montrer que, pour tout point $y_0 \in Y$, cette propriété équivaut à la suivante : les vecteurs $y - y_0$ $(y \in Y)$ engendrent E. Montrer qu'alors $\operatorname{card} Y \geqslant \dim E + 1$.
(ii) Montrer que le sous-groupe $\bigcap\limits_{y \in Y} \mathcal{G}A_y(X)$ (formé des transformations affines qui laissent invariant tout point de Y) est trivial si, et seulement si, Y engendre affinement X.

I.6.15 (i) Soient H l'hyperplan (affine) $K^n \times \{1\}$ de K^{n+1} (il est donc défini par l'équation $x_{n+1} = 1$) et G le sous-groupe de $GL_{n+1}(K)$ formé des automorphismes qui laissent H globalement invariant. Montrer que les éléments de G sont les matrices $\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix}$, où $A \in GL_n(K)$ et $B \in M_{n,1}(K)$. Calculer le produit de deux matrices de G.
(ii) Puisque G laisse H globalement invariant il agit sur K^n. Montrer que l'on obtient ainsi un isomorphisme de G sur $\mathcal{G}A(K^n)$.
(iii) Soient E un K-espace vectoriel de dimension finie et H un hyperplan de E. Montrer que $G := \{\varphi \in \mathcal{GL}(E) \mid \varphi(H) = H\}$ est isomorphe à $\mathcal{G}A(E)$.

I.6.16 (i) Soit $x \in \mathbf{P}^{n-1}(K)$. On suppose que x appartient aux cartes U_i et U_j et l'on note $p_i(x) = (y_1, \ldots, y_{n-1})$ et $p_j(x) = (z_1, \ldots, z_{n-1})$. Comment sont reliés ces deux jeux de coordonnées ?
(ii) Décrire l'effet de $[A]$ dans une carte U_i quelconque.

I.6.17 (i) Le groupe $SL_n(K)$ est-il ouvert ? Et le groupe $GL_n^+(\mathbb{R})$?
(ii) Le groupe $GL_n(K)$ est-il borné ?
(iii) Montrer que $GL_n(\mathbb{Q})$ est dense dans $GL_n(\mathbb{R})$, qui est fermé dans $GL_n(\mathbb{C})$ et que $SL_n(\mathbb{Q})$ est dense dans $SL_n(\mathbb{R})$, qui est fermé dans $SL_n(\mathbb{C})$.
(iv) Caractériser topologiquement (fermés ? ouverts ? denses ? bornés ? intérieur vide ? connexité ?) les sous-ensembles $\mathbf{Tsr_n}(K)$ et $\mathbf{Tsu_n}(K)$ de $T_n(K)$.

I.6.18 Dans la preuve de la proposition 14 de la page 273, justifier l'existence des a_i.

I.6.19
Étudier la différentiabilité et la différentielle de l'application $(A, B) \mapsto AB$ de $M_n(K) \times M_n(K)$ dans $M_n(K)$ et de l'application $A \mapsto A^2$ de $M_n(K)$ dans $M_n(K)$.

I.6.20 Soient $B \in M_n(K)$ et $D \in K[B]$ inversible. Montrer que $D^{-1} \in K[B]$.

Exercices sur la section 2

I.6.21 (i) Vérifier que $O_n(\mathbb{R})/SO_n(\mathbb{R}) \simeq \{+1, -1\}$ et que $U_n(\mathbb{C})/SU_n(\mathbb{C}) \simeq \mathbf{U}$.
(ii) On suppose n impair. Démontrer que $O_n(\mathbb{R})$ est isomorphe à $SO_n(\mathbb{R}) \times \{+1, -1\}$.
Déduire du théorème 23 de la page 283 que c'est faux si n est pair.

I.6.22 $*$ Démontrer que le centre de $U_n(\mathbb{C})$ est formé des matrices scalaires dont le rapport est dans $\mathbf{U}$. Quel est le centre de $SU_n(\mathbb{C})$? Le groupe $U_n(\mathbb{C})$ est-il isomorphe à $SU_n(\mathbb{C}) \times \mathbf{U}$?

I.6.23 Toute matrice de $M_n(\mathbb{C})$ s'écrit sous la forme QR, où $Q \in U_n(\mathbb{C})$ et R est triangulaire supérieure à éléments diagonaux réels positifs ou nuls ; pour une matrice inversible, cette décomposition est unique (décomposition QR complexe).

I.6.24 Déduire de la factorisation QR réelle la *décomposition d'Iwasawa* : toute matrice de $GL_n(\mathbb{R})$ s'écrit de manière unique sous la forme ODT, où $O \in O_n(\mathbb{R})$, D est diagonale à coefficients strictement positifs et $T \in \mathbf{Tsu_n}(\mathbb{R})$.

I.6.25 $*$ Démontrer la *décomposition polaire complexe* : toute matrice de $GL_n(\mathbb{C})$ s'écrit de manière unique sous la forme UH, où $U \in U_n(\mathbb{C})$ et $H \in H_n^{++}(\mathbb{C})$.

I.6.26 Soient a, b, c trois réels strictement positifs et E un espace euclidien. On note $T_{a,b,c}(E) := \{(u, v) \in E^2 \mid \|u\| = a, \|v\| = b, \|u + v\| = c\}$ (« triangles de côtés de longueurs a, b, c »). À quelle condition cet ensemble est-il non vide ? On suppose cette condition vérifiée. Montrer que $\mathrm{O}(E)$ agit transitivement sur $T_{a,b,c}(E)$ (*troisième cas d'égalité des triangles*). L'action de $\mathcal{SO}(E)$ est-elle transitive ?

I.6.27 (i) Donner une paramétrisation trigonométrique de la sphère unité $S^{n-1} \subset \mathbb{R}^n$.

(ii) Démontrer que $SU_2(\mathbb{C}) = \left\{ \begin{pmatrix} \lambda & \mu \\ -\overline{\mu} & \overline{\lambda} \end{pmatrix} \mid \lambda, \mu \in \mathbb{C} \text{ et } |\lambda|^2 + |\mu^2| = 1 \right\}$. En déduire un homéomorphisme de $SU_2(\mathbb{C})$ sur la sphère S^3 de $\mathbb{R}^4$. En paramétrant cette dernière, retrouver la connexité par arcs de $SU_2(\mathbb{C})$.

I.6.28 $**$ 1) Soit $(O_k, S_k)_{k\in\mathbb{N}}$ une suite d'éléments de $O_n(\mathbb{R}) \times S_n^{++}(\mathbb{R})$. On pose $A_k = O_k S_k$ et l'on suppose que $(A_k)_{k\in\mathbb{N}}$ converge vers $A \in GL_n(\mathbb{R})$. Soit $A = OS$ la décomposition polaire de A. Montrer que $(O_k)_{k\in\mathbb{N}}$ admet une sous-suite convergente et que toute sous-suite convergente de $(O_k)_{k\in\mathbb{N}}$ converge vers O.
2) En déduire que $(O_k)_{k\in\mathbb{N}}$ et $(S_k)_{k\in\mathbb{N}}$ convergent respectivement vers O et S. Vérifier que cela achève la preuve de la proposition 28 de la page 287.

I.6.29 Calculer l'exponentielle de $\begin{pmatrix} 0 & -a \\ a & 0 \end{pmatrix}$.

I.6.30 $*$ Démontrer que l'exponentielle est surjective de $\mathrm{o}_n(\mathbb{R})$ sur $SO_n(\mathbb{R})$, de $\mathrm{u}_n(\mathbb{C})$ sur $U_n(\mathbb{C})$ et de $\mathrm{su}_n(\mathbb{C})$ sur $SU_n(\mathbb{C})$.

I.6.31 $*$ (i) Démontrer que tout automorphisme de l'espace E (espace-temps de la relativité) qui préserve le cône de lumière est, à un scalaire près, q-orthogonal.
(ii) Démontrer que les vecteurs de type temps forment un cône C, c'est-à-dire une partie de E invariante par les homothéties non triviales. Montrer que les composantes connexes de C sont $C^+ := \{(x, y, z, t) \in \mathbb{R}^4 \mid t > \sqrt{x^2 + y^2 + z^2}\}$ et $C^- := -C^+$, qui sont convexes.
(iii) Soit $\gamma : [0, 1] \to E$ une courbe de classe $\mathcal{C}^1$ telle que, pour tout s, le vecteur vitesse $\gamma'(s)$ est dans C^+. Montrer que le vecteur $\overrightarrow{\gamma(0)\gamma(1)}$ est dans C^+.

I.6.32 $*$

(i) Montrer que tout élément de $SO_{1,1}(\mathbb{R})$ peut se mettre sous la forme $\pm \begin{pmatrix} \cosh a & \sinh a \\ \sinh a & \cosh a \end{pmatrix}$.

Vérifier que les éléments affectés du signe + forment un sous-groupe L d'indice 2, qui est la composante connexe de l'identité ; quelles sont les composantes connexes de $SO_{1,1}(\mathbb{R})$?

(ii) Écrire la transformation des coordonnées (z', t' en fonction de z, t) correspondant à un élément de L . On considère cette transformation comme reliant deux repères galiléens. Montrer que la le second repère se déplace le long de l'axe des z du premier à vitesse $U := \tanh a$. Si un point se déplace dans le second repère à vitesse V, autrement dit, si $z' = Vt'$, calculer sa vitesse dans le premier repère. Retrouver la loi relativiste d'addition des vitesses : $\dfrac{U + V}{1 + UV}$.

Exercices sur la section 3

I.6.33 À l'aide du théorème 4 de la page 263, démontrer que les homographies $x \mapsto x+1$, $x \mapsto 1/x$ et $x \mapsto \lambda x$ (où λ parcourt K^*) engendrent $\mathbf{P}GL_2(K)$.

I.6.34
Démontrer qu'une homographie non triviale de $\mathbf{P}GL_2(\mathbb{C})$ ne peut avoir qu'un ou deux points fixes. Dans le premier cas, elle est conjuguée (dans le groupe $\mathbf{P}GL_2(\mathbb{C})$) à l'homographie $x \mapsto x+1$. Dans le deuxième cas, elle est conjuguée à une homographie de la forme $x \mapsto qx$, où $q \notin \{0,1\}$.

I.6.35 (i) La droite projective $\mathbf{P}^1(K)$ a-t-elle nécessairement trois points distincts ?
(ii)A-t-elle nécessairement quatre points distincts ?

I.6.36 Décrire l'orbite d'un élément $x \in K$ par le groupe de six homographies de l'exercice 22 de la page 299.

I.6.37 Donner le formulaire complet des transformations du birapport selon les vingt-quatre permutations des arguments. Quelles permutations ne modifient pas le birapport ?

I.6.38 $*$ Déduire de l'exercice précédent et de l'exercice 22 de la page 299 un morphisme de groupes de $\mathfrak{S}_4$ dans $\mathfrak{S}_3$. Vérifier que ce morphisme est surjectif et déterminer son noyau.

I.6.39 $*$ Soit $g \in \mathbf{P}GL_2(\mathbb{C})$ une homographie qui laisse stable le demi-plan de Poincaré $\mathcal{H}$. Démontrer *topologiquement* que g laisse stable $\mathbf{P}^1(\mathbb{R})$.

I.6.40 Déduire de l'exercice 23 de la page 302 que le groupe des homographies qui laissent invariant $\overset{\circ}{D}(0,1)$ est $\left\{ \theta \, \dfrac{z-\alpha}{1-\overline{\alpha}z} \;\middle|\; \theta \in \mathbf{U}, \alpha \in \overset{\circ}{D}(0,1) \right\}$.

I.6.41 Déterminer les images dans $\overset{\circ}{D}(0,1)$ des droites hyperboliques de $\mathcal{H}$ par la bijection décrite à la suite de l'exercice 23 de la page 302.

I.6.42 (i) Notons $d(z,z')$ la longueur du plus court chemin joignant z et z' dans le « plan hyperbolique » $\mathcal{H}$. Montrer que c'est une distance sur $\mathcal{H}$.
(ii) Montrer que $\mathbf{P}SL_2(\mathbb{R})$ agit transitivement sur l'ensemble des couples (z,z') tels que $d(z,z') = \ell$ fixé.

I.6.43 Déterminer (en géométrie hyperbolique) toutes les parallèles à une droite Δ passant par un point $M \notin \Delta$.

Polynômes à plusieurs indéterminées

I.7

Dans tout ce module, K désigne un corps commutatif. Soit $A := K[X]$ l'anneau des polynômes en l'indéterminée X à coefficients dans K. Pour calculer avec des polynômes en une indéterminée *à coefficients dans* A, on ne peut évidemment pas employer à nouveau la lettre X. On notera donc $A[Y]$ (ou $A[Z]$, ou $A[T]$ ...) le nouvel anneau. On obtient ainsi une K*-algèbre* habituellement notée $K[X,Y]$ (ou $K[X,Z]$, ou $K[X,T]$ ...) des polynômes à *deux* indéterminées. Dans ce module, nous allons généraliser cette construction.

1 Calcul dans $K[X_1,\dots,X_n]$

1.1 Construction de $K[X_1,\dots,X_n]$

Soient $X_1,\dots,X_n$ des lettres appelées à désigner des *indéterminées* : leur rôle est analogue à celui de l'unique indéterminée X dans l'anneau des polynômes $K[X]$. En vertu de la section 2.2 de la page 22, on peut successivement construire les anneaux de polynômes $A_1 := K[X_1]$, $A_2 := A_1[X_2],\dots,A_n := A_{n-1}[X_n]$. Le dernier anneau ainsi obtenu est noté $K[X_1,\dots,X_n]$. Il est évidemment commutatif et K en est un sous-anneau : c'est donc une K-algèbre. On peut de même successivement construire les corps de fractions rationnelles $K_1 := K(X_1)$, $K_2 := K_1(X_2),\dots,K_n := K_{n-1}(X_n)$. Le dernier corps ainsi obtenu est noté $K(X_1,\dots,X_n)$. C'est encore une K-algèbre. Il paraît évident que ces constructions ne dépendent pas de l'ordre dans lequel on a introduit les indéterminées X_i. On a donc, pour tout $i \in [\![1,n]\!]$:

$$K[X_1,\dots,X_n] = K[X_1,\dots,\widehat{X_i},\dots,X_n][X_i],$$

où l'on a noté $K[X_1,\dots,\widehat{X_i},\dots,X_n] := K[X_1,\dots,X_{i-1},X_{i+1},\dots,X_n]$ la K-algèbre des polynômes en les indéterminées autres que X_i (ce que symbolise le chapeau sur $\widehat{X_i}$).

Théorème et définition 1.
(i) L'anneau $K[X_1,\dots,X_n]$ est appelée *anneau (ou K-algèbre) des polynômes en les n indéterminées* $X_1,\dots,X_n$. C'est un anneau intègre.
(ii) Le corps $K(X_1,\dots,X_n)$ est appelé *corps des fractions rationnelles en les n indéterminées* $X_1,\dots,X_n$. C'est le corps des fractions de $K[X_1,\dots,X_n]$.
(iii) Les éléments inversibles de l'anneau $K[X_1,\dots,X_n]$ sont les constantes non nulles : $K[X_1,\dots,X_n]^* = K^*$.

Démonstration. Chacune de ces affirmations se démontre par récurrence à l'aide de la proposition 41 de la page 24. ∎

Exemple. Le polynôme $\prod_{i=1}^{n}(T-X_i) \in K[X_1,\dots,X_n,T] = K[X_1,\dots,X_n][T]$ s'écrit :

$$\prod_{i=1}^{n}(T-X_i) = T^n - \sigma_1 T^{n-1} + \cdots + (-1)^n\sigma_n,$$

où $\sigma_1,\dots,\sigma_n \in K[X_1,\dots,X_n]$. Il y a une formule explicite pour les *polynômes symétriques élémentaires* : $\sigma_k = \sum_{1\leqslant i_1<\cdots<i_k\leqslant n} X_{i_1}\cdots X_{i_k}$ (même calcul que dans [L1] module II.6, section 1.4) ; par exemple, $\sigma_1 = \sum_{i=1}^{n} X_i$ et $\sigma_n = \prod_{i=1}^{n} X_i$.

Exercice 1.
Soit $\Delta_n := \prod_{1\leqslant i<j\leqslant n}(X_j - X_i) \in K[X_1,\dots,X_n]$ le *polynôme discriminant* en les indéterminées $X_1,\dots,X_n$ (en particulier, $\Delta_1 = 1$).
Exprimer $\prod_{\substack{1\leqslant i,j\leqslant n\\ i\neq j}}(X_i - X_j)$ en fonction de Δ_n.

Solution. On a :

$$\begin{aligned}\prod_{\substack{1\leqslant i,j\leqslant n\\ i\neq j}}(X_i - X_j) &= \prod_{1\leqslant i<j\leqslant n}(X_i - X_j)\prod_{1\leqslant j<i\leqslant n}(X_i - X_j)\\ &= \prod_{1\leqslant i<j\leqslant n}(X_j - X_i)\prod_{1\leqslant i<j\leqslant n}(X_i - X_j).\end{aligned}$$

Le dernier produit a $n(n-1)/2$ facteurs et vaut donc $(-1)^{n(n-1)/2}\Delta_n$, d'où l'égalité :

$$\prod_{\substack{1\leqslant i,j\leqslant n\\ i\neq j}}(X_i - X_j) = (-1)^{n(n-1)/2}\Delta_n^2.$$

Naturellement, on notera souvent $K[X,Y]$, $K[X,Y,Z]$ … des algèbres de polynômes en peu d'indéterminées. Par exemple :

$$\Delta_2(X,Y) = Y - X \qquad \text{ou} \qquad \Delta_3(X,Y,Z) = (Y-X)(Z-X)(Z-Y).$$

Mais, s'il y a « beaucoup » d'indéterminées, une gestion un peu lourde des indices est nécessaire. Les notations qui suivent sont adaptées au cas de n indéterminées indexées de 1 à n, mais elle pourront être étendues à toute famille finie d'indéterminées.

Exemple. Soit $A := K[(X_{i,j})_{1\leqslant i,j\leqslant n}]$ la K-algèbre des polynômes en les n^2 indéterminées $X_{1,1}, X_{1,2}, \dots, X_{n,n}$. La matrice $M := (X_{i,j})_{1\leqslant i,j\leqslant n} \in M_n(A)$ a pour déterminant un polynôme $\mathrm{Det}_n \in K[(X_{i,j})_{1\leqslant i,j\leqslant n}]$ en les $X_{i,j}$. Par exemple, pour $n = 2$, c'est le polynôme en 4 indéterminées $\mathrm{Det}_2 = X_{1,1}X_{2,2} - X_{1,2}X_{2,1} \in K[X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2}]$. De même, notons $c_i(M) \in K[(X_{i,j})_{1\leqslant i,j\leqslant n}]$ (pour $i = 1, \dots, n$) le coefficient de degré $n - i$ du polynôme caractéristique de M :

$$\chi_M(T) := \det(TI_n - M) = T^n + c_1(M)T^{n-1} + \cdots + c_n(M) \in K[(X_{i,j})_{1\leqslant i,j\leqslant n}][T].$$

Il aurait d'ailleurs été plus rigoureux d'écrire $c_{i,n}$, mais nous nous en tiendrons à cette notation allégée. Chaque c_i est encore un polynôme en les $X_{i,j}$. Par exemple, pour tout n, on a $c_1 = -(X_{1,1} + \cdots + X_{n,n})$ et $c_n = (-1)^n \mathrm{Det}_n$.

Définition 1. Les *monômes en les* n *indéterminées* $X_1, \dots, X_n$ sont les polynômes de la forme $X_1^{\alpha_1} \cdots X_n^{\alpha_n}$, où $(\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$.

Le monôme $X_1^{\alpha_1} \cdots X_n^{\alpha_n}$ sera fréquemment noté X^α, où :

1. la lettre X est *ici* une abréviation pour la famille d'indéterminées $(X_1, \dots, X_n)$ (il faut évidemment prendre garde au risque de confusion avec une indéterminée X) ;
2. la lettre α est une abréviation pour le *multiindice* $(\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$.

L'addition sur $\mathbb{N}$ s'étend naturellement en une addition sur $\mathbb{N}^n$ définie par :

$$(\alpha_1, \dots, \alpha_n) + (\beta_1, \dots, \beta_n) = (\alpha_1 + \beta_1, \dots, \alpha_n + \beta_n), \tag{1}$$

(loi produit), qui fait de $\mathbb{N}^n$ un monoïde commutatif (voir le paragraphe 1.2.1 de la page 8). La multiplication des monômes est alors donnée par la règle suivante :

$$\forall \alpha, \beta \in \mathbb{N}^n, X^\alpha X^\beta = X^{\alpha+\beta}. \tag{2}$$

Proposition 2.
La famille $(X^\alpha)_{\alpha\in\mathbb{N}^n}$ est une base du K-espace vectoriel $K[X_1, \dots, X_n]$.

Démonstration. On démontre par récurrence sur n que cette famille est libre et génératrice : nous laissons les détails au lecteur. ∎

En conséquence, tout polynôme admet une unique écriture $P = \sum p_\alpha X^\alpha$, où $(p_\alpha)_{\alpha\in\mathbb{N}^n}$ est une famille à support fini d'éléments de K, les *coefficients* du polynôme P. Les *termes* de P sont les $p_\alpha X^\alpha$ (ces termes sont parfois appelés abusivement monômes de P). Le *terme constant* de P est p_0, où l'on note ici 0 le multiindice $(0, \dots, 0) \in \mathbb{N}^n$. On a alors les règles d'addition et de multiplication suivantes :

$$\left(\sum p_\alpha X^\alpha\right) + \left(\sum q_\alpha X^\alpha\right) := \sum (p_\alpha + q_\alpha) X^\alpha, \quad \text{et}$$

$$\left(\sum p_\alpha X^\alpha\right)\left(\sum q_\alpha X^\alpha\right) := \sum r_\alpha X^\alpha, \text{ où } r_\alpha := \sum_{\beta+\gamma=\alpha} p_\beta q_\gamma.$$

La propriété universelle et ses conséquences

Soient $x_1, \dots, x_n$ des éléments d'une K-algèbre commutative A et soit $\alpha \in \mathbb{N}^n$ un multiindice. On peut étendre la notation condensée des monômes et poser : $x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. Plus généralement, pour tout polynôme $P := \sum p_\alpha X^\alpha \in K[X_1, \dots, X_n]$, on peut calculer *l'évaluation de* P *en* $(x_1, \dots, x_n) \in A^n$; c'est l'expression suivante dans l'algèbre A :

$$P(x) := P(x_1, \dots, x_n) := \sum p_\alpha x^\alpha.$$

Théorème 3 (Propriété universelle de $K[X_1, \dots, X_n]$).
Soit A une K-algèbre commutative et soient $x_1, \dots, x_n$ des éléments de A. Il existe alors un unique morphisme de K-algèbres $\varphi : K[X_1, \dots, X_n] \to A$ *associé à la substitution* $X_i \mapsto x_i$, c'est-à-dire tel que $\varphi(X_i) = x_i$ pour $i = 1, \dots, n$.

Démonstration. L'image de $P := \sum p_\alpha X^\alpha$ par le morphisme φ est nécessairement $\varphi(P) = P(x)$. Réciproquement, l'application $\varphi : P \mapsto P(x)$ satisfait aux propriétés requises : cela traduit la structure d'algèbre de $K[X_1, \dots, X_n]$. ∎

Exemple. Pour toute K-algèbre commutative A, il revient au même de se donner un morphisme $\varphi : K[(X_{i,j})_{1 \leqslant i,j \leqslant n}] \to A$ ou une famille $(m_{i,j})_{1 \leqslant i,j \leqslant n}$ d'éléments de A, c'est-à-dire une matrice $M \in M_n(A)$.
L'image dans A du polynôme Det_n (exemple de la page 315) est le déterminant $\det M \in A$. De même, les images $c_i(M)$ des polynômes c_i (*loc. cit.*) sont les coefficients du polynôme caractéristique $\chi_M(T) = T^n + c_1(M)T^{n-1} + \cdots + c_n(M) \in K[T]$.

Exercice 2.
Soit f une fonction n fois dérivable sur un ouvert U de $\mathbb{R}$ ou de $\mathbb{C}$ et à valeurs dans $K = \mathbb{R}$ ou $\mathbb{C}$. On appelle *relation différentielle algébrique (à coefficients constants) d'ordre* n *satisfaite par* f un polynôme $P \in K[X_0, \dots, X_n]$ tel que la fonction $P(f, f', \dots, f^{(n)})$ est identiquement nulle. En quoi est-ce une illustration du théorème ?

Solution. Prenons $A = \mathcal{F}(U, K)$ et $X_i \mapsto f^{(i)}$. Le noyau du morphisme obtenu est l'idéal des relations différentielles algébriques d'ordre n satisfaites par f.

Définition 2. Avec les notations du théorème, les éléments $x_1, \dots, x_n$ de A sont dits *algébriquement indépendants* si le morphisme φ est injectif, *algébriquement liés* dans le cas contraire. Un élément de $\mathrm{Ker}\,\varphi$ est appelé *relation algébrique entre les* x_i et l'idéal $\mathrm{Ker}\,\varphi$ est *l'idéal des relations algébriques entre les* x_i

Ainsi, l'élément unique x est algébriquement indépendant (resp. algébriquement lié) si, et seulement s'il est transcendant (resp. algébrique) (proposition 47 de la page 29).

Exemple. Dans la $\mathbb{R}$-algèbre $\mathcal{F}(\mathbb{R}, \mathbb{R})$, les fonctions $\cos$ et $\sin$ sont algébriquement liées, et $P := X^2 + Y^2 - 1$ est une relation algébrique entre elles, que l'on écrit plutôt $P(\cos, \sin) = 0$. Les fonctions $\exp$ et x (c'est-à-dire, bien sûr, $x \mapsto x$) sont algébriquement indépendantes. En effet, une relation algébrique non triviale serait un polynôme

non nul $P \in \mathbb{R}[X,Y]$ tel que $\forall x \in \mathbb{R}\ ,\ P(x,e^x) = 0$. Anticipant un peu sur le paragraphe 1.2.1 de la page 319, nous écrivons $P = \sum_{i=0}^{n} A_i(X)Y^i$, où les A_i sont des polynômes et $A_n \neq 0$. On sait alors que $P(x,e^x) \sim A_n(x)e^{nx}$ lorsque $x \to +\infty$, ce qui est incompatible avec la nullité.

Le théorème s'étend au cas d'une K-algèbre A non commutative, *à condition d'imposer que les x_i commutent entre eux :* $\forall i,j \in [\![1,n]\!]\ ,\ x_ix_j = x_jx_i$. Il est en effet nécessaire d'avoir l'égalité : $(x_1^{\alpha_1}\cdots x_n^{\alpha_n})(x_1^{\beta_1}\cdots x_n^{\beta_n}) = x_1^{\alpha_1+\beta_1}\cdots x_n^{\alpha_n+\beta_n}$; or, celle-ci n'est assurée que si les x_i commutent, comme le montre l'exemple suivant. Soient $M,N \in M_2(K)$. Il n'existe pas de morphisme φ de $K[X,Y]$ dans $A := M_2(K)$ tel que $\varphi(X) = M$ et $\varphi(Y) = N$ si par exemple $M := \begin{pmatrix}0&1\\0&0\end{pmatrix}$ et $N := \begin{pmatrix}0&0\\1&0\end{pmatrix}$. On aurait en effet $\varphi(XY) = MN = \begin{pmatrix}1&0\\0&0\end{pmatrix}$ et $\varphi(YX) = NM = \begin{pmatrix}0&0\\0&1\end{pmatrix}$, donc $\varphi(YX) \neq \varphi(XY)$, alors que $XY = YX$. En revanche, on peut toujours prendre $\varphi(X) = \varphi(Y) = M$ d'après le théorème, puisque M commute avec elle-même. Dans notre exemple, comme $M^2 = 0$, l'image du polynôme $P = a + bX + cY + \cdots$ (termes de degrés supérieurs) est $aI_2 + bM + cM = \begin{pmatrix}a&b+c\\0&a\end{pmatrix}$, les autres termes s'annulant.

Première conséquence : fonctions polynomiales. Prenons $A := K$. Pour tout polynôme $P \in K[X_1,\dots,X_n]$, on définit l'*application* ou *fonction polynomiale associée à* P : c'est l'application $\tilde{P} : x \mapsto P(x)$ de K^n dans K. L'application $P \mapsto \tilde{P}$ de $K[X_1,\dots,X_n]$ dans $\mathcal{F}(K^n,K)$ est un morphisme d'algèbres (le lecteur vérifiera que c'est le morphisme φ du théorème, lorsque l'on prend $A := \mathcal{F}(K^n,K)$, et pour x_i la $i^{ème}$ projection de K^n sur K). Les fonctions polynomiales seront étudiées en détail à la section 2, et l'on verra que, si K est infini, l'application $P \mapsto \tilde{P}$ est injective. Il n'y aura alors pas d'inconvénient à identifier le polynôme P à la fonction polynomiale associée, que nous noterons donc P.

On définit de même, pour toute fraction rationnelle R, la *fonction rationnelle* $\tilde{R} : x \mapsto R(x)$. On a posé $R(x) := \dfrac{P(x)}{Q(x)}$, où $\dfrac{P}{Q}$ est une représentation de R telle que $Q(x) \neq 0$; celle-ci est alors dite définie en x, et la valeur de $R(x)$ ne dépend pas du choix d'une telle représentation. Si de plus $P(x) = 0$, c'est un zéro de R. Si au contraire $P(x) \neq 0$, $Q(x) = 0$, c'est un pôle de R. Mais il y a aussi des *points d'indétermination* de R, en lesquels R n'est pas défini mais qui ne sont pas des pôles. Cela signifie que, pour toute représentation $R = \dfrac{P}{Q}$, on a $P(x) = Q(x) = 0$ (ce qui ne se produisait pas avec une seule indéterminée). Par exemple, $(0,0)$ est un point d'indétermination de $\dfrac{XY}{X^2+Y^2}$.

Soit E un K-espace vectoriel de dimension finie n. Pour chaque choix $\mathcal{B}$ d'une base de E, on a un isomorphisme $\varphi_{\mathcal{B}} : K^n \to E$. Disons qu'une application $f : E \to K$ est polynomiale si, et

seulement si, l'application $f \circ \varphi_{\mathcal{B}} : K^n \to K$ l'est. Cette définition ne dépend pas de la base $\mathcal{B}$ choisie. En effet, si $\mathcal{B}'$ est une autre base, les applications $g := f \circ \varphi_{\mathcal{B}}$ et $g' := f \circ \varphi_{\mathcal{B}'}$ sont telles que $g' = g \circ u$ et $g = g' \circ v$, où u et $v = u^{-1}$ sont des automorphismes linéaires de K^n : il est donc clair que g est polynomiale si, et seulement si, g' l'est.

Deuxième conséquence : substitution de polynômes.
Prenons $A := K[Y_1, \ldots, Y_p]$. On voit que, pour tout $P := \sum p_\alpha X^\alpha \in K[X_1, \ldots, X_n]$ et pour tout n-uplet $(Q_1, \ldots, Q_n)$ de polynômes en les indéterminées $Y_1, \ldots, Y_p$, la *substitution de $Q_1, \ldots, Q_n$ à $X_1, \ldots, X_n$* donne un polynôme :

$$P(Q_1, \ldots, Q_n) := \sum p_\alpha \, Q_1^{\alpha_1} \cdots Q_n^{\alpha_n} \in K[Y_1, \ldots, Y_p].$$

L'application $P \mapsto P(Q_1, \ldots, Q_n)$ est alors un morphisme de K-algèbres de $K[X_1, \ldots, X_n]$ dans $K[Y_1, \ldots, Y_p]$.
Si par exemple $A := K[X_1, \ldots, X_n]$, la donnée de $P_1, \ldots, P_n \in K[X_1, \ldots, X_n]$ définit un unique endomorphisme de la K-algèbre $K[X_1, \ldots, X_n]$ tel que $X_1 \mapsto P_1$,..., $X_n \mapsto P_n$. Un autre exemple est celui du morphisme :

$$\varphi : K[(X_{i,j})_{1 \leqslant i,j \leqslant n}] \to K[(Y_{i,j})_{1 \leqslant i,j \leqslant n}, (Z_{i,j})_{1 \leqslant i,j \leqslant n}]$$

tel que $X_{i,j} \mapsto \sum_{k=1}^{n} Y_{i,k} Z_{k,j}$. Alors l'image de Det_n est le produit $\mathrm{Det}_n(Y_{i,j}) \, \mathrm{Det}_n(Z_{i,j})$: c'est la formule $\det MN = \det M \, \det N$ de [L1], module II.7. Naturellement, la substitution est compatible avec l'évaluation : si $a_1, \ldots, a_p$ sont des éléments d'une K-algèbre B, et si l'on note $a = (a_1, \ldots, a_p)$, on peut calculer les $Q_i(a) \in B$, puis $P\big(Q_1(a), \ldots, Q_n(a)\big) \in B$, qui est évidemment égal à $P(Q_1, \ldots, Q_n)(a) \in B$.

1.2 Règles de calcul

Congruences dans $K[X_1, \ldots, X_n]$

Rappelons que, si I est un idéal d'un anneau commutatif A, on note $x \equiv y \pmod{I}$ (que l'on lit « x est congru à y modulo I ») si, et seulement si, $x - y \in I$.
Soient $x, y \in A^n$ tels que $\forall i \in [\![1, n]\!]$, $x_i \equiv y_i \pmod{I}$. Alors :

$$\forall P \in K[X_1, \ldots, X_n] \ , \ P(x) \equiv P(y) \pmod{I}.$$

Cela découle immédiatement du fait que $\equiv$ est une relation d'équivalence compatible avec l'addition et les multiplications interne et externe de A (module I.1). Énumérons quelques règles élémentaires.

PROPRIÉTÉ. *(i) Soit $a \in K^n$. Le morphisme de K-algèbres $P \mapsto P(a)$ de $K[X_1, \ldots, X_n]$ sur K est surjectif ; d'après le premier théorème d'isomorphisme pour les anneaux et l'équivalence : idéal maximal $\Leftrightarrow$ le quotient est un corps (module I.1), le noyau est un idéal maximal, qui contient évidemment l'idéal $\mathfrak{M}_a := <X_1 - a_1, \ldots, X_n - a_n>$. Comme $X_i \equiv a_i \pmod{\mathfrak{M}_a}$, on a $P \equiv P(a) \pmod{\mathfrak{M}_a}$ quel que soit P, donc $P \in \mathfrak{M}_a \Leftrightarrow P(a) \in \mathfrak{M}_a$. Mais un polynôme constant s'annule en a si, et seulement s'il est nul, donc $P \in \mathfrak{M}_a \Leftrightarrow P(a) = 0$.*

(ii) Soient $P,Q \in K[X_1,\dots,X_n]$ *tels que* $P(X_1,\dots,X_{i-1},Q,X_{i+1},\dots,X_n) = 0$. *Alors* P *est un multiple de* $X_i - Q$.
On a en effet : $P \equiv P(X_1,\dots,X_{i-1},Q,X_{i+1},\dots,X_n) \pmod I$, *où* $I = <X_i - Q>$. *La réciproque est vraie si* $Q \in K[X_1,\dots,\widehat{X_i},\dots,X_n]$ *: dans ce cas, si* $P = (X_i - Q)R$, *la substitution* $X_j \mapsto X_j$ $(j \neq i)$ *et* $X_i \mapsto Q$ *définit un endomorphisme* φ *de* $K[X_1,\dots,X_n]$ *tel que* $\varphi(X_i - Q) = 0$ *(parce que* Q *ne dépend pas de* X_i*), d'où :*

$$\varphi(P) = P(X_1,\dots,X_{i-1},Q,X_{i+1},\dots,X_n) = \varphi(X_i - Q)\varphi(R) = 0.$$

(iii) Pour tout $Q \in K[X_1,\dots,\widehat{X_i},\dots,X_n]$, *le morphisme :*

$$P \mapsto P(X_1,\dots,X_{i-1},Q,X_{i+1},\dots,X_n) \text{ de } K[X_1,\dots,X_n] \text{ sur } K[X_1,\dots,\widehat{X_i},\dots,X_n]$$

est surjectif et son noyau est l'idéal $<X_i - Q>$, *qui est donc premier d'après le premier théorème d'isomorphisme pour les anneaux et l'équivalence : idéal premier* $\Leftrightarrow$ *le quotient est intègre (module I.1). Par exemple, pour* $i \neq j$, *le polynôme* $X_i - X_j$ *est premier.*
(iv) Le noyau du morphisme surjectif $P \mapsto P(X,\dots,X)$ *de* $K[X_1,\dots,X_n]$ *sur* $K[X]$ *est l'idéal* $<(X_i - X_j)_{1\leqslant i,j\leqslant n}>$, *(puisque, notant* $I := <(X_i - X_j)_{1\leqslant i,j\leqslant n}>$, *on a* $P \equiv P(X_1,\dots,X_1) \pmod I$*) qui est donc premier.*
(v) Les polynômes $X_i - X_j$ *avec* $i < j$ *sont tous premiers et deux à deux non associés. Si un polynôme* P *est divisible par plusieurs* $X_i - X_j$ *distincts, il est donc divisible par leur produit. Cela vient du fait général suivant : si* $p_1,\dots,p_k$ *sont des éléments premiers deux à deux non associés de l'anneau intègre* A, *et si chacun des* p_i *divise* $a \in A$, *alors* $p_1 \cdots p_k$ *divise* a. *En effet, on peut écrire* $a = p_k b$. *Chaque* p_i $(i < k)$ *divise* $p_k b$, *donc divise* p_k *ou* b *(puisqu'il est premier) donc divise* b *(s'il divisait* p_k, *ils seraient associés). Ce processus s'itère facilement. Il est essentiel de noter que c'est bien la primalité des* p_i *et non seulement leur irréductibilité que l'on invoque ici.*

Exercice 3.
Déterminer toutes les relations différentielles algébriques satisfaites par e^x (*cf.* exercice 2).

Solution. Puisque $f := \exp = f' = \cdots = f^{(n)}$, une relation différentielle algébrique d'ordre n satisfaite par $\exp$ est de la forme : $\forall t \in \mathbb{R}\ ,\ P(e^t,\dots,e^t) = 0$. Le polynôme $P(X,\dots,X) \in \mathbb{R}[X]$ a une infinité de racines, donc il est nul. L'idéal des relations différentielles algébriques d'ordre n satisfaites par l'exponentielle réelle est donc l'idéal $<(X_i - X_j)_{0\leqslant i,j\leqslant n}>$ de $\mathbb{R}[X_0,\dots,X_n]$.

1.2.1 Degrés partiels, degré total

Soit $i \in [\![1,n]\!]$. Puisque $K[X_1,\dots,X_n] = K[X_1,\dots,\widehat{X_i},\dots,X_n][X_i]$, tout polynôme $P \in K[X_1,\dots,X_n]$ peut être considéré comme un polynôme en l'indéterminée X_i à coefficients dans l'anneau intègre $K[X_1,\dots,\widehat{X_i},\dots,X_n]$. On peut donc lui appliquer les définitions du paragraphe 2.2 de la page 22. En particulier, si $P \neq 0$, on l'écrit :

$$P = \sum_{k=0}^{d} P_k X_i^k, \quad \text{où les } P_k \in K[X_1,\dots,\widehat{X_i},\dots,X_n]\ ,\ P_d \neq 0.$$

On dit que l'on a *ordonné P suivant les puissances (ici croissantes) de X_i*. Le *degré (partiel) en X_i de P* est $\deg_{X_i} P := d$, son *coefficient dominant en X_i* est $\mathrm{cd}_{X_i} P := P_d$ et son *terme dominant en X_i* est $\mathrm{td}_{X_i} P := P_d X_i^d$. Par convention, $\deg_{X_i} 0 := -\infty$. L'anneau $K[X_1, \ldots, \widehat{X_i}, \ldots, X_n]$ étant intègre, les règles de calcul usuelles sont valides.

Exemple. Le polynôme Det_n est du premier degré en $X_{n,n}$ et son coefficient dominant est Det_{n-1} : cela se voit en le développant par rapport à la dernière ligne ou à la dernière colonne. (Et par rapport à un $X_{i,j}$ quelconque ?)

Exercice 4.

Reconnaître et factoriser l'image $V_n(X_1, \ldots, X_n)$ de Det_n dans $K[X_1, \ldots, X_n]$ par le morphisme associé à la substitution $X_{i,j} \mapsto X_j^{i-1}$.

Solution. Il s'agit bien sûr du déterminant de Vandermonde :

$$\det \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_1 & X_2 & \ldots & X_n \\ \vdots & \vdots & \ddots & \vdots \\ X_1^{n-1} & X_2^{n-1} & \ldots & X_n^{n-1} \end{pmatrix}.$$

Nous allons retrouver par récurrence la formule (déjà rencontrée) : $V_n = \prod_{1 \leqslant i < j \leqslant n} (X_j - X_i)$.

Pour $n = 1$ ou 2, c'est immédiat. Supposons le vrai au rang $n-1$. En développant par rapport à la dernière ligne ou la dernière colonne, on voit que V_n est de degré $n-1$ et de coefficient dominant $V_{n-1}(X_1, \ldots, X_{n-1})$ en X_n (qui est non nul par hypothèse). De plus, V_n s'annule lorsque l'on remplace X_i (où $i < n$) par X_n ; il est donc divisible par chaque $X_n - X_i$, donc par $\prod_{i=1}^{n-1} (X_n - X_i)$ (*cf.* la dernière propriété de la page 318). On a donc nécessairement : $V_n = V_{n-1} \prod_{i=1}^{n-1} (X_n - X_i)$, d'où la conclusion.

L'égalité $K[X_1, \ldots, X_n] = K[X_1, \ldots, \widehat{X_i}, \ldots, X_n][X_i]$ permet encore d'appliquer les résultats du paragraphe 2.2 sur la division euclidienne (proposition 43 de la page 24).

Exercice 5.

Déterminer l'idéal des relations algébriques entre $\cos$ et $\sin$.

Solution. On a déjà vu (exemple de la page 316) que cet idéal de $\mathbb{R}[X, Y]$ contient $X^2 + Y^2 - 1$. Soit $P \in \mathbb{R}[X, Y]$ tel que $P(\cos, \sin) = 0$. La division euclidienne par rapport à Y permet d'écrire $P = (X^2 + Y^2 - 1)Q + R$, où $\deg_Y(R) < 2$, donc $R = A(X)Y + B(X)$. On a évidemment $R(\cos, \sin) = 0$, c'est-à-dire :

$$\forall x \in \mathbb{R}\ ,\ A(\cos x) + \sin x B(\cos x) = 0.$$

Remplaçant x par $-x$, on voit que l'on a aussi identiquement $A(\cos x) - \sin x B(\cos x) = 0$. Prenant $x \in \mathbb{R} \setminus \pi\mathbb{Z}$, on voit que A et B s'annulent sur $]-1, 1[$, donc $A = B = 0$, donc $R = 0$. L'idéal recherché est donc $<X^2 + Y^2 - 1>$.

Appelons *poids* du multiindice $\alpha := (\alpha_1,\dots,\alpha_n)$ l'entier $|\alpha| := \alpha_1 + \cdots + \alpha_n$. On peut alors définir le *degré total*, ou simplement *degré*, du monôme X^α comme le poids de son exposant α, puis le *degré total* ou *degré* d'un polynôme non nul $P \in K[X_1,\dots,X_n]$ comme le plus grand des poids de ses monômes :

$$\forall \alpha \in \mathbb{N}^n\ ,\ \deg X^\alpha := |\alpha| \quad \text{et} \quad \forall P := \sum p_\alpha X^\alpha \in K[X_1,\dots,X_n]\ ,\ \deg P := \max_{p_\alpha \neq 0} |\alpha|.$$

Par convention, le degré du polynôme nul est $-\infty$. Les règles suivantes étendent celles sur les polynômes en une indéterminée : $\deg(P+Q) \leqslant \max(\deg P, \deg Q)$ et $\deg(PQ) = \deg P + \deg Q$. Les polynômes de degré 0 sont les constantes non nulles, autrement dit, les éléments inversibles de $K[X_1,\dots,X_n]$ (théorème 1 de la page 314). On en déduit immédiatement les faits suivants :

1. Tout polynôme de degré 1 de $K[X_1,\dots,X_n]$ est irréductible.
2. Si un polynôme P non nul divise Q, alors $\deg P \leqslant \deg Q$ ou $Q = 0$.
3. On peut définir sans ambiguïté le degré d'une fraction rationnelle $F = \dfrac{P}{Q}$ en posant $\deg F = \deg P - \deg Q$.

En revanche, comme plusieurs monômes X^α de P peuvent avoir pour degré $|\alpha| = \deg P$, il n'est pas possible de parler de terme dominant ou de coefficient dominant : examiner par exemple le cas de $X^2 - Y^2 \in K[X,Y]$.

On définit également la *valuation* (ou *ordre*) d'un polynôme non nul $P \in K[X_1,\dots,X_n]$ comme le plus petit des poids de ses monômes ; si $P = \sum p_\alpha X^\alpha$: $\omega(P) := \min_{p_\alpha \neq 0} |\alpha|$.
Par convention, l'ordre du polynôme nul est $+\infty$. Les règles suivantes étendent celles sur les polynômes en une indéterminée : $\omega(P+Q) \geqslant \min\big(\omega(P),\omega(Q)\big)$ et $\omega(PQ) = \omega(P) + \omega(Q)$. Les polynômes de valuation non nulle sont les polynômes sans terme constant, c'est-à-dire les éléments de l'idéal maximal $\mathfrak{M}_0$ (défini page 318). Lorsque $n := 1$, les définitions du degré et de l'ordre données ici coïncident avec celles données dans le module sur les polynômes de [L1].
Tout polynôme non constant de $K[X_1,\dots,X_n]$ est produit de polynômes irréductibles. On le montre facilement par récurrence forte sur le degré. Si $\deg P = 1$, on l'a déjà remarqué. Sinon, soit P est irréductible, soit $P = QR$ avec $\deg Q, \deg R < \deg P$, donc Q et R sont produits de polynômes irréductibles, donc P l'est également.

Théorème 4. L'anneau $K[X_1,\dots,X_n]$ est *factoriel*, c'est-à-dire :
(i) Tout polynôme irréductible de $K[X_1,\dots,X_n]$ est premier (lemme d'Euclide).
(ii) La décomposition d'un polynôme en facteurs irréductibles est essentiellement unique : si $P = P_1 \dots P_r = Q_1 \dots Q_s$ sont deux telles décompositions, alors $r = s$ et, quitte à renuméroter les indices, P_i et Q_i sont associés pour $i \in [\![1, r]\!]$.
(iii) Si A et B sont étrangers (*i.e.* n'ont pas de diviseur commun non inversible) et si $A|BC$, alors $A|C$ (lemme de Gauß).

Démonstration. Une fois l'existence d'une décomposition en irréductibles acquise, il est très facile de voir que chacune des trois propriétés entraîne les deux autres. Elles découlent,

par récurrence sur n, du théorème suivant, qui sera démontré dans le chapitre d'arithmétique du [L2-L3] : si A est un anneau factoriel (dans le sens ci-dessus) alors $A[X]$ l'est encore. L'initialisation de la récurrence peut se faire pour $n = 1$ sur la seule base des propriétés démontrées dans le module sur les polynômes de [L1]. ∎

Toute fraction rationnelle $F \in K(X_1, \ldots, X_n)$ peut donc s'écrire sous la forme $F = \dfrac{P}{Q}$ où $P, Q \in K[X_1, \ldots, X_n]$ sont étrangers. Une telle écriture est appelée *forme irréductible* de la fraction rationnelle F. Si $\dfrac{P}{Q}$ est une telle forme irréductible, alors, pour toute écriture $F = \dfrac{P_1}{Q_1}$, il existe $R \in K[X_1, \ldots, X_n]$ tel que $P_1 = RP$ et $Q_1 = RQ$. En particulier, si $\dfrac{P_1}{Q_1}$ est également une forme irréductible de F, alors R est constant.

1.2.2 Polynômes homogènes

Définition 3. Un polynôme $P \in K[X_1, \ldots, X_n]$ est dit *homogène de degré* k, (on dit aussi *forme de degré* k) s'il est non nul et si $\omega(P) = \deg P = k$, *i.e.* si tous ses monômes sont de degré k.

Une forme de degré 1 n'est donc pas autre chose qu'une forme linéaire en les X_i.

Exemple. Le polynôme Det_n (exemple de la page 315) est homogène de degré n.

Dans la suite de ce module, nous noterons S la K-algèbre $K[X_1, \ldots, X_n]$ et, pour tout entier k, S_k le sous–espace vectoriel de S engendré par les monômes X^α de poids $|\alpha| = k$. Les éléments de S_k sont donc 0 et les polynômes homogènes de degré k. Les S_k sont les *composantes homogènes* de S.

Proposition 5. L'espace vectoriel S_k est de dimension $\dbinom{n+k-1}{n-1}$.

Démonstration. Il admet pour base la famille des X^α de poids $|\alpha| = k$, dont le nombre d'éléments est le nombre de solutions dans $\mathbb{N}^n$ de l'équation $\alpha_1 + \cdots + \alpha_n = k$. C'est le « nombre de manières de répartir k objets entre n boîtes », qui a été calculé en exercice dans la section « Dénombrement » du module d'arithmétique de [L1]. ∎

Puisque l'ensemble de tous les monômes X^α est une base de S, on a une décomposition :

$$S = \bigoplus_{k \in \mathbb{N}} S_k.$$

Définition 4. Notons p_k les projecteurs associés à cette décomposition. La *composante homogène de degré* k d'un polynôme $P := \sum p_\alpha X^\alpha \in S$ est le polynôme :

$$p_k(P) = \sum_{|\alpha|=k} p_\alpha X^\alpha \in S_k.$$

La *forme initiale* d'un polynôme non nul P est sa composante homogène de plus bas degré : $\mathrm{In}(P) := p_k(P)$, où $k := \omega(P)$.

Ainsi, P est la somme de ses composantes homogènes. Outre la linéarité, la propriété suivante de p_k est la plus utile :

$$\forall P, Q \in K[X_1, \ldots, X_n]\ ,\ \forall k \in \mathbb{N}\ ,\ p_k(PQ) = \sum_{i+j=k} p_i(P) p_j(Q).$$

Elle repose sur le fait que le produit de deux polynômes homogènes de degrés k et ℓ et un polynôme homogène de degré $k + \ell$. Par exemple, si P et Q sont non nuls, $\mathrm{In}(PQ) = \mathrm{In}(P)\,\mathrm{In}(Q)$; et, si P est homogène de degré d, $p_k(PQ) = P\ p_{k-d}(Q)$ (ou 0 si $d > k$).

Exercice 6.

Soit $t \in K^*$. Quel est l'effet sur les composantes homogènes de l'automorphisme de $K[X_1, \ldots, X_n]$ associé à la substitution $X_i \mapsto tX_i$?

Solution. Il s'agit bien d'un automorphisme puisque la substitution $X_i \mapsto t^{-1}X_i$ définit son inverse. Notons $P(tX)$ l'image de $P(X) \in K[X_1, \ldots, X_n]$. Il est d'abord clair que $(tX)^\alpha = t^{|\alpha|}X^\alpha$. On en déduit que, si $P \in S_k$, alors $P(tX) = t^k P(X)$; plus généralement, les composantes homogènes de $P(tX)$ sont les $t^k p_k(P)$.

Homogénéisation. On peut établir un lien entre polynômes non nuls quelconques en n indéterminées et polynômes homogènes en $(n+1)$ indéterminées de la manière suivante. Soit $P \in K[X_1, \ldots, X_n]$ un polynôme non nul de degré d. Alors le polynôme :

$$P^h := \sum p_\alpha X^\alpha X_0^{d-|\alpha|} \in K[X_1, \ldots, X_n][X_0] = X_0^{\deg P} P\left(\frac{X_1}{X_0}, \ldots, \frac{X_n}{X_0}\right)$$

est bien homogène de degré d en les $(n+1)$ indéterminées $X_0, \ldots, X_n$. Si P est homogène, on a bien sûr $P^h = P$. Pour $P, Q \in K[X_1, \ldots, X_n]$ non nuls quelconques, on a $(PQ)^h = P^h Q^h$.

On retrouve le polynôme P en remarquant que c'est $P^h(1, X_1, \ldots, X_n)$.

Proposition 6. On suppose le corps K algébriquement clos. Alors toute forme de degré d en deux indéterminées est produit de d formes linéaires.

Démonstration. Écrivons $F := \sum_{i=0}^{d} a_i X^i Y^{d-i}$. Soit e le degré en X de F : donc $a_e \neq 0$ et $a_i = 0$ pour $e < i \leqslant d$. On a alors $F(X, Y) = Y^d P(X/Y)$, où $P(X) := \sum_{i=0}^{e} a_i X^i$ est un polynôme en une indéterminée de degré e. Le corps K étant algébriquement clos, on peut écrire : $P = a_e \prod_{i=1}^{e} (X - \lambda_i)$ (les $\lambda_i \in K$). Alors : $F = a_e Y^{d-e} \prod_{i=1}^{e} (X - \lambda_i Y)$. ∎

1.3 Dérivations

1.3.1 Dérivations de la K-algèbre $K[X_1, \ldots, X_n]$

Les dérivées partielles ont été introduites dans le module sur les fonctions de plusieurs variables de [L1]. Une bonne partie de leurs propriétés repose sur la règle de Leibniz, dont les conséquences algébriques (en ce qui concerne les polynômes) peuvent être étudiées sur un corps de base quelconque.

Définition 5. On appelle *dérivation* de la K-algèbre $K[X_1, \ldots, X_n]$ une application K-linéaire de $K[X_1, \ldots, X_n]$ dans lui même vérifiant la *règle de Leibniz* :

$$\forall P, Q \in K[X_1, \ldots, X_n]\ ,\ \Delta(PQ) = P\Delta(Q) + Q\Delta(P). \tag{3}$$

Il est facile de vérifier que si $P_1, \ldots, P_k \in K[X_1, \ldots, X_n]$ et si $\Delta_1, \ldots \Delta_k$ sont des dérivations, alors l'application $P \mapsto \sum_{i=1}^{k} P_i \Delta_i(P)$ est une dérivation, que l'on notera $\sum_{i=1}^{k} P_i \Delta_i$.

Une dérivation Δ de la K-algèbre $K[X_1, \ldots, X_n]$ admet une unique extension K-linéaire à $K(X_1, \ldots, X_n)$ vérifiant la règle de Leibniz. Remplaçant P par $\dfrac{P}{Q}$ dans (3), on voit en effet que l'on est tenu de poser $\Delta\left(\dfrac{P}{Q}\right) = \dfrac{Q\Delta(P) - P\Delta(Q)}{Q^2}$, d'où l'unicité.

Réciproquement, le lecteur prouvera que la formule ci-dessus est bien définie, *i.e.* $\Delta(F)$ ne dépend que de F et non de l'écriture particulière $F = \dfrac{P}{Q}$ choisie, et qu'elle répond aux conditions posées.

Il sera souvent commode d'utiliser la dérivée logarithmique $\dfrac{\Delta(P)}{P}$ $(P \neq 0)$; on voit en effet que, pour $P, Q \neq 0$, on a l'égalité suivante dans $K(X_1, \ldots, X_n)$:

$$\frac{\Delta(PQ)}{PQ} = \frac{\Delta(P)}{P} + \frac{\Delta(Q)}{Q}\cdot \tag{4}$$

Par récurrence, on en tire les formules plus générales :

$$\frac{\Delta(F_1 \cdots F_k)}{F_1 \cdots F_k} = \frac{\Delta(F_1)}{F_1} + \cdots + \frac{\Delta(F_k)}{F_k}, \quad \text{et} \quad \Delta(F_1 \cdots F_k) = \sum_{i=1}^{k} F_1 \cdots \Delta(F_i) \cdots F_k,$$

où, dans le $i^{\text{ème}}$ terme, seul le facteur F_i a été remplacé par $\Delta(F_i)$.

Lemme 7. Soient $P_1, \ldots, P_n \in K[X_1, \ldots, X_n]$. Il existe une unique dérivation Δ de la K-algèbre $K[X_1, \ldots, X_n]$ telle que $\Delta(X_i) = P_i$ pour $i \in [\![1, n]\!]$.

Démonstration. Soit Δ une telle application. L'application itérée de la règle (4) montre que l'on doit avoir : $\dfrac{\Delta(X_1^{\alpha_1} \cdots X_n^{\alpha_n})}{X_1^{\alpha_1} \cdots X_n^{\alpha_n}} = \sum_{i=1}^{n} \alpha_i \dfrac{\Delta(X_i)}{X_i} = \sum_{i=1}^{n} \alpha_i \dfrac{P_i}{X_i}\cdot$

On en tire la formule $\Delta(X_1^{\alpha_1} \cdots X_n^{\alpha_n}) = \sum_{i=1}^{n} \alpha_i X_1^{\alpha_1} \cdots X_i^{\alpha_i - 1} \cdots X_n^{\alpha_n} P_i$.

Notons que, dans cette somme, les monômes comportant un exposant négatif $\alpha_i - 1$ apparaissent toujours avec un facteur scalaire α_i nul : on a donc bien un polynôme. Ce calcul assure l'unicité de l'application Δ.
Réciproquement, définissons l'application linéaire Δ par son effet sur la base des monômes par la formule trouvée. On vérifie d'abord aisément l'égalité (3) dans le cas où P et Q sont des monômes ; puis on en déduit le cas général par bilinéarité. ∎

Théorème et définition 8. Soit $i \in [\![1, n]\!]$. Il existe une unique dérivation D_i de $K[X_1, \ldots, X_n]$ telle que $D_i(X_j) = \delta_{i,j}$ pour $j \in [\![1, n]\!]$. On l'appelle $i^{\text{ème}}$ *dérivée*, ou encore *dérivée partielle par rapport à* X_i, et l'on note $D_i(P) = \dfrac{\partial P}{\partial X_i}$.

Démonstration. On applique le lemme à $P_j := \delta_{i,j}$. ∎

On a alors : $\forall i \in [\![1, n]\!]\ ,\ D_i(X^\alpha) = \alpha_i X_1^{\alpha_1} \cdots X_i^{\alpha_i - 1} \cdots X_n$. L'indice i étant fixé, les polynômes $P \in K[X_1, \ldots, \widehat{X_i}, \ldots, X_n]$ (c'est-à-dire ceux où ne figure pas X_i) se comportent comme vis à vis de X_i comme des constantes : pour un tel polynôme, on a $D_i P = 0$ et, par application de la règle de Leibniz, $D_i(PQ) = PD_i(Q)$, quel que soit $Q \in K[X_1, \ldots, X_n]$.

Corollaire 9. Si Δ est une dérivation de $K[X_1, \ldots, X_n]$, on a la formule :

$$\forall P \in K[X_1, \ldots, X_n]\ ,\ \Delta(P) = \sum_{i=1}^{n} D_i(P)\Delta(X_i).$$

Formule d'Euler. Considérons la dérivation $\Delta := \sum_{i=1}^{n} X_i D_i$.
De l'égalité $X_i D_i(X^\alpha) = \alpha_i X^\alpha$, on tire $\Delta(X^\alpha) = |\alpha| X^\alpha$. Ainsi, les polynômes homogènes de degré k sont des vecteurs propres de Δ correspondant à la valeur propre k :

$$\forall P \in K[X_1, \ldots, X_n]\ ,\ P \in S_k \Longrightarrow \sum_{i=1}^{n} X_i \frac{\partial P}{\partial X_i} = kP \quad \text{(formule d'Euler).}$$

La réciproque est d'ailleurs valable si le corps K est de caractéristique 0 : on a alors $S_k = \operatorname{Ker}(\Delta - k \operatorname{Id}_S)$ (notations de la page 322).

Exercice 7.
Calculer $\dfrac{\partial \operatorname{Det}_n}{\partial X_{i,j}}$.

Solution. Selon les notations du module sur la dimension finie de [L1], soit $D_{i,j}$ le mineur d'ordre $n-1$ correspondant ; c'est un polynôme où ne figure pas $X_{i,j}$, et il en est de même de $\operatorname{Det}_n - (-1)^{i+j} X_{i,j} D_{i,j}$. On a donc : $\dfrac{\partial \operatorname{Det}_n}{\partial X_{i,j}} = (-1)^{i+j} D_{i,j}$.

Les formules de composition des différentielles de l'analyse ont leur équivalent algébrique (voir également à ce sujet les exercices I.7.20 de la page 348 et I.7.21 de la page 348).

Proposition 10. Soient $P \in K[X_1, \dots, X_n]$ et $Q_1, \dots, Q_n \in K[Y_1, \dots, Y_p]$. Alors :

$$\forall j \in [\![1, p]\!] \ , \ \frac{\partial P\big(Q_1, \dots, Q_n\big)}{\partial Y_j} = \sum_{i=1}^{n} \frac{\partial P}{\partial X_i}\big(Q_1, \dots, Q_n\big) \, \frac{\partial Q_i}{\partial Y_j}.$$

Démonstration. Comme $P(Q_1, \dots, Q_n) \in K[Y_1, \dots Y_p]$, cette formule a bien un sens. Par K-linéarité en P des deux membres, il suffit de la démontrer quand $P = X^\alpha$, autrement dit, de prouver : $\dfrac{\partial\big(Q_1^{\alpha_1} \cdots Q_k^{\alpha_k}\big)}{\partial Y_j} = \sum\limits_{i=1}^{n} \alpha_i \, Q_1^{\alpha_1} \cdots Q_i^{\alpha_{i-1}} \cdots Q_k^{\alpha_k} \, \dfrac{\partial Q_i}{\partial Y_j}$; mais cette formule s'obtient facilement par dérivation logarithmique. ∎

Exemple. Soit $(a_1, \dots, a_n) \in K^n$. Prenant $p = n$, $Y_i = X_i$ et $Q_i = X_i + a_i$ ($i \in [\![1, n]\!]$), on voit que : $\forall i \in [\![1, n]\!] \ , \ \dfrac{\partial\big(P(X_1 + a_1, \dots, X_n + a_n)\big)}{\partial X_i} = \dfrac{\partial P}{\partial X_i}(X_1 + a_1, \dots, X_n + a_n)$.

De même : $\dfrac{\partial\big(P(TX_1, \dots, TX_n)\big)}{\partial T} = \sum\limits_{i=1}^{n} X_i \dfrac{\partial P}{\partial X_i}(TX_1, \dots, TX_n)$.

1.3.2 Dérivées itérées et formules de Taylor

Les endomorphismes $D_1, \dots, D_n$ du K-espace vectoriel $K[X_1, \dots, X_n]$ commutent deux à deux. On vérifie en effet que :

$$\forall i \neq j \in [\![1, n]\!] \ , \ \forall \alpha \in \mathbb{N}^n \ , \ D_i \circ D_j(X^\alpha) = D_j \circ D_i(X^\alpha) = \frac{\alpha_i \alpha_j}{X_i X_j} X^\alpha = \alpha_i \alpha_j X^{\alpha - E_i - E_j}.$$

Dans cette dernière égalité, on a introduit une notation qui sera en vigueur dans tout ce paragraphe : E_i désigne le n-uplet $(0, \dots, 1, \dots, 0) \in \mathbb{N}^n$ (le 1 est en $i^{\text{ème}}$ position). On peut également vérifier l'égalité :

$$\forall i \in [\![1, n]\!] \ , \ \forall \alpha \in \mathbb{N}^n \ , \ D_i \circ D_i(X^\alpha) = \frac{\alpha_i(\alpha_i - 1)}{X_i^2} X^\alpha = \alpha_i(\alpha_i - 1) X^{\alpha - 2E_i}.$$

On peut alors définir les *dérivées itérées* en composant successivement les dérivations partielles. On notera :

$$\forall \alpha \in \mathbb{N}^n \ , \ D^\alpha := D_1^{\alpha_1} \cdots D_n^{\alpha_n}.$$

On note parfois également $D^\alpha P := \dfrac{\partial^{|\alpha|} P}{\partial X_1^{\alpha_1} \cdots \partial X_n^{\alpha_n}}$: par exemple, $\dfrac{\partial^3 P}{\partial X^2 \partial Y}$, etc.

La règle suivante est conséquence immédiate de la commutation des D_i :

$$\forall \alpha, \beta \in \mathbb{N}^n \ , \ D^\alpha \circ D^\beta = D^{\alpha + \beta}.$$

Pour calculer les dérivées itérées, définissons les factorielles des multiindices :

$$\forall \alpha := (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n \ , \ \alpha! := \alpha_1! \cdots \alpha_n!$$

Enfin, dans tout ce paragraphe, nous utiliserons l'*ordre produit* sur $\mathbb{N}^n$:

$\forall \alpha=(\alpha_1,\dots,\alpha_n),\beta=(\beta_1,\dots,\beta_n)\in\mathbb{N}^n\ ,\ \alpha\ll\beta \iff (\alpha_1\leqslant\beta_1 \text{ et } \cdots \text{ et } \alpha_n\leqslant\beta_n)$.

On a alors les formules suivantes :

$$\forall \alpha,\beta\in\mathbb{N}^n\ ,\ D^\alpha X^\beta=\begin{cases}\dfrac{\beta!}{(\beta-\alpha)!}X^{\beta-\alpha} \text{ si } \alpha\ll\beta,\\ 0 \text{ sinon.}\end{cases} \tag{5}$$

En effet, les α_i applications successives de D_i multiplient successivement X^β par $\dfrac{\beta_i}{X_i}$, $\dfrac{\beta_i-1}{X_i},\dots,\dfrac{\beta_i-\alpha_i+1}{X_i}$, donc, au total, par $\dfrac{\beta_i!}{(\beta_i-\alpha_i)!}X_i^{-\alpha_i}$ (si $\alpha_i\leqslant\beta_i$) ou par 0 (si $\alpha_i>\beta_i$). Comme les $D_i^{\alpha_i}$ commutent entre elles et que chacune n'a d'effet que sur l'indéterminée correspondante, la conclusion s'ensuit. En particulier, si $\deg P<|\alpha|$, alors $D^\alpha P=0$.

Proposition 11 (Formule de Leibniz). Soient $P,Q\in K[X_1,\dots,X_n]$. Alors :

$$\forall\alpha\in\mathbb{N}^n\ ,\ D^\alpha(PQ)=\sum_{\beta+\gamma=\alpha}\frac{\alpha!}{\beta!\,\gamma!}(D^\beta P)(D^\gamma Q).$$

Démonstration. Pour $\alpha=(0,\dots,0)$, c'est trivial. On va le démontrer par récurrence selon le schéma suivant : si la propriété est vraie pour $\alpha\in\mathbb{N}^n$, elle est vraie pour chaque $\alpha+E_i$ ($i\in[\![1,n]\!]$). Un instant de réflexion permettra au lecteur de se convaincre que cela revient à faire une récurrence sur le poids $|\alpha|$. Le calcul se mène ainsi :

$$\begin{aligned}
D^{\alpha+E_i}(PQ)&=D_i\big(D^\alpha(PQ)\big)=\sum_{\beta+\gamma=\alpha}\frac{\alpha!}{\beta!\gamma!}D_i\big((D^\beta P)(D^\gamma Q)\big)\\
&=\sum_{\beta+\gamma=\alpha}\frac{\alpha!}{\beta!\gamma!}\big((D^{\beta+E_i}P)(D^\gamma Q)+(D^\beta P)(D^{\gamma+E_i}Q)\big)\\
&=\sum_{\beta+\gamma=\alpha}\frac{\alpha!}{\beta!\gamma!}(D^{\beta+E_i}P)(D^\gamma Q)+\sum_{\beta+\gamma=\alpha}\frac{\alpha!}{\beta!\gamma!}(D^\beta P)(D^{\gamma+E_i}Q)\\
&=\sum_{\beta+\gamma=\alpha+E_i}\frac{\alpha!\beta_i}{\beta!\gamma!}(D^\beta P)(D^\gamma Q)+\sum_{\beta+\gamma=\alpha+E_i}\frac{\alpha!\gamma_i}{\beta!\gamma!}(D^\beta P)(D^\gamma Q)\\
&=\sum_{\beta+\gamma=\alpha+E_i}\frac{\alpha!(\beta_i+\gamma_i)}{\beta!\gamma!}(D^\beta P)(D^\gamma Q)\\
&=\sum_{\beta+\gamma=\alpha+E_i}\frac{(\alpha+E_i)!}{\beta!\gamma!}(D^\beta P)(D^\gamma Q),
\end{aligned}$$

ce que l'on désirait. Noter que l'antépénultième membre droit a été obtenu par changement d'indexation et qu'il n'y a pas d'erreur : les termes apparemment excédentaires sont nuls. ∎

On suppose maintenant le corps K de caractéristique nulle. Dans cette situation, la dérivée itérée $D_\alpha P$ est nulle si, et seulement si, tous les monômes de P sont strictement inférieurs à α (pour l'ordre produit) ; et la dérivée itérée $D_\alpha P$ est constante non nulle si, et seulement si, $|\alpha| = \deg P$ et X^α est un monôme de P. Plus précisément, si $P := \sum p_\alpha X^\alpha \in K[X_1, \dots, X_n]$, on tire de l'équation (5), page 327 :

$$\forall \alpha \in \mathbb{N}^n \ , \ p_\alpha = \frac{1}{\alpha!} D^\alpha P(0, \dots, 0). \tag{6}$$

Théorème 12 (Formule de Taylor).

On note $D_1, \dots, D_n$ les dérivations $\dfrac{\partial}{\partial X_i}$ de $K[X_1, \dots, X_n]$. Soit $P \in K[X_1, \dots, X_n]$. Dans la K-algèbre $K[X_1, \dots, X_n][Y_1, \dots, Y_n] = K[X_1, \dots, X_n, Y_1, \dots, Y_n]$, on a l'égalité :

$$P(X_1 + Y_1, \dots, X_n + Y_n) = \sum \frac{D^\alpha P(Y_1, \dots, Y_n)}{\alpha!} X^\alpha.$$

Démonstration. On a *a priori* une relation de la forme :

$$P(X_1 + Y_1, \dots, X_n + Y_n) = \sum p_\alpha X^\alpha, p_\alpha \in L := K(Y_1, \dots, Y_n).$$

Pour déterminer les p_α, il suffit alors d'appliquer la relation (6) au polynôme $P(X_1 + Y_1, \dots, X_n + Y_n)$ de $L[X_1, \dots, X_n]$, en invoquant de plus l'exemple de la page 326. ∎

Toutes les autres formules de Taylor (analogues à celles qui seront données dans le module II.5) s'obtiennent par substitution à partir de celle-là. Par exemple, en substituant $X_i \mapsto X_i - a_i$ et $Y_i \mapsto a_i$, on obtient, pour tout $a := (a_1, \dots, a_n) \in K^n$:

$$P = P(a) + \sum_{i=1}^{n} D_i P(a)(X_i - a_i) + \frac{1}{2} \sum_{i,j=1}^{n} D_i D_j P(a)(X_i - a_i)(X_j - a_j) + \cdots$$

Les termes suivants sont tous éléments de $\mathfrak{M}_a^3$ (notation introduite page 318) et l'on peut considérer une telle écriture comme une sorte de développement limité à l'ordre 2.

Exercice 8.

Démontrer que l'application $P \mapsto \sum \dfrac{D^\alpha P(X_1, \dots, X_n)}{\alpha!} Y^\alpha$ de $K[X_1, \dots, X_n]$ dans $K[X_1, \dots, X_n][Y_1, \dots, Y_n]$ est un morphisme, et en déduire une nouvelle preuve de la formule de Leibniz (proposition 11 de la page précédente).

Solution. D'après la formule de Taylor, notre application est simplement le morphisme associé à la substitution $X_i \mapsto X_i + Y_i$. Notons φ ce morphisme. Dans l'égalité $\varphi(PQ) = \varphi(P)\varphi(Q)$, si l'on identifie les coefficients des monômes Y^α des deux membres (coefficients qui appartiennent à $K[X_1, \dots, X_n]$), on obtient exactement la formule de Leibniz. Il faut cependant noter que cette preuve n'est valide qu'en caractéristique nulle.

1.4 Polynômes symétriques

1.4.1 L'action du groupe symétrique

Pour toute permutation $\sigma \in \mathfrak{S}_n$, notons Φ_σ l'endomorphisme de $S = K[X_1, \ldots, X_n]$ associé à la substitution $X_i \mapsto X_{\sigma(i)}$ $(i \in [\![1, n]\!])$.

On pose $\sigma P := \Phi_\sigma(P) = P(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$. Comme $\Phi_\sigma \circ \Phi_\tau(X_i) = \Phi_{\sigma\tau}(X_i)$ pour tout i, on a $\Phi_\sigma \circ \Phi_\tau = \Phi_{\sigma\tau}$; de même, Φ_{Id} est l'identité de $K[X_1, \ldots, X_n]$, donc Φ_σ est un automorphisme d'inverse $\Phi_{\sigma^{-1}}$.

On a donc un morphisme injectif de groupes : $\sigma \mapsto \Phi_\sigma$ de $\mathfrak{S}_n$ dans le groupe $\mathrm{Aut}_K(K[X_1, \ldots, X_n])$ des automorphismes de K-algèbre de $K[X_1, \ldots, X_n]$. Il s'agit évidemment d'une action de $\mathfrak{S}_n$ sur $K[X_1, \ldots, X_n]$, et l'on dit que *le groupe* $\mathfrak{S}_n$ *agit sur* $K[X_1, \ldots, X_n]$ *par automorphismes de* K*-algèbre.*

Pour tout $\sigma \in \mathfrak{S}_n$, l'ensemble $S^\sigma := \{P \in S \mid \sigma P = P\}$ des points fixes de σ est une sous-algèbre de S ; il en est donc de même de l'ensemble $\{P \in S \mid \forall \sigma \in \mathfrak{S}_n\ ,\ \sigma P = P\}$ des points fixes de $\mathfrak{S}_n$. On notera $K[X_1, \ldots, X_n]^{\mathrm{sym}}$ cette K-algèbre.

Définition 6. Les éléments de la K-algèbre $K[X_1, \ldots, X_n]^{\mathrm{sym}}$ sont appelés *polynômes symétriques en les indéterminées* $X_1, \ldots, X_n$.

Pour alléger les notations, nous noterons ici $S' := K[X_1, \ldots, X_n]^{\mathrm{sym}}$. Il est clair que chaque composante homogène S_k est invariante par chaque application Φ_σ. On a donc une décomposition en somme directe de sous-espaces vectoriels : $S' = \bigoplus_{k \in \mathbb{N}} S'_k$, où $S'_k := S_k \cap S'$. Autrement dit, pour que le polynôme $P \in S$ soit symétrique, il faut, et il suffit, que chacune de ses composantes homogènes $p_k(P)$ le soit.

Exemples. Les coefficients du polynôme $\prod_{i=1}^{n} (T + X_i) \in K[X_1, \ldots, X_n][T]$ en l'indéterminée T sont les *polynômes symétriques élémentaires* : $\sigma_k = \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} X_{i_1} \cdots X_{i_k} \in S'$.

Plus généralement, soient $P \in S$, et $E := \mathfrak{S}_n P$ son orbite. Les coefficients du polynôme $\prod_{Q \in E} (T + Q) \in K[X_1, \ldots, X_n][T]$ en l'indéterminée T sont symétriques. Par exemple, en prenant $P := X_1^k$ $(k \in \mathbb{N})$, on trouve pour coefficient de T^{n-1} la *somme de Newton*[1] : $s_k := X_1^k + \cdots + X_n^k$. De même, le polynôme $P := \prod_{\substack{1 \leqslant i, j \leqslant n \\ i \neq j}} (X_i - X_j)$ est symétrique, car l'ensemble de ses facteurs est l'orbite de $X_1 - X_2$ sous l'action de $\mathfrak{S}_n$.

Calcul de la dimension de S'_k. Notons $\mathbb{N}^n_k := \{\alpha \in \mathbb{N}^n \mid |\alpha| = k\}$ et $\mathbb{N}^n_k/\mathfrak{S}_n$ son quotient sous l'action de $\mathfrak{S}_n$ (c'est-à-dire l'ensemble des orbites). Il est évident que $\mathfrak{S}_n$ agit sur $\mathbb{N}^n$ par permutation des indices, et que le sous-ensemble $\mathbb{N}^n_k$ est stable. Soit $O \in \mathbb{N}^n_k/\mathfrak{S}_n$ une orbite. Le polynôme $P_O := \sum_{\alpha \in O} X^\alpha$ est symétrique et homogène

[1] Dans le [L1], les sommes de Newton étaient notées S_k, mais nous désirons ici éviter toute confusion avec l'espace vectoriel des polynômes homogènes de degré k.

de degré k. Lorsque O parcourt $\mathbb{N}_k^n/\mathfrak{S}_n$, la famille des P_O forme une base de S'_k : en effet, si $P := \sum p_\alpha X^\alpha \in S'_k$, tous les p_α correspondant aux α d'une même orbite sont égaux, et P est de la forme $\sum p_O P_O$; et il est clair que les P_O sont linéairement indépendants. La dimension du K-espace vectoriel S'_k est donc égale à $\operatorname{card}\left(\mathbb{N}_k^n/\mathfrak{S}_n\right)$. Lorsque par exemple $n := 2$, la base de monômes de S_3 est formée des deux orbites : $\{X^3, Y^3\}$ et $\{X^2Y, XY^2\}$, donc $\dim S'_3 = 2$. Lorsque $n := 3$, la base de monômes de S_3 est formée des trois orbites : $\{X^3, Y^3, Z^3\}$, $\{X^2Y, XY^2, X^2Z, XZ^2, Z^2Y, ZY^2\}$ et $\{XYZ\}$, donc $\dim S'_3 = 3$.

La définition de l'action de $\mathfrak{S}_n$ sur S s'étend aux fractions rationnelles en posant :

$$\sigma\left(\frac{P}{Q}\right) := \frac{\sigma P}{\sigma Q};$$

le lecteur vérifiera en effet que, pour toute fraction rationnelle R, la valeur de σR ne dépend pas de la représentation particulière $R = \dfrac{P}{Q}$.

On obtient ainsi des automorphismes de corps, et l'ensemble des points fixes (*fractions rationnelles symétriques*) est un sous corps $K(X_1, \ldots, X_n)^{sym}$ de $K(X_1, \ldots, X_n)$. En fait, c'est le corps des fractions de $K[X_1, \ldots, X_n]^{\mathrm{sym}}$. Posons en effet, pour un élément arbitraire $\dfrac{P}{Q}$ de ce corps, $Q_1 := \prod_{\sigma \in \mathfrak{S}_n \setminus \{\mathrm{Id}\}} \sigma Q$, de sorte que $QQ_1 = \prod_{\sigma \in \mathfrak{S}_n} \sigma Q \in S'$. De l'égalité $\dfrac{P}{Q} = \dfrac{PQ_1}{QQ_1}$, on déduit que $\dfrac{P}{Q}$ est symétrique si, et seulement si, $PQ_1 \in S'$. Les fractions rationnelles symétriques sont donc les quotients de polynômes symétriques (voir aussi l'exercice I.7.27 de la page 349).

1.4.2 Le théorème fondamental sur les polynômes symétriques

Comme chaque polynôme symétrique élémentaire σ_i est homogène de degré i, un monôme $\sigma^\beta := \sigma_1^{\beta_1} \cdots \sigma_n^{\beta_n}$ ($\beta \in \mathbb{N}^n$) en les σ_i est homogène de degré $\sum_{i=1}^n i\beta_i$.

Proposition et définition 13. On appelle *poids pondéré* du multiindice β l'entier $w(\beta) := \sum_{i=1}^n i\beta_i$. Le K-espace vectoriel S'_k des polynômes symétriques homogènes de degré k a pour base la famille des σ^β, où $\beta \in \mathbb{N}^n$ est de poids pondéré $w(\beta) := k$.

Démonstration. Il est clair que $\sigma^\beta \in S'_{w(\beta)}$. Nous allons d'abord prouver par double récurrence (sur n et sur k) que tout élément de S'_k est combinaison linéaire des σ^β tels que $w(\beta) = k$. Pour $n := 1$, on a $\sigma_1 = X_1$, $w(\beta) = |\beta|$ et, quel que soit k, $S_k = S'_k = KX_1^k$; la conclusion est alors facile. On suppose donc l'assertion d'existence vraie pour $n - 1 \geqslant 1$ indéterminées et l'on va la prouver pour n indéterminées, par récurrence forte sur k. Pour $k := 0$, on a $S_k = S'_k = K$ et $w(\beta) = 0 \Leftrightarrow \sigma^\beta = 1$: l'assertion est encore évidente. Supposons la vraie pour tous les degrés $k' < k$, où $k \geqslant 1$, et soit $P \in S'_k$.

Le polynôme $P(X_1, \ldots, X_{n-1}, 0)$ en les indéterminées $X_1, \ldots, X_{n-1}$ est symétrique et homogène de degré k ou bien nul. Notons $\sigma'_1, \ldots, \sigma'_{n-1}$ les polynômes symétriques élémentaires en $X_1, \ldots, X_{n-1}$: on a donc, pour tout entier i tel que $1 \leqslant i \leqslant n-1$, $\sigma'_i = \sigma_i(X_1, \ldots, X_{n-1}, 0)$. Par hypothèse de récurrence, $P(X_1, \ldots, X_{n-1}, 0)$ est une combinaison linéaire de monômes ${\sigma'_1}^{\beta_1} \cdots {\sigma'_{n-1}}^{\beta_{n-1}}$ de poids pondérés $\sum_{i=1}^{n-1} i\beta_i = k$; on écrit cela $P(X_1, \ldots, X_{n-1}, 0) = Q(\sigma'_1, \ldots, \sigma'_{n-1})$. On voit alors que le polynôme $P - Q(\sigma_1, \ldots, \sigma_{n-1}) \in S$ est symétrique et homogène de degré k. De plus, il est annulé par la substitution $X_n \mapsto 0$, donc est multiple de X_n, donc de tous les X_i (car symétrique), donc de leur produit $X_1 \cdots X_n = \sigma_n$ (car les X_i sont premiers). On peut donc écrire $P - Q(\sigma_1, \ldots, \sigma_{n-1}) = \sigma_n R$, et il est clair que R est symétrique et homogène de degré $k-n$, ou bien nul. Dans le dernier cas, la preuve est achevée. Dans le premier cas, l'hypothèse de récurrence forte entraîne que R est une combinaison linéaire de monômes de poids pondéré $k-n$ en les σ_i, donc que $\sigma_n R$ est une combinaison linéaire de monômes de poids pondéré k en les σ_i, donc $P = Q(\sigma_1, \ldots, \sigma_{n-1}) + \sigma_n R$ également, ce qui achève la démonstration d'existence.

Prouvons maintenant brièvement l'unicité de cette écriture par la même méthode. Si $\sum q_\beta \sigma^\beta = 0$ (avec tous les $w(\beta) = k$), en faisant $X_n \mapsto 0$ on trouve une relation similaire en les σ'_i d'où l'on tire (récurrence sur n) que q_β pour $\beta_n = 0$. La relation $\sum q_\beta \sigma^\beta = 0$ ne fait donc intervenir que des multiples de σ_n ; en la simplifiant, on obtient une relation de degré $k-n$ et l'on conclut par récurrence forte sur le degré. ∎

Théorème 14 (Théorème fondamental sur les polynômes symétriques).
L'algèbre $K[X_1, \ldots, X_n]^{\text{sym}}$ est librement engendrée par $\sigma_1, \ldots, \sigma_n$.

Cela signifie que tout polynôme symétrique $P \in K[X_1, \ldots, X_n]^{\text{sym}}$ s'écrit de manière unique comme polynôme en les σ_i : ceux-ci engendrent la K-algèbre $K[X_1, \ldots, X_n]^{\text{sym}}$ et ils sont algébriquement indépendants ; c'est une conséquence immédiate de la proposition 13. On peut le dire autrement : le morphisme de K-algèbres de $K[T_1, \ldots, T_n]$ dans $K[X_1, \ldots, X_n]$ associé à la substitution $T_i \mapsto \sigma_i$ ($i \in [\![1, n]\!]$) est un isomorphisme d'algèbre de $K[T_1, \ldots, T_n]$ sur $K[X_1, \ldots, X_n]^{\text{sym}}$.

Exemple. La somme de Newton $s_0 := \sum_{i=1}^{n} X_i^0$, qui est homogène de degré 0, est égale à n et $s_1 := \sum_{i=1}^{n} X_i$, qui est homogène de degré 1, est égale à σ_1. On vérifie facilement que $s_2 := \sum_{i=1}^{n} X_i^2$, qui est homogène de degré 2, est égale à $\sigma_1^2 - 2\sigma_2$. Dans le cas où $n := 3$ (indéterminées X, Y, Z), étudions $s_3 := X^3 + Y^3 + Z^3$ selon la méthode de la proposition 13. Le polynôme $s_3(X, Y, 0) = X^3 + Y^3$ est égal (à l'œil nu) à ${\sigma'_1}^3 - 3\sigma'_1\sigma'_2$. On introduit donc $s_3 - (\sigma_1^3 - 3\sigma_1\sigma_2) = 3XYZ = 3\sigma_3$, et l'on conclut que $s_3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3$.

Exercice 9.

Soient $\sigma_1, \dots \sigma_n$ (resp. $\sigma'_1, \dots \sigma'_p$) les polynômes symétriques élémentaires en $X_1, \dots, X_n$ (resp. $X'_1, \dots, X'_p$). Comment exprimer $\prod_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant p}} (X_i - X'_j)$ en fonction des σ_i et des σ'_j ?

Solution. Il suffit de le développer sous la forme $\prod_{i=1}^{n} \left(X_i^p - \sigma'_1 X_1^{p-1} + \cdots + (-1)^p \sigma'_p\right)$: on trouve un polynôme symétrique en les X_i, à coefficients des polynômes en les σ'_j. Une autre méthode apparaîtra à la section 2.2. Par exemple, avec les indéterminées X, Y et X', Y', Z', on écrit $(X - X')(X - Y')(X - Z')(Y - X')(Y - Y')(Y - Z')$ sous la forme d'un produit $(X^3 - \sigma'_1 X^2 + \sigma'_2 X - \sigma'_3)(Y^3 - \sigma'_1 Y^2 + \sigma'_2 Y - \sigma'_3)$, ce qui donne :

$$\begin{aligned}&\sigma_2^3 - \sigma'_1\sigma_1\sigma_2^2 + \sigma'_2(\sigma_1^2 - 2\sigma_2)\sigma_2 - \sigma'_3(\sigma_1^3 - 3\sigma_1\sigma_2) + \sigma'^2_1\sigma_2^2\\ &\quad - \sigma'_1\sigma'_2\sigma_1\sigma_2 + \sigma'_1\sigma'_3(\sigma_1^2 - 2\sigma_2) + \sigma'^2_2\sigma_2 - \sigma'_2\sigma'_3\sigma_1 + \sigma'^2_3.\end{aligned}$$

Théorème 15 (Formules de Newton). Entre les polynômes symétriques élémentaires et les sommes de Newton (page 329), on a les relations suivantes :

$$\begin{aligned}\forall k \in [\![1, n]\!]\ ,\ s_k &= \sum_{i=1}^{k} (-1)^{i-1}\sigma_i s_{k-i} + (-1)^k(n-k)\sigma_k,\\ &= \sum_{i=1}^{k-1} (-1)^{i-1}\sigma_i s_{k-i} + (-1)^{k-1}k\sigma_k,\\ \forall k \geqslant n\ ,\ s_k &= \sum_{i=1}^{n} (-1)^{i-1}\sigma_i s_{k-i}.\end{aligned}$$

Naturellement, ces formules peuvent servir de relations de récurrence pour calculer les s_k. La première et la troisième égalité ont été démontrées dans le module « Polynômes » de [L1], avec des éléments x_i d'un corps au lieu d'indéterminées. Pour en déduire le cas apparemment plus général énoncé ici, il suffit de prendre pour corps le corps $K(X_1, \dots, X_n)$ des fractions rationnelles et pour polynôme sur ce corps le polynôme $P(T) = (T - X_1)\cdots(T - X_n)$. La deuxième égalité est une conséquence immédiate de la première, car $s_0 = n$ et le terme en $i = k$ de la somme se combine avec le terme extérieur.

Exercice 10.

Calculer s_4 en fonction de σ_1, σ_2, σ_3 et σ_4.

Solution. On sait que $s_0 = 4$, $s_1 = \sigma_1$ et $s_2 = \sigma_1^2 - 2\sigma_2$; puis on calcule successivement :

$$\begin{aligned}s_3 &= \sigma_1 S_2 - \sigma_2 s_1 + \sigma_3 s_0 - \sigma_3 = \sigma_1(\sigma_1^2 - 2\sigma_2) - \sigma_2\sigma_1 + 3\sigma_3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3,\\ s_4 &= \sigma_1 s_3 - \sigma_2 s_2 + \sigma_3 s_1 - \sigma_4 s_0\\ &= \sigma_1(\sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3) - \sigma_2(\sigma_1^2 - 2\sigma_2) + \sigma_3\sigma_1 - \sigma_4 s_0\\ &= \sigma_1^4 - 4\sigma_1^2\sigma_2 + 4\sigma_1\sigma_3 + 2\sigma_2^2 - 4\sigma_4.\end{aligned}$$

Voici une autre méthode pour relier les σ_i et les s_i. On introduit le polynôme en une indéterminée T à coefficients dans le corps $L := K(X_1,\dots,X_n)$:

$$P(T) := \prod_{i=1}^{n}(1-X_iT) = 1-\sigma_1T+\sigma_2T^2+\cdots+(-1)^n\sigma_nT^n.$$

On écrit alors de deux manières la fraction rationnelle $-T\dfrac{P'}{P}$:

$$\frac{\sigma_1T-2\sigma_2T^2+\cdots+(-1)^{n-1}n\sigma_nT^n}{1-\sigma_1T+\sigma_2T^2+\cdots+(-1)^n\sigma_nT^n} = \sum_{i=1}^{n}\frac{X_iT}{1-X_iT}.$$

On compare alors les coefficients des développements limités des deux membres de cette égalité, en utilisant la relation $\dfrac{1}{1-u} \equiv 1+u+\cdots+u^p \pmod{u^{p+1}}$. Pour $k \geqslant 1$, le coefficient de T^k dans le développement limité du membre de droite est s_k. Pour le membre de gauche, il faut développer, en tronquant à l'ordre k, l'expression :

$$\bigl(\sigma_1T-2\sigma_2T^2+\cdots+(-1)^{n-1}n\sigma_nT^n\bigr)\sum_{i=0}^{k-1}\bigl(\sigma_1T-\sigma_2T^2+\cdots+(-1)^{n-1}\sigma_nT^n\bigr)^i.$$

Par exemple, pour $n := 3$ (indéterminées X, Y, Z), si l'on veut calculer s_3, on écrit :

$$\begin{aligned}&\frac{\sigma_1T-2\sigma_2T^2+3\sigma_3T^3}{1-(\sigma_1T-\sigma_2T^2+\sigma_3T^3)}\\ &\quad\equiv\bigl(\sigma_1T-2\sigma_2T^2+3\sigma_3T^3\bigr)\bigl(1+(\sigma_1T-\sigma_2T^2+\sigma_3T^3)+(\sigma_1T-\sigma_2T^2+\sigma_3T^3)^2\bigr)\\ &\qquad\qquad\pmod{T^4},\end{aligned}$$

d'où $s_3 = \sigma_1(-\sigma_2+\sigma_1^2)-2\sigma_2(\sigma_1)+3\sigma_3(1) = \sigma_1^3-3\sigma_1\sigma_2+3\sigma_3$.

1.4.3 Fonctions polynomiales symétriques des racines d'un polynôme

La principale application du théorème 14 de la page 331, à l'origine de la théorie de Galois, est le résultat suivant. Les fonctions rationnelles ont été définies à la page 317.

Théorème 16. Soient $P := a_0+\cdots+a_nT^n \in K[T]$ $(n \in \mathbb{N}^*)$ et L une extension de K dans lequel P se scinde : $P = a_n\prod_{i=1}^{n}(X-x_i)$ (avec $x_i \in L$). Alors toute expression rationnelle symétrique en les x_i (à coefficients dans K) est un élément de K.

Démonstration. Cela découle des relations entre coefficients et racines (module « Polynômes » de [L1]) : $\forall i \in [\![1,n]\!]$, $\sigma_i(x_1,\dots,x_n) = (-1)^i\dfrac{a_{n-i}}{a_n}$. ∎

Exemples. (i) Avec ces notations, $\prod_{1\leqslant i<j\leqslant n}(x_i-x_j)^2 \in K$.

(ii) Si l'on pose en outre $Q := b_0+\cdots+b_pT^p \in K[T]$ $(p \in \mathbb{N}^*)$ et que l'on suppose que Q est scindé dans L : $Q = b_p\prod_{j=1}^{p}(X-y_j)$ $(y_j \in L)$, on voit de même que $\prod(x_i-y_j) \in K$; c'est une variante de l'exercice 9 de la page précédente (voir également la section 2.2).

Première application : nombres algébriques. Soient x et y deux nombres algébriques ; on suppose x non nul. Il existe donc $P := a_0 + \cdots + a_n T^n \in \mathbb{Q}[T]$ ($n \in \mathbb{N}^*$, $a_0, a_n \neq 0$) dont x est racine et $Q := b_0 + \cdots + b_p T^p \in Q[T]$ ($p \in \mathbb{N}^*$, $b_p \neq 0$) dont y est racine. Ces polynômes sont scindés dans $\mathbb{C}$: $P = a_n \prod_{i=1}^{n} (X - x_i)$ (avec $x_1 = x$, tous les $x_i \neq 0$) et $Q = b_p \prod_{j=1}^{p} (X - y_j)$ (avec $y_1 = y$). Les coefficients des polynômes $R := \prod (T - x_i - y_j) = b_p^{-n} \prod Q(T - x_i)$ et $S := \prod (T - x_i y_j) = b_p^{-n} \prod \big(x_i^p Q(T/x_i)\big)$ sont des polynômes à coefficients rationnels en les $\dfrac{a_i}{a_n}$ et les $\dfrac{b_j}{b_p}$. On a donc $A, B \in \mathbb{Q}[T]$. Les nombres $x + y$ et xy, qui sont respectivement racine de A et de B, sont donc algébriques.

Deuxième application : théorème de d'Alembert-Gauß. Nous allons donner une nouvelle démonstration de ce théorème (déjà prouvé dans le module sur les complexes de [L1]), basée sur les deux faits suivants : tout polynôme de $\mathbb{R}[X]$ de degré impair admet une racine réelle ; tout polynôme du second degré dans $\mathbb{C}$ est scindé dans $\mathbb{C}$. Le premier fait découle du théorème des valeurs intermédiaires et relève donc de l'analyse. Le second fait est de nature algébrique (sauf en ce qui concerne l'existence des racines carrées des réels positifs). Nous devrons cependant *admettre* que le corps $\mathbb{C}$ est inclus dans un corps algébriquement clos L : c'est en fait vrai de tout corps et sera démontré dans [MPA1], de manière purement algébrique (voir cependant l'exercice I.7.32 de la page 349).
Remarquons tout d'abord que, si $P := a_0 + \cdots + a_n X^n \in \mathbb{C}[X]$ ($n \in \mathbb{N}^*$), notant $\overline{P} := \overline{a_0} + \cdots + \overline{a_n} X^n$, on a $Q := P\overline{P} \in \mathbb{R}[X]$ d'une part, et $Q(x) = 0$ si, et seulement si, $P(x) = 0$ ou $P(\overline{x}) = 0$. Il nous suffit donc de démontrer que tout polynôme non constant à coefficients réels admet au moins une racine complexe. Soit donc $P \in \mathbb{R}[X]$ un polynôme de degré $n \in \mathbb{N}^*$. Nous pouvons évidemment supposer P unitaire. On écrit $n = 2^k m$, où $k \in \mathbb{N}$ et $m \in \mathbb{N}^*$ est impair. La preuve procède par récurrence sur k. Si $k := 0$, c'est le premier fait rappelé ci-dessus.
Supposons $k \geqslant 1$. On écrit $P = \prod_{i=1}^{n} (X - x_i)$, où les x_i sont dans L (qui est algébriquement clos). Il nous faut prouver que l'un au moins des x_i est dans $\mathbb{C}$. Pour tout réel c, définissons : $y_{i,j}(c) := x_i + x_j + c x_i x_j$ et $Q_c(X) = \prod_{1 \leqslant i < j \leqslant n} \big(X - y_{i,j}(c)\big)$. Nous démontrerons plus loin que $Q_c \in \mathbb{R}[X]$. De plus, $\deg Q_c = \dfrac{n(n-1)}{2} = 2^{k-1} m'$, où $m' := m(n-1)$ est impair. Par hypothèse de récurrence, l'un des $y_{i,j}(c)$ au moins est dans $\mathbb{C}$. Mais cela est vrai pour tout réel c. Comme il y a une infinité de réels et un nombre fini de couples (i, j), on voit qu'il existe deux réels distincts c, d tels que $y_{i,j}(c), y_{i,j}(d) \in \mathbb{C}$. On en déduit que $s := x_i + x_j$ et $p := x_i x_j$ sont dans $\mathbb{C}$, donc racines de l'équation du second degré $X^2 - sX + p = 0$, à coefficients dans $\mathbb{C}$, donc $x_i, x_j \in \mathbb{C}$.
Reste à prouver que $Q_c \in \mathbb{R}[X]$. Les coefficients de Q_c sont des polynômes en les x_i ; il suffit donc de vérifier que ce sont des polynômes symétriques. La petite difficulté de ce point

est que $\sigma \in \mathfrak{S}_n$ ne permute *a priori* pas les couples (i,j) tels que $1 \leqslant i < j \leqslant n$, et l'indexation naturelle des facteurs de Q_c ne montre pas immédiatement l'invariance par l'action de $\mathfrak{S}_n$. Soit A l'ensemble des paires $\{i,j\} \subset [\![1,n]\!]$. Pour chaque paire $a := \{i,j\} \in A$, on peut poser : $z_a := x_i + x_j + cx_ix_j$; en effet, cette expression ne dépend par de l'ordre des éléments i,j. On a alors $Q_c = \prod_{a\in A}(X - z_a)$. Le groupe $\mathfrak{S}_n$ agit sur A. Toute permutation $\sigma \in \mathfrak{S}_n$ permute les z_a selon la règle : $z_{\sigma(a)} = x_{\sigma(i)} + x_{\sigma(j)} + cx_{\sigma(i)}x_{\sigma(j)}$. Si l'on remplace chaque x_i par $x_{\sigma(i)}$, l'expression $\prod_{a\in A}(X - z_a)$ est invariante.

Troisième application : matrices séparables. Soient K un corps algébriquement clos et $A \in M_n(K)$, de valeurs propres $\lambda_1, \ldots, \lambda_n$. Alors $\prod_{i\neq j}(\lambda_i - \lambda_j)$ est un polynôme en les coefficients $c_i(A)$ du polynôme caractéristique (définis dans les exemples des pages 315 et 316), donc également un polynôme en les coefficients de A. Notons le $\mathrm{Sep} \in K[X_{1,1}, \ldots, X_{n,n}]$ (en fait, il est à coefficients dans $\mathbb{Z}$ d'après les remarques antérieures). Nous avons donc trouvé un polynôme « universel » tel que :

$$\forall A := (a_{i,j})_{1\leqslant i,j\leqslant n} \in M_n(K)\ ,\ \mathrm{Sep}\big((a_{i,j})_{1\leqslant i,j\leqslant n}\big) \neq 0 \Longleftrightarrow A \text{ est séparable},$$

où l'on appelle *séparable* une matrice dont les valeurs propres sont deux à deux distinctes. En vertu de la théorie de la réduction, on a donc :

$$\forall A := (a_{i,j})_{1\leqslant i,j\leqslant n} \in M_n(K)\ ,\ \mathrm{Sep}\big((a_{i,j})_{1\leqslant i,j\leqslant n}\big) \neq 0 \Longrightarrow A \text{ est diagonalisable}.$$

Nous en déduirons dans l'exercice I.7.41 de la page 350 une nouvelle preuve du théorème de Cayley-Hamilton. Lorsque par exemple $n := 2$, le polynôme caractéristique de $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ est $T^2 - (a+d)T + (ad-bc)$. Soient λ, μ les valeurs propres. Alors :

$$\begin{aligned}\mathrm{Sep}(a,b,c,d) = -(\lambda-\mu)^2 &= -\sigma_1^2(\lambda,\mu) + 4\sigma_2(\lambda,\mu) \\ &= -(a+d)^2 + 4(ad-bc) = -(a-d)^2 - 4bc.\end{aligned}$$

2 Fonctions polynomiales

Dans toute cette section, le corps K sera supposé infini. Le cas d'un corps fini est évoqué dans l'exercice I.7.37 de la page 350.

2.1 Prolongement des identités algébriques

La fonction polynomiale $\tilde{P} \in \mathcal{F}(K^n, K)$ associée à un polynôme $P \in K[X_1, \ldots, X_n]$ a été définie à la page 317 comme l'application $\tilde{P} : (x_1, \ldots, x_n) \mapsto P(x_1, \ldots, x_n)$. Contrairement au cas des polynômes à une indéterminée, *l'application $\tilde{P}$ peut s'annuler en une infinité de points même si le polynôme P est non nul* : il suffit de considérer par exemple les polynômes $Y - aX - b$ ou $XY - 1$ de $\mathbb{R}[X]$, dont les lieux d'annulation sont respectivement la droite $y = ax + b$ et l'hyperbole $y = 1/x$ de $\mathbb{R}^2$.

Lemme 17. Soient $A_1, \dots, A_n$ des sous-ensembles infinis de K. Si $\tilde{P}$ s'annule sur le *pavé* $A_1 \times \cdots \times A_n$, alors $P = 0$.

Démonstration.

Pour $n := 1$, cela a été démontré dans le [L1]. Supposons l'implication vraie pour des polynômes en $n-1$ indéterminées. Le polynôme $P \in K[X_1, \dots, X_n]$ peut s'écrire : $P := \sum_{i=0}^{d} P_i X_n^i$, où $P_0, \dots, P_d \in K[X_1, \dots, X_{n-1}]$. Pour tout $a := (a_1, \dots, a_n) \in K^n$, on a l'égalité : $\tilde{P}(a) = \sum_{i=0}^{d} \tilde{P}_i(a_1, \dots, a_{n-1}) a_n^i$. Fixons $(a_1, \dots, a_{n-1}) \in A_1 \times \cdots \times A_{n-1}$. Le polynôme $\sum_{i=0}^{d} \tilde{P}_i(a_1, \dots, a_{n-1}) X^i \in K[X]$ est nul sur l'ensemble infini A_n, donc nul. Ses coefficients $\tilde{P}_i(a_1, \dots, a_{n-1})$ sont donc nuls. Cela étant vrai quelque soit $(a_1, \dots, a_{n-1}) \in A_1 \times \cdots \times A_{n-1}$, on a donc $P_0 = \cdots = P_d = 0$ (hypothèse de récurrence), donc $P = 0$. ∎

Une version « quantitative » de ce lemme est proposée dans l'exercice I.7.38 de la page 350.

Exercice 11.
La conclusion est-elle valide si l'un des A_i est fini ?
Solution. Non, bien sûr : si A_i est fini, le polynôme $P := \prod_{a \in A_i} (X_i - a) \in K[X_1, \dots, X_n]$ est non nul mais $\tilde{P}$ s'annule sur le pavé $A_1 \times \cdots \times A_n$.

Théorème 18. Le corps K ayant été supposé infini, l'application $P \mapsto \tilde{P}$ de $K[X_1, \dots, X_n]$ dans $\mathcal{F}(K^n, K)$ est un morphisme injectif d'algèbres. On peut donc *identifier* le polynôme P à la fonction polynomiale $\tilde{P}$.

Démonstration. Il suffit d'appliquer le lemme avec $A_1 := \cdots := A_n := K$. ∎

Exercice 12.
L'application $P \mapsto \tilde{P}$ est-elle surjective ?
Solution. Non : considérer la fonction $a \mapsto \delta_{a,0}$, qui vaut 1 en $0 \in K^n$ et est nulle ailleurs. Si elle était de la forme $\tilde{P}$, en appliquant le lemme avec $A_1 := \cdots := A_n := K^*$, on aurait $P = 0$, donc $\tilde{P} = 0$; mais notre fonction n'est pas identiquement nulle.

Théorème 19 (Principe de prolongement des identités algébriques).
Soient $Q_1, Q_2 \in K[X_1, \ldots, X_n]$ deux polynômes qui prennent les mêmes valeurs sur le complémentaire du lieu des zéros d'un polynôme non nul $P \in K[X_1, \ldots, X_n]$:

$$\forall a \in K^n \setminus P^{-1}(0)\ ,\ Q_1(a) = Q_2(a), \tag{7}$$

Alors $Q_1 = Q_2$: autrement dit, on peut *prolonger l'identité algébrique* (7) à K^n.

Démonstration. Le polynôme $(Q_1 - Q_2)P$ s'annule (évidemment) sur $P^{-1}(0)$ et (par hypothèse) sur $K^n \setminus P^{-1}(0)$, donc sur K^n tout entier. Il est donc nul (théorème 18). Comme l'anneau $K[X_1, \ldots, X_n]$ est intègre, on déduit de l'hypothèse $P \neq 0$ que $Q_1 = Q_2$. ∎

Définition 7. On dit qu'une partie $A \subset K^n$ est *Zariski-dense* si l'on peut prolonger les identités algébriques valides sur A :

$$\forall Q_1, Q_2 \in K[X_1, \ldots, X_n]\ ,\ \big(\forall a \in A\ ,\ Q_1(a) = Q_2(a)\big) \Longrightarrow Q_1 = Q_2. \tag{8}$$

La terminologie est justifiée par la « topologie de Zariski » (voir [MPA1]). On peut cependant s'en forger une idée intuitive en comparant les exercices 13 et I.7.41 de la page 350 à leurs analogues « topologiques » du module I.6 (exercices 7 et 8 de la page 273).

Exemples.

1. Toute partie dense de $\mathbb{R}^n$ ou de $\mathbb{C}^n$ est Zariski-dense : en effet, les fonctions polynomiales sont continues (exercice I.7.40 de la page 350).
2. Toute partie qui contient une partie Zariski-dense est elle-même Zariski-dense.
3. Tout pavé $A_1 \times \cdots \times A_n$ tel que chaque A_i est infini est Zariski-dense dans K^n.
4. Selon le principe de prolongement des identités algébriques, le complémentaire du lieu des zéros de $P \in K[X_1, \ldots, X_n]$ ($P \neq 0$) est Zariski-dense dans K^n.
5. Le sous-ensemble $GL_n(K)$ de $M_n(K)$ est Zariski-dense : c'est en effet le complémentaire du lieu des zéros du polynôme déterminant, et celui-ci n'est pas nul puisque $\det I_n = 1$.

Exercice 13.
Soient $A, B \in M_n(K)$. Si A ou B est inversible, montrer que AB et BA sont semblables ; en déduire que AB et BA ont même polynôme caractéristique. À l'aide du dernier exemple ci-dessus, le démontrer sans l'hypothèse d'inversibilité.

Solution. Si, par exemple, A est inversible, alors $BA = A^{-1}(AB)A$; les matrices AB et BA sont donc semblables, donc ont même polynôme caractéristique.
Dans le cas général, fixons $B \in M_n(K)$ arbitraire et notons $f_i(A) := c_i(AB)$ et $g_i(A) := c_i(BA)$, pour $i = 1, \ldots, n$ (notation des exemples des pages 315 et 316) : ce sont donc des polynômes en les coefficients de A. De plus, f_i et g_i sont égaux sur $GL_n(K)$, donc égaux.

2.2 Résultant et discriminant

Nous allons étudier des critères pour décider si deux polynômes ont une racine commune. Ces critères fournissent des méthodes élémentaires mais très utiles de résolution de systèmes d'équations algébriques. Cette question présente toutefois des aspects délicats :

Exemple. Les polynômes $A := aX^2+bX+c$ et $B := \lambda X+\mu$ ont-ils une racine commune ? L'unique racine $-\mu/\lambda$ de B est racine de A si, et seulement si, $a\mu^2/\lambda^2 - b\mu/\lambda + c = 0$, c'est-à-dire $a\mu^2 - b\lambda\mu + c\lambda^2 = 0$. On peut également considérer les racines x' et x'' de A et poser la condition (nécessaire et suffisante) $B(x')B(x'') = 0$. Celle-ci équivaut à $0 = \lambda^2 x'x'' + \lambda\mu(x'+x'') + \mu^2 = \lambda^2 c/a - \lambda\mu b/a + \mu^2$ c'est-à-dire à $a\mu^2 - b\lambda\mu + c\lambda^2 = 0$ encore une fois. Cependant, dans le premier calcul, on a implicitement supposé que $\lambda \neq 0$ et, dans le second, que $a \neq 0$. Si $a = 0 \neq \lambda$ ou si $a \neq 0 = \lambda$, on constate que le critère $a\mu^2 - b\lambda\mu + c\lambda^2 = 0$ est encore valable. Cependant, si $a = \lambda = 0$, l'expression $a\mu^2 - b\lambda\mu + c\lambda^2$ est certainement nulle, alors que A et B n'ont pas nécessairement de zéro commun (par exemple si $\mu = 1$).

2.2.1 Matrice de Sylvester et résultant

Soient R un anneau commutatif intègre de corps des fractions L ainsi que :

$$A := a_0 + \cdots + a_p X^p \in R_p[X] \qquad \text{et} \qquad B := b_0 + \cdots + b_q X^q \in R_q[X].$$

On introduit alors la *matrice de Sylvester* :

$$S_{p,q}(A,B) := \begin{pmatrix} a_0 & 0 & \ddots & 0 & b_0 & 0 & 0 & \dots & 0 & 0 \\ a_1 & a_0 & \ddots & 0 & b_1 & b_0 & 0 & \ddots & 0 & 0 \\ \vdots & a_1 & \ddots & 0 & \vdots & b_1 & b_0 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & a_0 & \vdots & \vdots & b_1 & \ddots & 0 & 0 \\ a_{p-1} & \vdots & \ddots & a_1 & b_q & \vdots & \vdots & \ddots & b_0 & 0 \\ a_p & a_{p-1} & \ddots & \vdots & 0 & b_q & \vdots & \ddots & b_1 & b_0 \\ 0 & a_p & \ddots & \vdots & 0 & 0 & b_q & \ddots & \vdots & b_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & a_{p-1} & \vdots & \vdots & \vdots & \ddots & b_q & \vdots \\ 0 & 0 & \dots & a_p & 0 & 0 & 0 & \dots & 0 & b_q \end{pmatrix} \in M_{p+q}(R). \quad (9)$$

Les q premières colonnes sont formées avec les coefficients de A, les p suivantes avec les coefficients de B. Nous noterons $\mathrm{Res}_{p,q}(A,B) := \det S_{p,q}(A,B)$ le déterminant de la matrice de Sylvester. Noter que, dans le [L1], c'est la transposée de cette matrice que l'on avait étudiée en exercice (modules « Polynômes » et « Dimension finie »).

Exemples. Si $A := a \in R$, $S_{0,q}(A,B) = aI_q$ et $\mathrm{Res}_{0,q}(A,B) = a^q$.
De même, si $B = b \in R$, $S_{p,0}(A,B) = bI_p$ et $\mathrm{Res}_{p,0}(A,B) = b^p$.

Exercice 14.
On prend $A := aX^2 + bX + c$ et $B := \lambda X + \mu$. Calculer $\mathrm{Res}_{2,1}(A,B)$.

Solution. On trouve : $\mathrm{Res}_{2,1}(A,B) = \det \begin{pmatrix} a & \lambda & 0 \\ b & \mu & \lambda \\ c & 0 & \mu \end{pmatrix} = a\mu^2 - b\lambda\mu + c\lambda^2$.

Proposition 20. Pour que $\mathrm{Res}_{p,q}(A,B)$ soit nul, il faut, et il suffit, qu'il existe $P \in R_{q-1}[X]$ et $Q \in R_{p-1}[X]$ non tous deux nuls tels que $AP + BQ = 0$.

Démonstration. Quitte à chasser les dénominateurs, il revient au même (dans cette équivalence logique) de prendre $P \in R_{q-1}[X]$ et $Q \in R_{p-1}[X]$ ou $P \in L_{q-1}[X]$ et $Q \in L_{p-1}[X]$. On remarque que $S_{p,q}(A,B)$ est la matrice de la famille $(A, XA, \ldots, X^{q-1}A, B, XB, \ldots, X^{p-1}B)$ dans la base canonique du L-espace vectoriel $L_{p+q-1}[X]$. En conséquence, si $P := \lambda_0 + \cdots + \lambda_{q-1}X^{q-1} \in L_{q-1}[X]$ et $Q := \mu_0 + \cdots + \mu_{p-1}X^{p-1} \in L_{p-1}[X]$, alors le vecteur colonne des coordonnées de $AP + BQ$ dans la base canonique de $L_{p+q-1}[X]$ est $S_{p,q}(A,B)C_{\lambda,\mu}$, où l'on a noté $C_{\lambda,\mu}$ le vecteur colonne transposé du vecteur ligne $\begin{pmatrix} \lambda_0 & \ldots & \lambda_{q-1} & \mu_0 & \ldots & \mu_{p-1} \end{pmatrix}$. L'affirmation de la proposition en découle. ∎

Proposition 21.
Il existe $P \in R_{q-1}[X]$ et $Q \in R_{p-1}[X]$ tels que $AP + BQ = \mathrm{Res}_{p,q}(A,B)$.

Démonstration.
Des résultats sur le déterminant et la comatrice dans [L1], on déduit l'existence d'une matrice $\tilde{S}_{p,q}(A,B) \in M_{p+q}(R)$ telle que $S_{p,q}(A,B)\tilde{S}_{p,q}(A,B) = \mathrm{Res}_{p,q}(A,B)I_{p+q}$.

Notons $\lambda_0, \ldots, \lambda_{q-1}, \mu_0, \ldots, \mu_{p-1}$ les coefficients de la première colonne de $\tilde{S}_{p,q}(A,B)$. D'après la preuve de la proposition précédente, les polynômes :
$P := \lambda_0 + \cdots + \lambda_{q-1}X^{q-1} \in R_{q-1}[X]$ et $Q := \mu_0 + \cdots + \mu_{p-1}X^{p-1} \in R_{p-1}[X]$
conviennent. ∎

En combinant les équations $A(x) = 0$ et $B(x) = 0$, on peut donc se ramener à une équation $r = 0$, où $r \in R$ ne fait pas intervenir x : on dit que l'on a *éliminé* l'inconnue x.

Définition 8. Le *résultant* des polynômes $A, B \in R[X]$ de degrés respectifs $p, q \geqslant 0$ est l'élément $\mathrm{Res}(A,B) := \mathrm{Res}_{p,q}(A,B) \in R$.

Exercice 15.
Calculer le résultant de $A := aX^2 + bX + c$ et $B := \lambda X + \mu$.
Solution. Supposons d'abord λ non nul, donc $q = 1$.
Si a est non nul, $p = 2$ et $\mathrm{Res}(A,B) = \mathrm{Res}_{2,1}(A,B) = a\mu^2 - b\lambda\mu + c\lambda^2$.
Si $a := 0$ et $b \neq 0$, $p = 1$ et $\mathrm{Res}(A,B) = \mathrm{Res}_{1,1}(A,B) = b\mu - c\lambda$.
Si $a := b := 0$, donc $c \neq 0$, $p = 0$ et $\mathrm{Res}(A,B) = \mathrm{Res}_{0,1}(A,B) = c$.
Si maintenant $\lambda := 0$, donc $\mu \neq 0$ et $q = 0$, on a $\mathrm{Res}(A,B) = \mathrm{Res}_{p,0}(A,B) = \mu^p$, où, à nouveau, $p := \deg A$ vaut 2, 1 ou 0.

Voici le lien entre $\mathrm{Res}_{p,q}(A,B)$ et $\mathrm{Res}(A,B)$ dans tous les cas :

1. Si $p := \deg A$ et $q := \deg B$, par définition, $\mathrm{Res}_{p,q}(A,B) = \mathrm{Res}(A,B)$.
2. Si $p := \deg A$ et $q > \deg B$, la dernière ligne de $S_{p,q}(A,B)$ n'a qu'un coefficient non nul et, en développant le déterminant par rapport à cette dernière ligne, on voit que $\mathrm{Res}_{p,q}(A,B) = (-1)^p a_p \,\mathrm{Res}_{p,q-1}(A,B)$. En itérant cette relation, on se ramène au cas précédent et l'on obtient : $\mathrm{Res}_{p,q}(A,B) = \big((-1)^p a_p\big)^{q-\deg B} \mathrm{Res}(A,B)$.
3. De même, si $p > \deg A$ et $q := \deg B$, on trouve :
$$\mathrm{Res}_{p,q}(A,B) = (b_q)^{p-\deg A} \mathrm{Res}(A,B).$$
4. Si $p > \deg A$ et $q > \deg B$, la dernière ligne de $S_{p,q}(A,B)$ est nulle et par conséquent $\mathrm{Res}_{p,q}(A,B) = 0$.

Mentionnons encore la formule : $\mathrm{Res}(B,A) = (-1)^{\deg A \deg B} \mathrm{Res}(A,B)$, qui se prouve en permutant les colonnes de $S_{p,q}(A,B)$ pour obtenir $S_{q,p}(B,A)$, et, pour $a, b \in R \setminus \{0\}$, les formules : $\mathrm{Res}(a,B) = a^{\deg B}$ et $\mathrm{Res}(b,A) = b^{\deg A}$.

Considéré comme expression en les coefficients de $A \in R_p[X]$ et de $B \in R_q[X]$, $\mathrm{Res}_{p,q}(A,B)$ est un polynôme en les $p+q+2$ indéterminées a_i, b_j. Mais le résultant $\mathrm{Res}(A,B)$ n'est pas un tel polynôme, car son expression algébrique dépend du degré de A et de B, donc la nullité ou non-nullité de certains coefficients.

Proposition 22. Soit $A = QB + A_1$ une division euclidienne, avec $A_1 \neq 0$. Alors, avec les mêmes notations que précédemment, $\mathrm{Res}(A,B) = b_q^{degA - \deg A_1} \mathrm{Res}(A_1,B)$.

Démonstration. Pour $0 \leqslant k \leqslant q - p = \deg B - \deg A$, remplacer A par $A - cX^k B$ revient à effectuer les opérations élémentaires suivantes sur les colonnes de $S_{p,q}(A,B)$: $C_i \leftarrow C_i - cC_{p+i+k}$, pour $i \in [\![1,p]\!]$ (et dans un ordre quelconque). Ces opérations ne changent pas le déterminant, d'où l'égalité $\mathrm{Res}_{p,q}(A,B) = \mathrm{Res}_{p,q}(A_1,B)$. Il suffit alors d'appliquer les règles déjà énoncées. ∎

On en déduit un algorithme de calcul du résultant des polynômes A et B.

```
Resultant(A,B)
U := A ; V := B ; W := 1 ;
tant que deg V > 0 faire
   ((Q,A1) := diveucl(U,V) ;
    W := W * (-1)^(deg V deg A1) * (cd(V))^(deg U - deg A1) ;
    U := V ; V := R) ;
si V = 0 alors rendre 0 sinon rendre W * V^(deg U) ;
```

La preuve de la terminaison est la même que pour l'algorithme d'Euclide. Des diverses règles ci-dessus, on tire sans peine l'invariant de boucle suivant : $\mathrm{Res}(A,B) = W \,\mathrm{Res}(U,V)$.

Lemme 23. Si $B = (X - \beta)C$, alors $\mathrm{Res}(A, B) = A(\beta)\,\mathrm{Res}(A, C)$.

Démonstration.
Écrivons $A := a_0 + \cdots + a_p X^p$, $B := b_0 + \cdots + b_q X^q$ et $C := c_0 + \cdots + c_{q-1}X^{q-1}$: on a donc $b_i = c_{i-1} - \beta c_i$ pour $i \in [\![0, q]\!]$, en convenant que $c_{-1} := c_q := 0$. On effectue sur $S_{p,q}(A, B)$ les opérations élémentaires suivantes sur les lignes : $L_i \leftarrow L_i + \beta L_{i+1}$, pour i de $p+q-1$ à 1 (dans cet ordre). Dans la matrice obtenue, chaque a_i a été remplacé par $A_i(\beta)$, où $A_i := \sum_{k=i}^{p} a_k X^{k-i}$ (de sorte que $A_0 = A$, $A_p = a_p$ et $A_i = a_i + XA_{i+1}$) ; de même, chaque b_j a été remplacé par c_{j-1}. On effectue ensuite sur la matrice obtenue les opérations élémentaires suivantes sur les colonnes : $C_j \leftarrow C_j - \beta C_{j-1}$ pour j de q à 2 (dans cet ordre). Les c_j restent en place et les $A_i(\beta)$ redeviennent des a_i, sauf ceux de la première colonne, qui ne sont pas affectés.
Sur la matrice finale, les coefficients de la première ligne sont nuls, sauf le premier, qui vaut $A(\beta)$; le mineur correspondant est $S_{p,q-1}(A, C)$. La formule annoncée en découle. ∎

Théorème 24.
Si $A := a(X - \alpha_1)\cdots(X - \alpha_p)$ et $B := b(X - \beta_1)\cdots(X - \beta_q)$, alors :

$$\begin{aligned}\mathrm{Res}(A, B) &= b^p A(\beta_1)\cdots A(\beta_q)\\ &= b^p a^q \prod_{\substack{1\leqslant i\leqslant p\\ 1\leqslant j\leqslant q}} (\beta_j - \alpha_i)\\ &= (-1)^{pq} a^q B(\alpha_1)\cdots B(\alpha_p).\end{aligned}$$

Démonstration. La première formule s'obtient en itérant le lemme. La seconde découle alors de l'égalité $A(\beta) = a \prod_{1\leqslant i\leqslant p} (\beta - \alpha_i)$. La dernière est laissée en exercice au lecteur. ∎

Corollaire 25. Supposons le corps L algébriquement clos. Alors $\mathrm{Res}(A, B) = 0$ si, et seulement si, les polynômes A et B ont une racine commune.

On suppose maintenant que le corps L (qui contient R) est algébriquement clos et de caractéristique nulle. Soit $A = a_0 + \cdots + a_p X^p \in R_p[X]$ un polynôme de degré p.

Définition 9. Le *discriminant* de A est défini par la formule :

$$\mathrm{Dis}(A) := \frac{(-1)^{p(p-1)/2}}{a_p}\,\mathrm{Res}(A, A').$$

En fait, si $A \in R_p[X]$, la dernière ligne de $S_{p,p-1}(A, A')$ a pour seuls coefficients non nuls a_p et pa_p ; le résultant de A et de A' est donc multiple de a_p dans R, et l'on en déduit que $\mathrm{Dis}(A) \in R$. Par définition, $\mathrm{Dis}(A)$ est un polynôme en les a_i, à coefficients entiers et homogène de degré $2p - 2$. Pour que le polynôme A admette des racines multiples, il faut, et il suffit, que A et A' aient des racines communes, autrement dit, que $\mathrm{Dis}(A)$ soit nul. Cela découle également de la factorisation $A = a_p \prod_{i=1}^{p}(X - \alpha_i)$, des égalités

$A'(\alpha_i) = a_p \prod_{j\neq i}(\alpha_i - \alpha_j)$ et des formules du théorème 24, dont on tire :

$$\mathrm{Dis}(A) = a_p^{2p-2} \prod_{1\leqslant i<j\leqslant p} (\alpha_i - \alpha_j)^2.$$

Exemples. Soit $A := aX^2 + bX + c$, avec $a \neq 0$. Alors $\mathrm{Res}(A, A') = a(4ac - b^2)$ et $\mathrm{Dis}(A) = b^2 - 4ac$, qui est bien égal à $a^2(\alpha_1 - \alpha_2)^2$. Soit $A := X^3 + pX + q$. Alors :

$$\mathrm{Res}(A, A') = 4p^3+27q^2 \quad \text{et} \quad \mathrm{Dis}(A) = -(4p^3+27q^2) = \big((\alpha_1-\alpha_2)(\alpha_1-\alpha_3)(\alpha_2-\alpha_2)\big)^2.$$

2.2.2 Résultant et élimination

La proposition suivante généralise le corollaire 25 au cas où le corps L n'est pas supposé algébriquement clos.

Proposition 26. Pour que $\mathrm{Res}(A, B) = 0$, il faut, et il suffit, que A et B aient un facteur commun non constant dans $L[X]$.

Démonstration. Si Δ est le pgcd de A et B et qu'il est non trivial, en posant $P = B/\Delta$ et $Q = -A/\Delta$, on déduit de la proposition 20 de la page 339 que le résultant est nul. Si $\Delta = 1$, l'égalité $AP + BQ = 0$ entraîne $A|BQ$ donc $A|Q$ (lemme de Gauß) ; si $\deg Q \leqslant p - 1$, on a donc $Q = 0$. D'après la proposition 20 de la page 339, le résultant est donc non nul. ∎

Selon les notations de la page 313, pour tous $f, g \in K[X_1, \ldots, X_n]$, nous noterons $\mathrm{Res}_{X_i}(f, g) \in K[X_1, \ldots, \widehat{X_i}, \ldots, X_n]$ le résultant de f et g considérés comme des polynômes en X_i.

Théorème 27 (d'extension, ou de relèvement).
Soient $A, B \in K[X_0, \ldots, X_n]$ de degrés en X_0 $p := \deg_{X_0} A$ et $q := \deg_{X_0} B$. On écrit :

$$A := \sum_{i=0}^{p} a_i X_0^i \quad \text{et} \quad B := \sum_{j=0}^{q} b_j X_0^j,$$

où les $a_i, b_j \in K[X_1, \ldots, X_n]$ et $a_p, b_q \neq 0$. On suppose de plus le corps K algébriquement clos. Pour tout point $(c_1, \ldots, c_n)$ de K^n, on a alors l'équivalence logique suivante : $\mathrm{Res}_{X_0}(A, B)(c_1, \ldots, c_n) = 0$ si, et seulement si,

- $\big(a_p(c_1, \ldots, c_n) = 0 \quad \text{et} \quad b_q(c_1, \ldots, c_n) = 0\big)$
- ou $\big(\exists c_0 \in K \ : \ A(c_0, c_1, \ldots, c_n) = B(c_0, c_1, \ldots, c_n) = 0\big)$.

Démonstration. Le morphisme de $K[X_0, \ldots, X_n]$ associé à la substitution $X_i \mapsto c_i$ $(1 \leqslant i \leqslant n)$ envoie $\mathrm{Res}_{X_0}(A, B)$ sur :

$$\mathrm{Res}_{X_0}(A, B)(c_1, \ldots, c_n) = \mathrm{Res}_{p,q}\big(A(X_0, c_1, \ldots, c_n), B(X_0, c_1, \ldots, c_n)\big).$$

Ce dernier est nul si, et seulement si, l'un des coefficients $a_p(c_1, \ldots, c_n)$, $b_q(c_1, \ldots, c_n)$ est nul ou si ces polynômes ont une racine commune c_0. ∎

Exemples.

1. Soient $A := X^2 + Y^2 - 4$ et $B := XY - 1$. Alors :

$$\mathrm{Res}_Y(A, B) = \det \begin{pmatrix} X^2 - 4 & -1 & 0 \\ 0 & X & -1 \\ 1 & 0 & X \end{pmatrix} = X^4 - 4X^2 + 1.$$

   Pour chaque x tel que $x^4 - 4x^2 + 1 = 0$, les coefficients dominants 1 et x étant non nuls, le théorème garantit l'existence de y tel que (x, y) est zéro commun de A et B. En fait, il suffit de prendre $y := 1/x$. Le cercle $A = 0$ et l'hyperbole $B = 0$ ont donc quatre points communs sur $\mathbb{R}$, d'abscisses $x := \pm\sqrt{2 \pm \sqrt{3}}$.

2. Soient $A := XY^2 - 1$ et $B := XY - 1$. Alors

$$\mathrm{Res}_Y(A, B) = \det \begin{pmatrix} -1 & -1 & 0 \\ 0 & X & -1 \\ X & 0 & X \end{pmatrix} = -X^2 + X.$$

   Pour $x := 1$, les coefficients dominants étant non nuls, le théorème garantit l'existence de y tel que (x, y) est zéro commun de A et B (il suffit de prendre $y := 1/x = 1$). Pour $x := 0$, ces coefficients dominants s'annulent et le théorème ne permet pas de conclure : et il est en effet visible qu'il n'y a aucun tel y.

3. Soient $A := X^2Y^2 - 1$ et $B := XY - 1$. Alors :

$$\mathrm{Res}_Y(A, B) = \det \begin{pmatrix} -1 & -1 & 0 \\ 0 & X & -1 \\ X^2 & 0 & X \end{pmatrix} = 0,$$

   reflétant le facteur commun $XY - 1$.

4. Soient $A := X^2Y^2 - 2$ et $B := XY - 1$. Alors :

$$\mathrm{Res}_Y(A, B) = \det \begin{pmatrix} -2 & -1 & 0 \\ 0 & X & -1 \\ X^2 & 0 & X \end{pmatrix} = -X^2.$$

   Pour l'annuler, on doit prendre $x = 0$; mais alors les coefficients dominants (en Y) de A et B sont nuls, et le théorème ne permet pas de conclure à l'existence d'un zéro commun (x, y)de A et B ; il est d'ailleurs bien évident qu'il n'en existe pas !

Voici, pour finir, quelques unes des *nombreuses* applications de l'élimination.

Nombres algébriques.
Soient α et β deux nombres algébriques. On a donc $A(\alpha) = B(\beta) = 0$ avec $A, B \in \mathbb{Q}[X]$ unitaires, de degrés respectifs $p, q \geqslant 1$. En éliminant X entre $A(X)$ et $B(T - X)$, on obtient un polynôme $C \in \mathbb{Q}[X]$ tel que $C(\alpha + \beta) = 0$, ce qui prouve que $\alpha + \beta$ est algébrique. Prenons par exemple $A := X^3 - 2$ et $B := X^3 - 3$. Le résultant $C(T)$ est :

$$\det\begin{pmatrix} -2 & 0 & 0 & T^3-3 & 0 & 0 \\ 0 & -2 & 0 & -3T^2 & T^3-3 & 0 \\ 0 & 0 & -2 & 3T & -3T^2 & T^3-3 \\ 1 & 0 & 0 & -1 & 3T & -3T^2 \\ 0 & 1 & 0 & 0 & -1 & 3T \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix} = T^9 - 15T^6 - 87T^3 - 125.$$

Implicitation de paramétrage.

On cherche une équation cartésienne de la courbe paramétrée $x(t) := \dfrac{2t}{1+t^2}$, $y(t) := \dfrac{1-t^2}{1+t^2}$.
Il suffit pour cela d'éliminer T entre les équations :

$$A(T) := (1+T^2)X - 2T = XT^2 - 2T + X = 0$$
$$B(T) := (1+T^2)Y - (1-T^2) = (Y+1)T^2 + (Y-1) = 0.$$

On trouve pour résultant : $\det\begin{pmatrix} X & 0 & Y-1 & 0 \\ -2 & X & 0 & Y-1 \\ X & -2 & Y+1 & 0 \\ 0 & X & 0 & Y+1 \end{pmatrix} = 4(X^2+Y^2-1)$; on reconnaît un cercle. Réciproquement, si $x^2+y^2=1$, le théorème d'extension 27 de la page 342 ne garantit l'existence d'un t que si x et $y+1$ ne sont pas tous les deux nuls. Et, en effet, le paramétrage ci-dessus « n'atteint pas » le point $(0,-1)$ (sauf pour $t=\infty$).

Enveloppes. On se donne, pour tout $t \in I$ (intervalle ouvert de $\mathbb{R}$), une droite D_t de $\mathbb{R}^2$, d'équation $a(t)X + b(t)Y + c(t) = 0$. On suppose que a, b, c sont des fonctions de classe $\mathcal{C}^1$ sur I telles que la fonction $ab' - a'b$ ne s'annule pas. On peut alors démontrer les faits suivants (exercice I.7.53 de la page 351) : soit $t \in I$; alors, pour h assez petit, les droites D_t et D_{t+h} admettent un unique point d'intersection $M(t,h)$. Lorsque h tend vers 0, le point $M(t,h)$ tend vers un point $M(t)$ qui est l'unique point d'intersection de la droite D_t et de la droite D'_t d'équation $a'(t)X + b'(t)Y + c'(t) = 0$. L'ensemble des points $M(t)$ est appelé *enveloppe de la famille de droites* $(D_t)_{t\in I}$. C'est une courbe paramétrée ; un paramétrage explicite s'obtient en résolvant, à l'aide des formules, de Cramer le système $a(t)X + b(t)Y + c(t) = a'(t)X + b'(t)Y + c'(t) = 0$. Si les fonctions a, b, c sont de classe $\mathcal{C}^2$, l'enveloppe est une courbe de classe $\mathcal{C}^1$ et sa tangente en $M(t)$ est D_t. Lorsque a, b, c sont des fonctions rationnelles, on peut écrire les équations de D_t et de D'_t sous forme de polynômes en t. En éliminant t, on obtient une équation cartésienne de l'enveloppe. Prenons par exemple pour D la droite qui joint les points $M(0, \sin\theta)$ et $N(\cos\theta, 0)$, où $\theta \in]0, \pi/2[$. On parle de « glissement de l'échelle », car le segment MN est de longueur constante 1, son extrémité M est sur le « mur » $x = 0$ et son extrémité N sur le « sol » $y = 0$. L'équation de la droite D est $\dfrac{X}{\cos\theta} + \dfrac{Y}{\sin\theta} = 1$. Comme rien de tout cela n'est rationnel en θ, on pose $t := \tan\theta/2$, ce qui donne des expressions rationnelles $a(t)$, $b(t)$ et $c(t)$. Nous laissons au lecteur le soin d'achever le calcul (exercice I.7.54 de la page 351).

Exercice type corrigé

I.7.0 Pour toute permutation $\sigma \in \mathfrak{S}_n$, on a défini ([L1], exercices du module d'algorithmique) le tableau des inversions $b(\sigma) := \big(b_n(\sigma), \dots, b_1(\sigma)\big) \in \{0\} \times \{0,1\} \times \cdots \times \{0, \dots, n-1\}$ et démontré que l'application $\sigma \mapsto b(\sigma)$ de $\mathfrak{S}_n$ dans $E := \{0\} \times \{0,1\} \times \cdots \times \{0, \dots, n-1\}$ est une bijection. Le *polynôme générateur associé* est :

$$P_n := \sum_{\sigma \in \mathfrak{S}_n} X_{b_n(\sigma)} \cdots X_{b_1(\sigma)} \in \mathbb{Q}[X_0, \dots, X_n].$$

(i) Calculer P_1, P_2, P_3.
(ii) Déduire de l'exercice évoqué une factorisation de P_n.
(iii) Exprimer de deux manières l'image $Q_n \in \mathbb{Q}[X]$ du polynôme P_n par le morphisme $\mathbb{Q}[X_0, X_1, \dots, X_{n-1}]$ dans $\mathbb{Q}[X]$ associé à la substitution $X_i \mapsto X^i$.
(iv) Calculer et interpréter $Q_n(1)$ et $\dfrac{Q_n'(1)}{Q_n(1)}$.

Solution. (i) L'unique $\sigma \in \mathfrak{S}_1$ a pour tableau d'inversions (0), d'où $P_1 = X_0$. Les deux permutations de $\mathfrak{S}_2$ ont pour tableaux $(0,0)$ (l'identité) et $(0,1)$ (la transposition). On a donc $P_2 = X_0^2 + X_0X_1$. Le lecteur patient vérifiera de même (en étudiant chaque permutation de $\mathfrak{S}_3$) que $P_3 = X_0^3 + X_0^2X_1 + X_0^2X_2 + X_0^2X_1 + X_0X_1^2 + X_0X_1X_2$.
(ii) Comme l'application $\sigma \mapsto b(\sigma)$ est bijective, on a $P_n = \sum X_{i_1} \cdots X_{i_n}$, où les indices $i_k \in \{0, \dots, k-1\}$ varient indépendamment. Les règles de calcul dans un anneau (en fait, la distributivité) permettent donc de factoriser P_n en $\displaystyle\prod_{k=1}^{n} \sum_{i_k \in \{0,\dots,k-1\}} X_{i_k}$.
Finalement, $P_n = X_0(X_0 + X_1) \cdots (X_0 + \cdots + X_{n-1})$.
(iii) La forme développée de P_n donne :

$$Q_n = \sum_{\sigma \in \mathfrak{S}_n} X^{b_n(\sigma)} \cdots X^{b_1(\sigma)} = \sum_{\sigma \in \mathfrak{S}_n} X^{b_n(\sigma) + \cdots + b_1(\sigma)} = \sum_{\sigma \in \mathfrak{S}_n} X^{I_n(\sigma)},$$

où $I_n(\sigma) := b_n(\sigma) + \cdots + b_1(\sigma)$ est le nombre d'inversions de σ.
On peut d'ailleurs regrouper les monômes de même exposant et écrire $Q_n = \sum a_{n,k} X^k$, où $a_{n,k} := \operatorname{card}\{\sigma \in \mathfrak{S}_n \mid I_n(\sigma) = k\}$.
La forme produit de P_n donne : $Q_n = X^0(X^0 + X^1) \cdots (X^0 + \cdots + X^{n-1}) = \displaystyle\prod_{k=1}^{n} (1 + \cdots + X^{k-1})$.
(iv) La première expression donne : $Q_n(1) = \displaystyle\sum_{\sigma \in \mathfrak{S}_n} 1 = \operatorname{card} \mathfrak{S}_n$, et la deuxième : $Q_n(1) = \displaystyle\prod_{k=1}^{n} k = n!$, ce qui est raisonnable ! D'après la première expression : $\dfrac{Q_n'(1)}{Q_n(1)} = \dfrac{1}{n!} \displaystyle\sum_{\sigma \in \mathfrak{S}_n} I_n(\sigma)$, donc le nombre moyen d'inversions d'une permutation de $\mathfrak{S}_n$.
D'après la deuxième expression, par dérivation logarithmique : $\dfrac{Q_n'}{Q_n} = \displaystyle\sum_{k=1}^{n} \frac{0 + \cdots + (k-1)X^{k-2}}{1 + \cdots + X^{k-1}}$, d'où l'égalité : $\dfrac{Q_n'(1)}{Q_n(1)} = \displaystyle\sum_{k=1}^{n} \frac{k(k-1)/2}{k} = \sum_{k=1}^{n} \frac{k-1}{2} = \frac{n(n-1)}{4}$.

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.7.1 L'idéal $<(X_i - X_j)_{1\leqslant i,j\leqslant n}>$ de $K[X_1,\dots,X_n]$ est-il principal ?

I.7.2 $*$ (i) Soit k un entier supérieur ou égal à 2. On note I_k l'idéal de $K[X_1,\dots,X_n]$ engendré par les $X_i^k - X_i$, $i \in [\![1,n]\!]$ et E_k le sous–espace vectoriel de $K[X_1,\dots,X_n]$ engendré par les polynômes dont tous les degrés partiels sont strictement inférieurs à k. Autrement dit, E_k admet pour base la famille des monômes X^α tels que $\alpha_1,\dots,\alpha_n \in [\![0,k-1]\!]$. Démontrer que tout polynôme de $K[X_1,\dots,X_n]$ est congru modulo I_k à un unique élément de E_k. En déduire une base de $K[X_1,\dots,X_n]/I_k$ et sa dimension.
(ii) On note $A := K[X_1,\dots,X_n]/I_2$ et $x_i := X_i \pmod{I_k}$ ($i \in [\![1,n]\!]$) la classe de X_i dans A. Les x_i sont donc des idempotents de A ([L1], module « Fondements »). Pour tout $E \in \mathcal{P}\big([\![1,n]\!]\big)$, on note : $x^E := \prod_{i\in E} x_i$. Démontrer que les x^E forment une base de A. Comment calcule-t-on les produits dans cette base ?

I.7.3 $**$ (i) On reprend les notations de la question (ii) de l'exercice I.7.2. Pour tout $E \in \mathcal{P}\big([\![1,n]\!]\big)$, on note : $u_E := \prod_{i\in E} x_i \prod_{i\notin E}(1-x_i)$. Démontrer que les u_E forment une *famille d'idempotents orthogonaux*, autrement dit : ce sont des idempotents ; leur somme est 1 ; pour $E \neq F$, on a $u_E u_F = 0$. Démontrer que les u_E forment une base de A. En déduire que la projection de $a \in A$ sur la composante Ku_F de $A = \bigoplus Ku_E$ est au_F, puis la formule : $\forall E \in \mathcal{P}\big([\![1,n]\!]\big)$, $x^E = \sum_{E\subset F} u_F$. Exprimer de même chaque u_F dans la base des x^E. Exprimer $(x_1 + \cdots + x_n)^k$ dans les deux bases.
(ii) Proposer une application de cet exercice au « problème du collectionneur de coupons » (dernier exercice sur la première section du module « Polynômes » de [L1]).

I.7.4 Démontrer que Det_n est un polynôme irréductible.

I.7.5 $*$ Calculer le *déterminant de Cauchy* : $\det\left(\dfrac{1}{X_i+Y_j}\right)_{1\leqslant i,j\leqslant n} \in K(X_1,\dots,X_n,Y_1,\dots,Y_n)$.

I.7.6 $*$
Décrire l'idéal des relations différentielles algébriques d'ordre n satisfaites par $\sin$ pour $n = 0, 1, 2$.

I.7.7 Soit $f \in \mathcal{F}(\mathbb{R},\mathbb{R})$ la fonction $x \mapsto e^{x^2}$. Vérifier que f et f' sont algébriquement indépendantes et que $P_0 := X_0X_2 - X_1^2 - 2X_0^2$ appartient à l'idéal I des relations différentielles algébriques d'ordre 2 satisfaites par f.

I.7.8 $**$ Avec les mêmes notations, démontrer que P_0 engendre I.

I.7.9 Étudier la réciproque de l'exercice 6 de la page 323 : l'égalité $\forall t \in K^*\ ,\ P(tX) = t^k P(X)$ implique-t-elle que $P \in S_k$?

I.7.10 Démontrer que, si PQ est homogène, alors P et Q le sont.

I.7.11 On appelle *fraction rationnelle homogène* un quotient de deux polynômes homogènes. Démontrer que les fractions rationnelles homogènes de degré 0 forment, avec la fraction nulle, le sous-corps $K(X_1, \dots, X_n)_0$ de $K(X_1, \dots, X_n)$, engendré sur K par les X_i/X_j.

I.7.12 $*$ Soient $L_1, \dots, L_r$ des éléments linéairement indépendants de S_1 et soit F un polynôme en r indéterminées homogène de degré k. Démontrer que $F(L_1, \dots, L_r) \neq 0$.

I.7.13 $**$
(i) On suppose que $P_1, \dots, P_r \in K[X_1, \dots, X_n]$ sont d'ordre 1 et que $\mathrm{In}(P_1), \dots, \mathrm{In}(P_r) \in S_1$ sont linéairement indépendants. Soient $Q_1, \dots, Q_r \in K[X_1, \dots, X_n]$ d'ordre strictement plus grand que 1. Démontrer que les $P_i + Q_i$ sont algébriquement indépendants.
(ii) Soient $P_1, \dots, P_n \in K[X_1, \dots, X_n]$ des polynômes sans terme constant tels que la substitution $X_i \mapsto P_i$ définisse un automorphisme de $K[X_1, \dots, X_n]$. Démontrer que les P_i sont d'ordre 1 et que les $\mathrm{In}(P_i)$ forment une base de S_1.

I.7.14 Démontrer que les polynômes c_i de la page 315 sont homogènes de degré i.

I.7.15 Que peut-on dire d'un polynôme homogène $P \in K[X, Y, Z]$ tel que $P(X, Y, 1) = 0$?

I.7.16 L'application $Q \mapsto Q(1, X_1, \dots, X_n)$ est-elle la réciproque de $P \mapsto P^h$?

I.7.17 $*$ Soit I un idéal de $K[X_1, \dots, X_n]$. Démontrer que les propriétés suivantes sont équivalentes : (i) L'idéal I admet une famille génératrice formée de polynômes homogènes. (ii) Les composantes homogènes de tout polynôme de I sont dans I. (iii) On a l'égalité : $I = \bigoplus_{k \in \mathbb{N}} (I \cap S_k)$.
On dit alors que I est un *idéal homogène*.

I.7.18 Soient I un idéal de $K[X_1, \dots, X_n]$, $k \in \mathbb{N}$ et $P \in I^k$.
Démontrer que, pour tout $\alpha \in \mathbb{N}^n$ tel que $|\alpha| \leqslant k$, on a $D^\alpha P \in I^{k-|\alpha|}$.

I.7.19
(i) Calculer les n dérivées partielles de chacun des polynômes $\sigma_1, \dots, \sigma_n$, puis calculer les p dérivées partielles de chacun des polynômes $\sigma_i(Q_1, \dots, Q_n)$, où $Q_1, \dots, Q_n \in K[Y_1, \dots, Y_p]$.
(ii) Soient $Q_1, \dots, Q_n \in K[T]$. Calculer la dérivée de $P\big(Q_1(T), \dots, Q_n(T)\big) \in K[T]$.
(iii) Calculer $\dfrac{\partial\big(P(TX_1, \dots, TX_n)\big)}{\partial T}$ et en déduire la formule d'Euler.

I.7.20 $*$ On reprend les notations de la proposition 10 de la page 326.
On note de plus $P \circ Q$ le polynôme $P(Q_1, \dots, Q_n) \in K[Y_1, \dots, Y_p]$.
Notons $\overrightarrow{\operatorname{grad}} P$ le vecteur colonne ${}^t\left(\frac{\partial P}{\partial X_1}, \dots, \frac{\partial P}{\partial X_n}\right) \in M_{n,1}(K[X_1, \dots, X_n])$ et $\overrightarrow{\operatorname{grad}}(P \circ Q)$ le vecteur colonne ${}^t\left(\frac{\partial (P \circ Q)}{\partial Y_1}, \dots, \frac{\partial (P \circ Q)}{\partial Y_p}\right) \in M_{p,1}(K[Y_1, \dots, Y_p])$. On définit enfin la *matrice jacobienne* : $\frac{\partial (Q_1, \dots, Q_n)}{\partial (Y_1, \dots, Y_p)} := \left(\frac{\partial Q_j}{\partial Y_i}\right)_{i,j} \in M_{p,n}(K[Y_1, \dots, Y_p])$. Démontrer la formule : $\overrightarrow{\operatorname{grad}}(P \circ Q) = \frac{\partial (Q_1, \dots, Q_n)}{\partial (Y_1, \dots, Y_p)} \left((\overrightarrow{\operatorname{grad}} P) \circ Q\right)$ (on compose composante par composante).

I.7.21 $**$ Avec les notations de l'exercice précédent, on suppose $p = n$, $Y_i = X_i$ ($i \in [\![1, n]\!]$) et $Q_1, \dots, Q_n \in \mathfrak{M}_0$. Démontrer qu'une condition nécessaire pour que $P \mapsto P \circ Q$ soit un automorphisme de la K-algèbre $K[X_1, \dots, X_n]$ est que l'évaluation en $(0, \dots, 0)$ de $\frac{\partial (Q_1, \dots, Q_n)}{\partial (X_1, \dots, X_n)}$ soit inversible.

I.7.22 Démontrer les formules (valables en caractéristique 0) :

$$p_k(P) = \sum_{|\alpha| = k} \frac{D^\alpha P(0, \dots, 0)}{\alpha!} X^\alpha \quad \text{et} \quad \operatorname{In}(P) = \sum_{|\alpha| = \omega(P)} \frac{D^\alpha P(0, \dots, 0)}{\alpha!} X^\alpha.$$

I.7.23 $*$ Sur un corps quelconque, montrer que l'on peut définir des endomorphismes Δ_α du K-espace vectoriel $K[X_1, \dots, X_n]$ tels que : $P(X_1 + Y_1, \dots, X_n + Y_n) = \sum \Delta_\alpha P(Y_1, \dots, Y_n) X^\alpha$.
Démontrer les relations : $\Delta_\alpha(PQ) = \sum_{\beta + \gamma = \alpha} \Delta_\beta(P) \Delta_\gamma(Q)$ et $\Delta_{\beta+\gamma} = \frac{(\beta + \gamma)!}{\beta!\, \gamma!} \Delta_\beta \circ \Delta_\gamma$.
Comment retrouve-t-on les propriétés de la caractéristique nulle ?

I.7.24 Décrire explicitement l'action de S_n sur $\mathbb{N}^n$ évoquée dans le calcul de la page 329 et vérifier que c'est bien une action. Démontrer que, dans chaque orbite, il y a un unique multiindice α tel que $\alpha_1 \geqslant \alpha_2 \geqslant \cdots \geqslant \alpha_n$.
En déduire la dimension de S'_k, et comparer aux formules obtenues en *loc. cit.*.

I.7.25 $**$ Soient E un K-espace vectoriel de dimension finie et p un projecteur de E. Démontrer l'égalité $\dim \operatorname{Im} p = \operatorname{Tr} p$. On suppose K de caractéristique nulle. Pour tout $P \in S_k$, on pose : $sym_k(P) := \frac{1}{n!} \sum_{\sigma \in S_n} \sigma P$. Démontrer que sym_k est un projecteur de S_k d'image S'_k. Notant $\sigma^{(k)}$ la restriction de Φ_σ à S_k, en déduire la formule : $\dim S'_k = \frac{1}{n!} \sum_{\sigma \in S_n} \operatorname{Tr} \sigma^{(k)}$.
L'appliquer pour retrouver les formules de la page 329.

I.7.26 $*$ En étudiant l'effet de l'action de S_n sur le polynôme discriminant, retrouver une preuve de l'existence d'un morphisme non trivial ε de S_n dans $\{-1, 1\}$ (la signature) et le lien entre $\varepsilon(\sigma)$ et le nombre d'inversions de σ.

I.7.27 $*$ (i) On suppose que le corps K n'est pas de caractéristique 2. Soit $P \in K[X_1, \ldots, X_n]$ un vecteur propre commun à tous les Φ_σ ($\sigma \in S_n$). Notons $\lambda(\sigma) \in K^*$ la valeur propre de l'automorphisme Φ_σ associée au vecteur propre P. Démontrer que l'application $\sigma \mapsto \lambda(\sigma)$ est un morphisme de groupe de S_n dans K^*, donc, soit le morphisme trivial, soit le morphisme signature. Que conclure dans le premier cas ? Dans le deuxième cas, on dit que P est *alterné*.
(ii) Déduire de l'exercice 1 de la page 314 que Δ_n est un polynôme alterné.
(iii) Démontrer que les polynômes alternés sont les polynômes de la forme $P\Delta_n$, où P est un polynôme symétrique. Soit $R = \dfrac{P}{Q}$ une fraction rationnelle symétrique écrite sous forme irréductible. Démontrer que P et Q sont symétriques.

I.7.28 Déduire du théorème fondamental sur les polynômes symétriques une nouvelle méthode de calcul de la dimension de S'_k, et la comparer aux précédentes.

I.7.29 Exprimer Δ_n^2 en fonction des σ_i pour $n := 2$ et $n := 3$.

I.7.30 $**$ Soit $P \in S'_k$ et soit $p_\alpha X^\alpha$ son terme dominant pour l'ordre lexicographique. Démontrer que $\alpha_1 \geqslant \alpha_2 \geqslant \cdots \geqslant \alpha_n$ et que $P - c\sigma_1^{\alpha_1-\alpha_2} \cdots \sigma_n^{\alpha_n}$ est dans S'_k et de multidegré strictement plus petit. En déduire une nouvelle démonstration du théorème 14 de la page 331.

I.7.31 Soient K un corps de caractéristique nulle, $n, p \in \mathbb{N}$ et $\alpha_1, \ldots, \alpha_n \in K$ tels que : $\alpha_1^k + \cdots + \alpha_n^k = 0$ pour $k \in [\![p, p+n-1]\!]$. Démontrer que les α_i sont nuls. Donner un contre-exemple lorsque la caractéristique est 2 et $n := 2$.

I.7.32 $**$ (i) Soit $P \in K[X]$ un polynôme non constant. Démontrer qu'il existe une extension K' de K dans laquelle P admet une racine.
(ii) En déduire l'existence d'une extension L de K sur laquelle P est scindé.
(iii) Application : modifier la preuve du théorème de d'Alembert-Gauß (page 334) pour éviter l'hypothèse d'existence d'un corps algébriquement clos contenant $\mathbb{C}$.

I.7.33 $*$

(i) Soient $A \in M_n(\mathbb{C})$ et $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ ses valeurs propres. On pose $P(T) := \prod_{i=1}^{n} (1 - \lambda_i T)$. Reconnaître le polynôme $T^n P(1/T)$.

(ii) Démontrer l'égalité $-T\dfrac{P'}{P} = \sum_{k \geqslant 1} \mathrm{Tr}(A^k) T^k$. Quel est le rayon de convergence de cette série ? En déduire un procédé de calcul des sommes de Newton des λ_i.

I.7.34 (i) Étendre la proposition 13 de la page 330 au cas de l'anneau $S := A[X_1, \ldots, X_n]$ des polynômes en n indéterminées à coefficients dans un anneau commutatif A.
(ii) Étendre le théorème 16 de la page 333 au cas où P est unitaire et où les a_i appartiennent à un sous-anneau A de K : toute expression polynomiale symétrique en les x_i (à coefficients dans A) est un élément de A.

I.7.35 $*$ On appelle *entier algébrique* un nombre complexe qui est racine d'un polynôme unitaire à coefficients entiers. En s'inspirant de la première application du théorème 16 de la page 333, démontrer que la somme et le produit de deux entiers algébriques sont des entiers algébriques. En déduire que les entiers algébriques forment un sous-anneau du corps des nombres algébriques. Quel est son corps des fractions ?

I.7.36 ∗∗ Soit $P(T) \in \mathbb{Z}[T]$ un polynôme unitaire dont toutes les racines complexes sont non nulles et de module inférieur ou égal à 1. Démontrer que ces racines sont des racines de l'unité (Kronecker).

Exercices sur la section 2

I.7.37 ∗∗ (i) Soit K un corps fini de cardinal q. Démontrer que l'application de $K[X_1, \dots, X_n]$ dans $\mathcal{F}(K^n, K)$ qui, au polynôme $P \in K[X_1, \dots, X_n]$, associe la fonction polynomiale $\tilde{P} : (x_1, \dots, x_n) \mapsto P(x_1, \dots, x_n)$, est un morphisme surjectif d'algèbres de noyau I_q (exercice I.7.2 de la page 346).

(ii) Pour tout $a := (a_1, \dots, a_n) \in K^n$, on pose : $P_a(X) := \prod_{i=1}^{n} \left(1 - (X_i - a_i)^{q-1}\right)$. Démontrer que l'application polynomiale associée est, avec la notation de Kronecker, l'application $x \mapsto \delta_{a,x}$ qui vaut 1 en a et 0 ailleurs. En déduire une nouvelle preuve de la surjectivité.

I.7.38 ∗∗ (i) Soient $N \in \mathbb{N}^*$ un entier et K un corps de caractéristique nulle (donc tel que les images des entiers $1, \dots, N \in \mathbb{N}$ dans K sont distinctes). Soit $P \in K[X]$ un polynôme non nul de degré $d \geqslant 1$. Démontrer que le nombre de points d'annulation de $\tilde{P}$ dans le pavé $[\![1, N]\!]^n \subset K^n$ est majoré par dN^{d-1}.

(ii) On propose l'algorithme suivant pour tester si un polynôme $P \in K[X_1, \dots, X_n]$ de degré inférieur ou égal à d est nul : on choisit $a_1, \dots, a_n \in [\![1, N]\!]$ et l'on calcule $P(a_1, \dots, a_n) \in K$. Si l'on trouve 0, on déclare P nul, sinon, on le déclare non nul. Que pensez-vous de cette méthode ?

I.7.39 ∗ Soit $P \in K[X_1, \dots, X_n]$ un polynôme tel que la fonction associée $\tilde{P}$ est symétrique. Si le corps K est infini, démontrer que P est un polynôme symétrique. Si le corps K est fini, démontrer qu'il existe un polynôme symétrique $Q \in K[X_1, \dots, X_n]$ tel que $\tilde{Q} = \tilde{P}$. (Examiner l'action de $\mathfrak{S}_n$ sur un système de représentants pour la congruence modulo I_q, où $q := \operatorname{card} K$.) Donner un contre-exemple à la première affirmation.

I.7.40 Montrer que toute partie dense de $\mathbb{R}^n$ ou $\mathbb{C}^n$ est Zariski-dense.

I.7.41 ∗ (i) On suppose K algébriquement clos. À l'aide du polynôme Sep introduit dans l'exemple de la page 335, démontrer que l'ensemble des matrices diagonalisables de $M_n(K)$ est Zariski-dense.

(ii) Notons χ_A le polynôme caractéristique de la matrice A. Vérifier que la matrice $\chi_A(A)$ est nulle lorsque la matrice A est diagonalisable.

(iii) En déduire le théorème de Cayley-Hamilton. En admettant que tout corps commutatif est inclus dans un corps algébriquement clos, ou à l'aide de l'exercice I.7.32 de la page précédente, étendre ce résultat à un corps quelconque.

I.7.42 ∗∗ (i) Notons S la K-algèbre $K[X_{1,1}, \dots, X_{n,n}]$ et $M := (X_{i,j})_{1 \leqslant i,j \leqslant n} \in M_n(S)$. On fait agir le groupe $G := GL_n(K)$ sur la K-algèbre S de la manière suivante : pour tout $P \in GL_n(K)$, notons temporairement $X'_{i,j} \in S$ les coefficients de la matrice PMP^{-1} ; ce sont donc des formes linéaires en les $X_{i,j}$; on note alors Φ_P l'endomorphisme associé à la substitution $X_{i,j} \mapsto X'_{i,j}$ $(1 \leqslant i, j \leqslant n)$. Vérifier que l'on a ainsi défini une action de G sur S et

comparer ses propriétés à celles de l'action de $\mathfrak{S}_n$ sur $K[X_1, \ldots, X_n]$.
(ii) Démontrer que les polynômes c_i de l'exemple de la page 315 appartiennent à la sous-algèbre S^G des polynômes invariants sous cette action.
(iii) Formuler et démontrer un théorème analogue au théorème 14 de la page 331, dans lequel les c_i jouent le rôle des polynômes symétriques élémentaires.

I.7.43 Soient $A := a_0 + \cdots + a_p X^p$ ($a_p \neq 0$) et $B := \lambda X + \mu$ ($\lambda \neq 0$).
Démontrer la formule : $\mathrm{Res}(A, B) = \sum_{i=0}^{p} (-1)^i a_i \lambda^{p-i} \mu^i$.

I.7.44 $*$ Retrouver les propositions 20 de la page 339 et 21 de la page 339 en considérant la relation $AP + BQ = 0$ comme un système d'équations en les coefficients de P et Q.

I.7.45
(i) Soit $\varphi : R \to R'$ un morphisme d'anneaux intègres. Notons $\Phi(A)$ et $\Phi(B)$ les images respectives de $A \in R_p[X]$ et de $B \in R_q[X]$ dans $R'_p[X]$ et dans $R'_q[X]$. Comparer $\mathrm{Res}_{p,q}\big(\Phi(A), \Phi(B)\big)$ à $\varphi\big(\mathrm{Res}_{p,q}(A, B)\big)$. Comparer de même $\mathrm{Res}\big(\Phi(A), \Phi(B)\big)$ à $\varphi\big(\mathrm{Res}(A, B)\big)$. Même question pour $\varphi\big(\mathrm{Dis}(A)\big)$ et $\mathrm{Dis}\big(\Phi(A)\big)$.
(ii) Que valent $\mathrm{Res}_{p,q}(\lambda A, \mu B)$, $\mathrm{Res}(\lambda A, \mu B)$ et $\mathrm{Dis}(\lambda A)$, où $\lambda, \mu \in R$ sont non nuls ?

I.7.46 $*$ Montrer que $\mathrm{Res}_{p,q}(A, B)$ est un polynôme homogène en les coefficients de A et de B, et préciser son degré. Montrer qu'il est même *bihomogène*, en un sens que l'on explicitera, et préciser son bidegré.

I.7.47 Quel rapport y a-t-il entre le polynôme discriminant de l'exercice 1 de la page 314 et le discriminant de la définition 9 de la page 341 ?

I.7.48 $*$ Démontrer la formule : $\mathrm{Dis}(AB) = \mathrm{Dis}(A)\,\mathrm{Dis}(B)\big(\mathrm{Res}(A, B)\big)^2$.

I.7.49 On appelle *fenêtre de Viviani* l'intersection de la sphère $x^2 + y^2 + z^2 = 1$ et du cylindre $x^2 - x + y^2 = 0$. Décrire ses projections sur les trois plans de coordonnées.

I.7.50 Déterminer l'équation cartésienne de la surface paramétrée : $x = uv$, $y = v$, $z = u^2$ (*parapluie de Whitney*). Tous les points qui satisfont à l'équation cartésienne proviennent-ils de la surface de départ ?

I.7.51 $*$ Éliminer t dans les équations $x = t^n$, $y = t^p$, où $n, p \in \mathbb{N}^*$.

I.7.52 $*$ Démontrer par élimination que le produit de deux nombres algébriques est algébrique. Appliquer à $\sqrt{3} + \sqrt{2}$ et $\sqrt{3} - \sqrt{2}$ (ainsi, la vérification sera aisée).

I.7.53 $*$ Démontrer les propriétés de $M(t)$ et de l'enveloppe énoncées lors de l'étude des enveloppes, page 344

I.7.54 Achever les calculs concernant l'échelle glissante (page 344).

I.7.55 $**$ Calculer l'enveloppe des rayons réfléchis des rayons solaires par le bord d'un bol circulaire (« caustique »). Pour tout point $M(\cos\theta, \sin\theta)$ avec $\theta \in]-\pi/2, \pi/2[$, le rayon solaire est supposé horizontal : $y = \sin\theta$; le rayon réfléchi est le symétrique du rayon solaire par rapport au rayon du cercle passant par M.

Structures discrètes et récursivité

I.8

Les mathématiciens qui s'intéressent à la réalisabilité de leurs constructions et, plus récemment, les informaticiens qui veulent comprendre la structure de leurs programmes, ont mis en place un ensemble d'outils et de théories dont nous abordons quelques exemples fondamentaux dans ce module. Nous nous contenterons d'effleurer ce domaine immense, laissant en particulier de côté l'algorithmique « fine » (reconnaissance de motifs, parcours de graphes…) ; mais nous aborderons la programmation des algorithmes (des programmes en CAML seront fournis avec la solution des exercices de fin de chapitre).
Dans tout ce module, on notera $|E|$ le nombre d'éléments d'un ensemble fini E.

1 Mots et langages

Toute théorie mathématique peut être écrite et tout programme doit être écrit dans un *langage formel*. Une partie de la logique et une partie de l'informatique s'occupent donc des règles de *production* et de *reconnaissance* des objets des langages formels.
Dans un même texte, plusieurs niveaux de langage peuvent être imbriqués : la phrase « qui, demain, mangera le loup ? » (ou l'agneau) met en jeu, au niveau le plus bas, la production (et la reconnaissance) de *mots* et, au niveau supérieur, la production (et la reconnaissance) de *phrases* ; de même, lorsque l'on écrit $23, 1 + 35, 1 = 58, 2$, on rencontre, au niveau le plus bas, des *nombres décimaux*, au niveau supérieur une *expression arithmétique* (le membre gauche) et, au niveau le plus élevé, une *relation* (ici, une égalité). Nous nous concentrerons sur un seul niveau de langage, et ses éléments seront appelés *mots*.
Autre complication : pour étudier les langages formels, nous aurons besoin … d'un langage ! Il y a donc risque de confusion entre les symboles du langage que nous étudions et ceux du langage qui nous sert à cette étude, que l'on appelle parfois *métalangage*. Certains symboles seront donc réservés au métalangage et nous tâcherons de les éviter dans les exemples. (Il s'agit principalement de notations de la théorie des ensembles.)

Une méthode typique de la logique et de l'informatique pour construire des objets nouveaux est la *définition inductive*. Supposons que l'on veuille décrire l'ensemble L_0 de tous les mots ne comportant que la lettre a (et peut-être même aucune lettre). On pourrait dire :
(i) le « mot vide » ε est dans le langage L_0 (cas de base) ;
(ii) si le mot w est dans L_0, alors aw est dans L_0 (étape d'induction) ;
(iii) rien d'autre n'est dans L_0.
Cette dernière règle (ou plutôt « méta-règle ») dit le mode d'emploi des deux premières : un mot est valide si, et seulement s'il peut être obtenu par un nombre fini d'applications des règles (i) et (ii). Voici une autre description, qui relève de la *grammaire formelle*. Dans un certain (méta)langage, on peut écrire la *règle de production* : $S \to aS \mid \varepsilon$. Son mode d'emploi est le suivant : on part du symbole S ; on lui substitue, au choix, soit le mot aS soit le mot vide ε, et cela, tant que l'on veut et peut. On n'a le droit (et le devoir) d'arrêter que lorsque l'on a un mot où ne figure plus S. Voici une suite possible de telles transformations : $S \to aS \to aaS \to aa\varepsilon = aa$. Ces règles permettent donc de produire le mot aa ; plus généralement, les mots que l'on peut ainsi produire sont tous les mots entièrement formés de a (y compris le mot vide). En abrégé, le *langage produit par cette grammaire* est $L_0 := \{a^n \mid n \in \mathbb{N}\}$.
Il est alors facile de reconnaître si un mot arbitraire appartient ou non à L_0 : on lit tous ses caractères, de gauche à droite et un seul à la fois, jusqu'à épuisement du texte analysé (on n'admet donc que des mots *finis*) ; si l'on n'a lu que des a, le mot est accepté (succès de la reconnaissance), sinon, il est refusé (échec de la reconnaissance).

Exercice 1.
Quel langage produit-on avec la règle $S \to aSb \mid \varepsilon$? Comment reconnaître les mots de ce langage ? Quelle différence profonde y a-t-il avec l'exemple de L_0 ?
Solution. Le langage produit est $L_1 := \{a^n b^n \mid n \in \mathbb{N}\}$. Pour reconnaître un mot de L_1, on lit des a en les comptant. À la fin des a, si l'on tombe sur un autre caractère que b, il y a échec. Ensuite, on lit des b en les comptant. À la fin des b, si l'on tombe sur un caractère (quel qu'il soit), il y a échec. Sinon, on est à la fin du mot, et l'on teste si les deux nombres trouvés sont égaux. Il nous faut donc gérer des compteurs de taille imprévisible : contrairement au précédent, cet algorithme utilise une mémoire *non bornée*.

Nous visons ici l'étude de la classe des langages que l'on peut reconnaître avec une mémoire bornée (*langages rationnels*, voir le paragraphe 1.2). Cette classe contient des exemples plus riches que L_0, et qui ont leur importance dans la pratique et dans la théorie .

Le lecteur est également invité à réfléchir à l'exemple du langage engendré par la règle $S \to aSbS \mid \varepsilon$. Voici une production : $S \to a\underline{S}bS \to aaSbSb\underline{S} \to aaSbSbaSbS \to aabbab$. Dans les deux premières *dérivations*, on a souligné l'occurrence de S qui allait être remplacée ; dans la dernière dérivation, on a remplacé simultanément les trois occurrences par le mot vide. Pour comprendre ce langage, il faut penser que a et b codent des parenthèses ouvrante et fermante : c'est le langage D des « chaînes de parenthèses bien formées », ou *langage de Dyck* ; mais comment reconnaître ses mots ? Cela apparaîtra dans le cours de ce module.

1.1 L'algèbre des mots

Nous partons d'un ensemble fini non vide X, appelé *alphabet*. Ses éléments sont appelés, au choix, *caractères*, ou *symboles* ou *lettres*, mais également *chiffres*, *atomes*, *lexèmes* et, en anglais, *token*, qui signifie « jeton », mot très neutre. Notons $|X|$ le nombre des éléments de X. Lorsque $|X| = 1$, presque tout ce qui va suivre est trivial. Mais, dès que $|X| = 2$, presque toutes les difficultés de la théorie apparaissent. Nous prendrons donc le plus souvent comme exemple $X := \{a, b\}$.

Rappelons qu'une suite $(u_0, u_1, \ldots) \in X^{\mathbb{N}}$ à valeurs dans X est tout simplement une application $u : \mathbb{N} \to X$, où l'on note u_i l'image $u(i)$ de l'indice $i \in \mathbb{N}$. De même, nous appellerons *suite finie de longueur* $n \in \mathbb{N}$ d'éléments de X une application $u : [\![0, n-1]\!] \to X$; nous la noterons $(u_0, \ldots, u_{n-1})$ et l'identifierons à un élément de X^n. Lorsque $n := 1$, on identifie u à un élément de X et X^1 à X. Lorsque $n := 0$, on conviendra de noter ε l'unique élément de X^0, c'est-à-dire l'unique suite de longueur 0 d'éléments de X.

Définition 1. On appelle *mot de longueur* n sur l'alphabet X une suite finie de longueur n de caractères de X. L'unique mot de longueur 0 est le *mot vide* ε. Les mots de longueur 1 sont identifiés aux caractères. Les caractères (ou les lettres) du mot $u := (u_0, \ldots, u_{n-1})$ sont $u_0, \ldots, u_{n-1}$; sa première lettre est donc $u(0)$. La longueur de u est notée $|u| := n$. L'ensemble $\bigcup_{n \in \mathbb{N}} X^n$ de tous les mots sur X est noté X^*.

Les mots ayant au moins deux lettres sont notés par juxtaposition de leurs lettres. Par exemple, le mot de quatre lettres défini par $0 \mapsto b$, $1 \mapsto b$, $2 \mapsto a$, et $3 \mapsto b$ est écrit $bbab$. Puisque chaque ensemble X^n est fini (il y a précisément $|X|^n$ mots de n lettres), l'ensemble X^* est dénombrable ([L1], module « Fondements »). *Il est donc impossible de dénoter par des mots différents tous les nombres réels*, et ce, quel que soit le langage employé. Cette affirmation « métamathématique » établit, par des moyens très simples, une limite de toute théorie mathématique !

Exercice 2.

Définir rigoureusement le *miroir* $\tilde{u}$, obtenu en renversant l'ordre des lettres du mot $u \in X^*$

Solution. C'est l'application de $[\![0, n-1]\!]$ dans X telle que $i \mapsto u(n-1-i)$, où $n := |u|$.

Nous munissons maintenant l'ensemble X^* d'une loi de composition interne : la *concaténation* $u.v$ (ou uv) des mots $u, v \in X^*$ de longueurs $n, p \in \mathbb{N}$ est le mot w de longueur $n + p$ défini par :

$$\forall i \in [\![0, n+p-1]\!] \ , \ w(i) := \begin{cases} u(i) \text{ si } 0 \leqslant i \leqslant n-1, \\ v(i-n) \text{ si } n \leqslant i \leqslant n+p-1. \end{cases}$$

Théorème 1. La concaténation est une loi de composition interne sur X^* dont le mot vide ε est élément neutre (à gauche et à droite) et qui est associative : on dit alors qu'elle fait de X^* un *monoïde* ; ce monoïde particulier est appelé *monoïde libre sur l'alphabet* X. Il est commutatif si, et seulement si, $|X| = 1$.

Démonstration. Nous laisserons au lecteur le soin de démontrer que ε est neutre (*i.e.* $u.\varepsilon = \varepsilon.u = u$) et établirons l'associativité. Soient $u, v, w \in X^*$ de longueurs n, p, q et $i \in [\![0, n+p+q-1]\!]$:

$$(u.v).w(i) = \begin{cases} (u.v)(i) \text{ si } 0 \leqslant i \leqslant n+p-1, \\ w\big(i-(n+p)\big) \text{ si } n+p \leqslant i \leqslant n+p+q-1, \end{cases}$$

$$= \begin{cases} u(i) \text{ si } 0 \leqslant i \leqslant n-1, \\ v(i-n) \text{ si } n \leqslant i \leqslant n+p-1, \\ w\big(i-(n+p)\big) \text{ si } n+p \leqslant i \leqslant n+p+q-1. \end{cases}$$

Un calcul similaire donne les mêmes formules pour $u.(v.w)(i)$. On a donc bien $(u.v).w = u.(v.w)$, que l'on notera uvw. Si $|X| = 1$, les mots uv et vu sont tous deux l'unique application (constante) de $[\![0, n+p-1]\!]$ dans X. Si $a, b \in X$ sont distincts, les mots de 1 lettre a et b ont comme produits ab et ba qui diffèrent (ils n'ont pas la même première lettre).

En concaténant les mots ab et ba on obtient le mot $abba$, de même qu'en concaténant les mots de 1 lettre a,b,b et a (ce qui a un sens, en vertu de l'associativité) : les notations sont donc cohérentes.
Comme dans tout monoïde, on peut poser $u^0 := \varepsilon$, $u^1 := u$ et $u^{n+1} := u^n u$; ainsi, $abba = ab^2a$. Les règles usuelles sur les puissances : $u^{n+p} = u^n u^p$ et $(u^n)^p = u^{np}$ sont alors valides, mais pas la règle sur $(uv)^n$, qui présupposerait la commutativité.

Proposition 2. (i) Soient $u, v \in X^*$. Alors $|uv| = |u| + |v|$.
(ii) Le seul élément inversible de X^* est le mot vide.
(iii) Tous les éléments de X^* sont simplifiables à droite et à gauche.
(iv) On a $u^p = v^p \Leftrightarrow p = 0 \quad \text{ou} \quad u = v$.

Démonstration. La première assertion découle tautologiquement de la définition. Si u a pour inverse v, on a donc $|u| + |v| = 0$, d'où $|u| = |v| = 0$, d'où $u = v = \varepsilon$. Supposons $uw = vw$ (resp. $wu = wv$, resp. $u^p = v^p$ et $p \neq 0$). Alors u et v ont même longueur n et, par définition, pour tout $i \in [\![0, n-1]\!]$, on a : $u(i) = uw(i) = vw(i) = v(i)$ (resp. $u(i) = wu(i+|w|) = wv(i+|w|) = v(i)$, resp. $u(i) = u^p(i) = v^p(i) = v(i)$.)

Théorème 3 (Propriété universelle).
Soit $(M, \star)$ un monoïde. alors toute application $\varphi : X \to M$ s'étend en un unique morphisme de monoïdes $\overline{\varphi} : (X^*, .) \to (M, \star)$.

Démonstration. Notant e le neutre de M, on doit avoir (par définition des morphismes de monoïdes) $\overline{\varphi}(\varepsilon) = e$; pour tout mot d'une lettre, $\overline{\varphi}(x) = \varphi(x)$ (car on requiert que $\overline{\varphi}$ étende φ) ; et, pour un mot de n lettres $w = w_1 \cdots w_n$ ($n \geqslant 2$), $\overline{\varphi}(w) = \varphi(w_1) \cdots \varphi(w_n)$ (car $\overline{\varphi}$ est un morphisme). On a donc bien l'unicité de $\overline{\varphi}$. Réciproquement, en définissant $\overline{\varphi}$ par ces formules, on voit facilement que l'on répond à la question.

Comme le monoïde X^* est engendré par X (voir l'exercice I.8.8 de la page 385), il est clair *a priori* que l'application de restriction $\overline{\varphi} \mapsto \varphi := \overline{\varphi}_{|X}$ est injective. Le fait qu'elle est surjective dit que l'on n'a besoin de poser aucune condition sur les images des lettres, autrement dit, que celles-ci engendrent *librement* X^* : comparer à la possibilité, lorsque E et F sont des espaces vectoriels et $A \subset F$, d'étendre une application $A \to F$ en une application linéaire de E dans F : il y a unicité si, et seulement si, A est génératrice et existence si, et seulement si, A est libre.

Exemple. Soit $a \in X$. L'extension à X^* de l'application $x \mapsto \delta_{a,x}$ de X dans $\mathbb{N}$ est l'application $u \mapsto |u|_a$, nombre d'occurrences de la lettre a dans le mot u.

Définition 2. On dit que le mot u est un *facteur* (resp. un *facteur gauche*, resp. un *facteur droit*) du mot v si $v \in X^*uX^*$ (resp. $v \in uX^*$, resp. $v \in X^*u$).

Rappelons ([L1], module « Fondements », section 4.1) que l'opération de concaténation s'étend comme suit à $\mathcal{P}(X^*)$: si $A, B \subset X^*$, on pose $AB := \{uv \mid u \in A, v \in B\}$. Ainsi, $X^*uX^* = \{u'uu'' \mid u', u'' \in X^*\}$, etc. Autrement dit, u est facteur de v s'il existe des mots u', u'' tels que $v = u'uu''$; u est facteur gauche (on dit aussi *préfixe*) s'il existe un mot u'' tels que $v = uu''$; u est facteur droit (on dit aussi *suffixe*) s'il existe un mot u' tel que $v = u'u$. Il existe également une notion de « sous-mot » (qui équivaut à « suite extraite »), mais nous n'en aurons pas l'emploi.
Les relations « facteur » (resp. « préfixe », resp. « suffixe ») sont des relations d'ordre sur X^* ; ce ne sont pas des ordres totaux (sauf si $|X| = 1$). En revanche, la restriction à $\mathrm{Pref}(v)$ (ensemble des préfixes de v) de la relation « préfixe » est un ordre total : si u_1 et u_2 sont deux préfixes de v, le plus court des deux est préfixe de l'autre. De même, la restriction à $\mathrm{Suff}(v)$ (ensemble des suffixes de v) de la relation « suffixe » est un ordre total. Le *très utile* résultat suivant en découle immédiatement :

Lemme 4. Si $uv = u'v'$, alors il y a une alternative : soit il existe $w \in X^*$ tel que $u' = uw$ et $v = wv'$; soit il existe $w \in X^*$ tel que $u = u'w$ et $wv = v'$.

1.2 Langages

Définition 3. Un *langage* sur l'alphabet X est une partie de X^*.

Ainsi X^* et X sont des langages ; ce dernier est fini, ainsi que $\{\varepsilon\}$, que l'on ne confondra pas avec le langage vide $\varnothing$. Un abus courant de notation consiste à identifier un mot $w \in X^*$ au langage $\{w\}$; par exemple, $\{\varepsilon\}$ est fréquemment noté ε.
Toutes les opérations ensemblistes : $L_1 \cup L_2$, $L_1 \cap L_2$, $L_1 \setminus L_2$ ... sont définies sur les langages, mais il est coutumier de noter la réunion comme une somme : $L_1 + L_2 := L_1 \cup L_2$.

On définit de plus une opération algébrique, la concaténation

$$L_1L_2 := \{w_1w_2 \mid w_1 \in L_1, w_2 \in L_2\}.$$

C'est une opération associative, d'élément neutre $\{\varepsilon\}$, ce qui permet de définir les puissances L^n d'un langage : par convention, $L^0 := \varepsilon$, $L_1 := L$ et, pour $n \geqslant 1$:

$$L^n = \{w_1 \cdots w_n \mid w_1, \ldots, w_n \in L\}.$$

Le langage vide $\varnothing$ est neutre pour l'addition et absorbant pour la concaténation. On a les distributivités : $L(L_1 + L_2) = LL_1 + LL_2$ et $(L_1 + L_2)L = L_1L + L_2L$.
Remarquons que $\varnothing$ et ε jouent, vis à vis de la somme et la concaténation des langages, le même rôle que 0 et 1 vis à vis de la somme et la multiplication des nombres (neutralités, absorption). En conséquence, nous interpréterons ici la fonction caractéristique ([L1], module « Fondements ») comme suit ; si L est un langage et w un mot :

$$\chi_L(w) = \begin{cases} \varepsilon & \text{si } w \in L, \\ \varnothing & \text{si } w \notin L. \end{cases} \tag{1}$$

$$\delta(L) := \chi_L(\varepsilon) = \begin{cases} \varepsilon & \text{si } \varepsilon \in L, \\ \varnothing & \text{si } \varepsilon \notin L. \end{cases} \tag{2}$$

On définit enfin l'*étoile*, ou *itération* de L ; c'est le langage :

$$L^* := \bigcup_{n \geqslant 0} L^n = \{w_1 \cdots w_n \mid n \in \mathbb{N} \quad \text{et} \quad w_1, \ldots, w_n \in A\}.$$

On vérifie facilement que l'étoile de X est bien ce que nous avons noté jusqu'ici X^* ! On a les égalités : $L^* = \varepsilon + LL^* = \varepsilon + L^*L$.

Exemples.

1. Sur l'alphabet $X := \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, le langage des écritures décimales propres des entiers est $0 + \{1, 2, 3, 4, 5, 6, 7, 8, 9\}\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}^*$.
2. Si $X = \{a, b\}$, on a $(a + b)^* = X^*$.

   Les mots qui contiennent deux lettres consécutives égales forment le langage $X^*(aa + bb)X^*$, dont le complémentaire est $(\varepsilon + a)(ba)^* + (\varepsilon + b)(ab)^*$.

Définition 4. On dit que le langage L est *propre* si $\varepsilon \notin L$.

Ainsi, l'expression $L_1 + \delta(L_2)L_3$ vaut L_1 si L_2 est propre et $L_1 + L_3$ autrement. Cette notion joue un rôle clé, comme le montreront les applications de la proposition 6.

Lemme 5. Le langage $L_1 + L_2$ est propre si, et seulement si, L_1 et L_2 le sont. Le langage L_1L_2 est propre si, et seulement si, L_1 ou L_2 l'est.

La preuve est immédiate et laissée au lecteur. (Le langage L^* n'est jamais propre.)

Proposition 6 (Lemme d'Arden).
Soient L_1 et L_2 deux langages sur X. L'équation $L = L_1L+L_2$ (d'inconnue L) admet pour plus petite solution le langage $L_1^*L_2$. Si L_1 est propre, c'est la seule solution.

Démonstration.
Soit $L := L_1^*L_2$. Alors $L_1L + L_2 = L_1L_1^*L_2 + L_2 = (L_1L_1^* + \varepsilon)L_2 = L_1^*L_2$: c'est bien une solution. Soit maintenant L une solution quelconque. Alors $L_1L \subset L$ et, en itérant, $L_1^nL \subset L$ pour tout n, donc $L_1^*L \subset L$. Comme $L_2 \subset L$, on a bien $L_1^*L_2 \subset L$. Supposons enfin que L_1 est propre et que L est une solution. Il s'agit de montrer que tout mot de L est dans $L_1^*L_2$. Par l'absurde, soit $w \in L \setminus L_1^*L_2$ de longueur minimale. Comme $w \notin L_2$ par hypothèse, on déduit de l'équation que $w \in L_1L$, donc $w = w_1w'$ où $w_1 \in L_1$ et $w' \in L$. Comme L_1 est propre, $|w'| < |w|$ et (hypothèse de minimalité) $w' \in L_1^*L_2$; mais cela entraîne que $w = w_1w' \in L_1^*L_2$, contradiction. (Pour le cas où L_1 n'est pas propre, voir l'exercice I.8.12 de la page 386.) ∎

Proposition 7. On se donne un système d'équations d'inconnues $L_1, \ldots, L_n$:

$$\begin{cases} L_1 = A_{1,1}L_1 + \cdots + A_{1,n}L_n + B_1, \\ \cdots\cdots\cdots \\ L_n = A_{n,1}L_1 + \cdots + A_{n,n}L_n + B_n, \end{cases} \tag{3}$$

dans lequel les $A_{i,j}$ sont supposés propres. Alors ce système admet une unique solution $(L_1, \ldots, L_n)$ et les L_i s'expriment *rationnellement* en fonction des $A_{i,j}$ et des B_i, c'est-à-dire à l'aide des seules opérations : somme, concaténation, étoile.

Démonstration. On fait une récurrence sur n. Pour $n := 1$, cela découle du lemme d'Arden. Supposons $n \geqslant 2$. Du lemme d'Arden et de la dernière équation, on tire : $L_n = A_{n,n}^*(A_{n,1}L_1 + \cdots + A_{n,n-1}L_{n-1} + B_n)$. Reportant dans les équations précédentes, on voit que le système est équivalent à la conjonction de cette égalité et du système : $L_1 = A'_{1,1}L_1 + \cdots + A'_{1,n-1}L_{n-1} + B'_1, \ldots, L_{n-1} = A'_{n-1,1}L_1 + \cdots + A'_{n-1,n-1}L_{n-1} + B'_{n-1}$, où l'on a posé : $A'_{i,j} := A_{i,j} + A_{i,n}A_{n,n}^*A_{n,j}$ et $B'_i := B_i + A_{i,n}A_{n,n}^*B_n$. Les $A'_{i,j}$ sont tous propres (lemme 5). La conclusion s'ensuit par récurrence. ∎

Exemple. Notons $L_0 := \{u \in \{a,b\}^* \mid |u|_a \text{ pair}\}$ et $L_1 := \{u \in \{a,b\}^* \mid |u|_a \text{ impair}\}$. En examinant par quelle lettre commence un mot, on voit que $L_0 = aL_1 + bL_0 + \varepsilon$ et $L_1 = aL_0 + bL_1$. De la deuxième équation on tire : $L_1 = b^*aL_0$, puis de la première : $L_0 = (ab^*a + b)L_0 + \varepsilon$, d'où enfin : $L_0 = (ab^*a + b)^*$ et $L_1 = b^*a(ab^*a + b)^*$. Si l'on avait utilisé les équations en ordre inverse, on aurait obtenu : $L_0 = b^*a(ab^*a+b)^*ab^* + b^*$ et $L_1 = (ab^*a+b)^*ab^*$. C'est la même solution, mais décrite par des expressions différentes.

Définition 5. On appelle *résiduel (à gauche) du langage L sur l'alphabet X par le mot $u \in X^*$* le langage : $u^{-1}L := \{v \in X^* \mid uv \in L\}$.

Il y a également une notion de résiduel à droite, mais nous n'en aurons pas besoin. On a alors les formules suivantes :

$$\begin{aligned}
\varepsilon^{-1}L &= L,\\
(uv)^{-1}L &= v^{-1}(u^{-1}L),\\
u^{-1}(L_1 + L_2) &= u^{-1}L_1 + u^{-1}L_2,\\
u^{-1}(L_1 \cap L_2) &= u^{-1}L_1 \cap u^{-1}L_2,\\
u^{-1}(X^* \setminus L) &= X^* \setminus u^{-1}L,\\
\forall x \in X\ ,\ x^{-1}(L_1L_2) &= (x^{-1}L_1)L_2 + \delta(L_1)(x^{-1}L_2),\\
\forall x \in X\ ,\ x^{-1}L^* &= (x^{-1}L)L^*.
\end{aligned}$$

L'intérêt de la deuxième règle est qu'elle permet de calculer les résiduels lettre par lettre. Démontrons l'avant-dernière règle, qui fait appel à la notation introduite dans l'équation (2) page 358. La relation $v \in x^{-1}(L_1L_2)$ équivaut à l'existence d'une décomposition $xv = w_1w_2$, $w_1 \in L_1$, $w_2 \in L_2$. Si $w_1 \neq \varepsilon$, on a donc $w_1 = xw_1'$ et $v = w_1'w_2$ avec $w_1' \in x^{-1}L_1$, donc $v \in (x^{-1}L_1)L_2$. Si L_1 n'est pas propre, il y a également la possibilité que $w_1 = \varepsilon$, et $xv = w_2$, donc $v \in x^{-1}L_2$. Le terme $\delta(L_1)(x^{-1}L_2)$, qui vaut $\varnothing$ ou $x^{-1}L_2$ selon que L_1 est ou non propre, en tient compte. La dernière règle se démontre de manière analogue.

Exemples.

1. Soit $L := a^*b^*$. Alors $a^{-1}L = L$ et $b^{-1}L = b^* =: L'$; puis $a^{-1}L' = \varnothing =: L''$ et $b^{-1}L' = L'$; enfin, $a^{-1}L'' = b^{-1}L'' = L''$. On a donc calculé tous les résiduels de L. Par exemple : $(abba)^{-1}L = (bba)^{-1}L = (ba)^{-1}L' = a^{-1}L' = L''$. Le langage L n'a donc qu'un nombre fini de résiduels, à savoir L, L' et L''.

2. Soit $L := \{a^nb^n \mid n \in \mathbb{N}\}$. On vérifie que $a^{-1}L = \{a^nb^{n+1} \mid n \in \mathbb{N}\}$ et $b^{-1}L = \varnothing$, ce qui conduit à poser, pour tout $i \in \mathbb{N}$: $L_i := \{a^nb^{n+i} \mid n \in \mathbb{N}\}$ (donc $L = L_0$). On a alors : $a^{-1}L_i = L_{i+1}$, $L_i' := b^{-1}L_i = b^{i-1}$, ou $\varnothing$ si $i = 0$ et $b^{-1}L_i' = L_{i-1}'$, ou $\varnothing$ si $i = 0$. Les résiduels ainsi calculés sont deux à deux distincts : par exemple, $b^i \in L_j \Leftrightarrow i = j$. Le langage L a donc une infinité de résiduels.

Proposition 8. Pour tout langage L sur X, on a l'égalité :

$$L = \sum_{x \in X} x(x^{-1}L) + \delta(L).$$

Démonstration. Il est évident que $x(x^{-1}L) \subset L$. Réciproquement, tout mot non vide de L est de la forme xv pour un $x \in X$, et $v \in x^{-1}L$. Le terme $\delta(L)$ prend en compte la possibilité que $\varepsilon \in L$. ∎

Définition 6. On définit inductivement les *langages rationnels (ou réguliers)* sur l'alphabet X comme suit : tout langage fini est rationnel ; si les langages L_1 et L_2 sont rationnels, alors $L_1 + L_2$ et L_1L_2 le sont ; si le langage L est rationnel, alors L^* l'est. La classe des langages rationnels sur X est notée $\mathrm{Rat}(X) \subset \mathcal{P}(X^*)$.

Il revient au même de dire que $\mathrm{Rat}(X)$ est la plus petite partie de $\mathcal{P}(X^*)$ qui contient les langages finis et qui est stable par somme (c'est-à-dire union), concaténation et étoile. Nous avons rencontré pas mal d'exemples de langages rationnels : ce sont ceux que l'on peut décrire par des expressions du type $0 + \{1,2,3,4,5,6,7,8,9\}\{0,1,2,3,4,5,6,7,8,9\}^*$, $(\varepsilon + a)(ba)^* + (\varepsilon + b)(ab)^*$, $(ab^*a + b)^*ab^*$, etc. De telles expressions sont appelées *expressions régulières* et seront étudiées un peu plus formellement à la section 3.3. Il en découle notamment que la classe $\mathrm{Rat}(X)$ est dénombrable ; comme $\mathcal{P}(X^*)$ est équipotent à $\mathbb{R}$ (puisque X^* est dénombrable), la plupart des langages ne sont pas rationnels.

De la définition inductive ci-dessus découle un utile *principe d'induction structurelle pour les langages rationnels*. Soit P un *prédicat* défini sur $\mathcal{P}(X^*)$ (autrement dit, une propriété qui a un sens pour les langages). On suppose P vérifié par les langages finis (cas de base). On suppose également (cas d'induction) que, si P est vérifié par L_1 et L_2, il est vérifié par $L_1 + L_2$ et par L_1L_2 ; et que, s'il est vérifié par L, il est vérifié par L^*. *Alors P est vérifié par tous les langages rationnels*.

Exemple. Démontrons que, si L est rationnel, alors le langage miroir $\tilde{L} := \{\tilde{w} \mid w \in L\}$ l'est aussi. Si L est fini, $\tilde{L}$ est fini ; de plus, on a les égalités : $\widetilde{L_1 + L_2} = \tilde{L_1} + \tilde{L_2}$, $\widetilde{L_1L_2} = \tilde{L_2}\tilde{L_1}$ et $\widetilde{L^*} = \tilde{L}^*$. On conclut à l'aide du prédicat « $\tilde{L}$ est rationnel ».

Notre but est maintenant d'établir un important critère de rationalité ; d'autres critères seront donnés à la section 2, qui expliqueront en quoi les langages rationnels sont bien ceux que l'on peut reconnaître avec une mémoire bornée.

Lemme 9. Notons $R(L)$ l'ensemble des résiduels du langage L et, s'il est fini, $r(L)$ le nombre de ces résiduels. On a alors les inégalités : $r(L_1 + L_2) \leqslant r(L_1)r(L_2)$, $r(L_1L_2) \leqslant r(L_1)2^{r(L_2)}$ et $r(L^*) \leqslant 2^{r(L)} + 1$.

Démonstration. Tout d'abord, on prouve par récurrence sur $|u|$, en partant des formules déjà établies, les égalités suivantes : $u^{-1}(L_1L_2) = (u^{-1}L_1)L_2 + \sum_{\substack{u=u_1u_2 \\ u_1 \in L_1}} u_2^{-1}L_2$, et, pour $u \neq \varepsilon$, $u^{-1}L^* = \sum_{\substack{u=u_1u_2 \\ u_1 \in L^*}} (u_2^{-1}L)L^*$ (pour $u := \varepsilon$, le résiduel est L^*). À l'aide de toutes les formules, on voit alors que l'on a les relations :

$$R(L_1 + L_2) \subset \big\{A_1 + A_2 \mid A_1 \in R(L_1), A_2 \in R(L_2)\big\},$$
$$R(L_1L_2) \subset \{AL_2 + \sum_{B \in R'} B \mid A \in R(L_1), R' \subset R(L_2)\},$$
$$R(L^*) \subset \{L^*\} \cup \Big\{\sum_{A \in R'} AL^* \mid R' \subset R(L)\Big\}.$$

Théorème 10. Un langage est rationnel si, et seulement si, $R(L)$ est fini.

Démonstration. Si le langage L est rationnel, on déduit par induction structurelle du lemme 9 qu'il a un nombre fini de résiduels (le cas de base des langages finis étant évident). Si le langage L a un nombre fini de résiduels $L_1, \ldots, L_n$, on déduit de la proposition 8 de la page 360 que ceux-ci vérifient le système d'équations (3) page 359, où, avec la notation de l'équation (2) page 358, $B_i := \delta(L_i)$ et où $A_{i,j} := \{x \in X \mid x^{-1}L_i = L_j\}$, langage qui peut être vide mais qui est certainement propre. La conclusion découle alors de la proposition 7 de la page 359. ∎

Corollaire 11. L'intersection de deux langages rationnels est un langage rationnel. Le complémentaire d'un langage rationnel est un langage rationnel.

Démonstration. Appliquer les formules de la page 360. ∎

Exemple. D'après les exemples de la page 360, le langage a^*b^* (qui est rationnel) a trois résiduels ; mais le langage $\{a^nb^n \mid n \in \mathbb{N}\}$ en a une infinité et n'est donc pas rationnel.

Exercice 3.
Soient $X := \{a, b\}$ et $L := \{u \in X^* \mid |u|_a = |u|_b\}$. Calculer $L \cap a^*b^*$. Le langage L est-il rationnel ?
Solution. On voit facilement que $L \cap a^*b^* = \{a^nb^n \mid n \in \mathbb{N}\}$, qui n'est pas rationnel. Comme a^*b^* est rationnel, L ne l'est pas d'après le corollaire ci-dessus.

---

2 Graphes

Nous allons décrire la modélisation mathématique et quelques applications élémentaires des graphes. Un graphe est constitué de sommets (ou nœuds) reliés par des arêtes. Par exemple le réseau routier peut être considéré comme un graphe dont les sommets seraient les villes et dont les arêtes seraient les routes reliant ces villes. Il en est de même du réseau SNCF, du réseau téléphonique et du réseau Internet : en première approximation, leurs nœuds sont respectivement les gares, les centraux (et, comme nœuds terminaux, les téléphones) et les serveurs (nœuds terminaux, les PC). Les deux premiers réseaux ne sont pas connexes (ils ne traversent en général pas les mers), les deux derniers le sont.
Lorsqu'un sens de parcours des arêtes est privilégié, on dit que le graphe est orienté (et l'on parle plutôt d'arcs). Par exemple, la succession des tâches dans une activité collective (chantier, système informatique réparti sur plusieurs processeurs) est soumise à des relations de dépendance qui forment un graphe orienté. Les arbres généalogiques sont également des graphes orientés.

2.1 Graphes non orientés

2.1.1 Généralités

Définition 7. Un *graphe non orienté* est un couple (S, A), où S est un ensemble fini non vide (ses éléments sont les *sommets*) et A une partie de l'ensemble $\mathcal{P}_2(S)$ des paires d'éléments de S (les éléments de A sont les *arêtes*). L'arête $a := \{s, t\} \in A$ est dite *incidente* aux sommets s et t, et ceux-ci sont les *extrémités* de a. Le nombre d'arêtes incidentes au sommet s est le *degré* $\deg(s)$ de s. Le sommet s est dit *isolé* si $\deg(s) = 0$ et *pair* (resp. *impair*) si $\deg(s)$ est pair (resp. impair).

Cette définition exclut donc les arêtes reliant un sommet à lui-même ; elle exclut aussi la possibilité d'avoir plusieurs arêtes entre deux sommets donnés.

Exemples. Le *graphe complet* sur S est celui qui a toutes les arêtes possibles : $A := \mathcal{P}_2(S)$. Le *graphe induit* par $G := (S, A)$ sur $S' \subset S$ est $G' := (S', A')$, où $A' := A \cap \mathcal{P}_2(S')$. Un *graphe biparti* est un graphe tel que :

$$S = S_1 \vee S_2 \quad \text{(union disjointe)} \qquad \text{et} \qquad A \subset \big\{\{x_1, x_2\} \mid x_1 \in S_1, x_2 \in S_2\big\}.$$

C'est un *graphe biparti complet* si de plus $A := \big\{\{x_1, x_2\} \mid x_1 \in S_1, x_2 \in S_2\big\}$.

Tout graphe peut être représenté dans $\mathbb{R}^2$ ou dans $\mathbb{R}^3$: les sommets sont figurés par des points, les arêtes par des lignes reliant ces points. Dans $\mathbb{R}^3$, on peut représenter les arêtes par des segments qui ne se rencontrent pas. il suffit de le montrer pour le graphe complet à n sommets, et, pour celui-ci, de le prouver par récurrence. Supposons donc placés n points distincts $A_1, \ldots, A_n$ dans $\mathbb{R}^3$ tels que les segments A_iA_j ne se rencontrent qu'en les A_i. La réunion des $\dfrac{n(n-1)(n-2)}{6}$ plans $A_iA_jA_k$ ne recouvre pas $\mathbb{R}^3$ (c'est un exercice élémentaire de géométrie) et il suffit de prendre A_{n+1} hors de cette réunion. Le lecteur dessinera dans $\mathbb{R}^2$ le graphe complet sur 4 sommets. En revanche, on peut démontrer que le graphe complet sur 5 sommets ou le graphe biparti complet pour lequel $|S_1| = |S_2| = 3$ ne peuvent être réalisés dans $\mathbb{R}^2$, même avec des arêtes courbes, mais cela dépasse le cadre de cet ouvrage.

Proposition 12. Pour tout graphe $G := (S, A)$, on a $\sum_{s \in S} \deg(s) = 2|A|$. Par conséquent, le nombre des sommets impairs de G est pair.

Démonstration. Dans la somme $\sum_{s \in S} \deg(s)$, chaque arête $\{s, t\} \in A$ est comptée deux fois : une fois pour son incidence en s, une fois pour son incidence en t. ∎

Définition 8. Les sommets s et t du graphe non orienté G sont dits *adjacents* ou *voisins* si $\{s,t\} \in A$. Un *chemin de longueur* n de s à t est une suite $s_0,\dots,s_n$ de sommets tels que $s_0 = s$, $s_n = t$ et s_i, s_{i+1} sont adjacents pour $0 \leqslant i \leqslant n-1$. Ce chemin est dit *élémentaire* ou *simple* si les s_i sont deux à deux distincts. Les sommets s et t sont dits *connectés* s'ils sont reliés par un chemin ; c'est évidemment une relation d'équivalence (c'est la clôture réflexive transitive de la relation symétrique de voisinage), dont les classes sont appelées *composantes connexes* du graphe G. Un *circuit de longueur* n est un chemin $s_0,\dots,s_n$ tel que $s_0 = s_n$. Il est dit *simple* ou *élémentaire* si $s_i = s_j$ implique $i = j$ ou $\{i,j\} = \{0,n\}$.

Lemme 13. De tout chemin reliant deux sommets distincts, on peut extraire un chemin élémentaire. La longueur d'un chemin élémentaire est inférieure ou égale à $|S|-1$.

Démonstration. Si $s_0,\dots,s_n$ n'est pas élémentaire, soient $i<j$ tels que $s_i = s_j$. On en extrait alors $s_0,\dots,s_i,s_{j+1},\dots,s_n$ qui est strictement plus court, et l'on réitère tant que c'est nécessaire. Si $s_0,\dots,s_n$ est élémentaire, les s_i sont distincts et $n+1 \leqslant |S|$. ∎

Pour décrire algorithmiquement un graphe non orienté $G := (S,A)$, on convient de renommer les sommets pour avoir $S := \{1,\dots,n\}$. La *matrice d'adjacence* de G est la matrice $M_G := (m_{i,j}) \in M_n(\mathbb{R})$ telle que $m_{i,j} := 1$ si i et j sont adjacents, $m_{i,j} := 0$ sinon. Par définition des graphes non orientés, elle est symétrique et sa diagonale de comporte que des 0 : l'espace n'y est donc pas très bien utilisé.

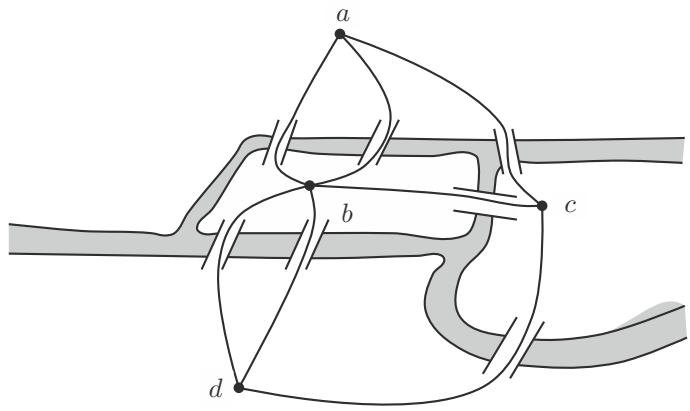
Proposition 14. Le nombre $m_{i,j}(\ell)$ de chemins de longueur l reliant i à j est égal au coefficient d'indices (i,j) de M_G^ℓ.

Démonstration. Pour $k := 0$ et $k := 1$, c'est tautologique. On procède ensuite par récurrence, en utilisant la relation : $m_{i,j}(\ell+1) = \sum_{k=1}^{n} m_{i,k}(\ell) m_{k,j}(1)$. ∎

On utilise parfois la matrice de booléens $\overline{M}_G$ obtenue en remplaçant les 1 par des **`vrai`** et les 0 par des **`faux`**. Les sommes et les produits de telles matrices se font de la manière usuelle, en interprétant la somme comme « ou » et le produit comme « et ». Le même raisonnement que dans la proposition montre que le coefficient d'indices (i,j) de $\overline{M}_G^\ell$ vaut **`vrai`** si i et j sont connectés par un chemin de longueur ℓ et **`faux`** sinon ; de même, le coefficient d'indices (i,j) de $\sum_{k=0}^{\ell} \overline{M}_G^k$ vaut **`vrai`** si i et j sont connectés par un chemin de longueur inférieure ou égale à ℓ et **`faux`** sinon. En vertu du lemme 13, on en déduit un algorithme pour décider si deux sommets sont connectés : on pose $M_0 := \overline{I}_n$, puis $M_{\ell+1} := \overline{I}_n + M_\ell \overline{M}_G$; la matrice M_{n-1} (avec $n := |S|$) contient la réponse. On l'appelle *matrice de connexion* du graphe. Elle est d'ailleurs égale à $(\overline{I}_n + \overline{M}_G)^{n-1}$.

2.1.2 Circuits euleriens

La théorie des graphes a été inventée par Euler pour résoudre le problème des ponts de Königsberg suivant : peut-on parcourir la ville de Königsberg, traversée par le fleuve Pregel et ses multiples bras, en franchissant chacun des 7 ponts une fois et une seule ? Nous allons voir que c'est impossible. Pour cela, modélisons le problème par le graphe de sommets a, b, c et d superposé au plan (simplifié) de la ville.



Comme il y a deux ponts entre l'île de gauche et chacune des deux rives, donc deux arêtes entre le sommet b et chacun des deux sommets a et d, le dessin de droite ne représente pas un graphe, mais un *multigraphe* $G := (S, A)$. Cela signifie que A est un *multiensemble* (en anglais : multiset, ou *bag*) d'arêtes, qui sont encore ici des paires de sommets. Un multiensemble est un ensemble dans lequel chaque élément peut être présent en plusieurs exemplaires : il a donc une multiplicité, qui est un entier. On peut décrire formellement un multisensemble comme un couple (E, f), où E est un ensemble et f une application $E \to \mathbb{N}^*$. Toutes les définitions et tous les résultats antérieurs s'étendent sans problème aux multigraphes.

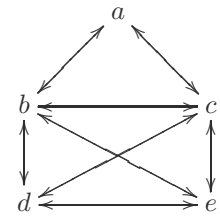
Théorème et définition 15 (Euler).
Un chemin (ou un circuit) du multigraphe $G := (S, A)$ est dit *eulerien* s'il emprunte chaque arête une fois et une seule. Le multigraphe connexe $G := (S, A)$ admet un circuit eulerien si, et seulement si, tous ses sommets sont pairs. Il admet un chemin eulerien reliant les sommets distincts s et t si, et seulement si, ce sont les seuls sommets impairs.

Puisque les quatre sommets du graphe des ponts de Königsberg sont impairs (de degrés respectifs 3, 5, 3 et 3), celui-ci ne comporte aucun chemin eulerien.

Démonstration. Il est facile de déduire la deuxième affirmation de la première (ajouter une arête reliant s et t). Supposant tous les sommets de G pairs, nous allons construire un circuit eulerien partant d'un sommet arbitraire s ; le même argument permet de prouver la réciproque. On part donc de s droit devant soi, en choisissant en chaque sommet visité une arête non encore utilisée, jusqu'à ce l'on arrive en t et que cela ne soit plus possible. Pour chaque sommet visité, le nombre d'entrées et de sorties est égal, sauf peut-être pour s et t ; puisque tous les sommets sont pairs, l'impossibilité de repartir de t signifie que $s = t$: on a donc parcouru un circuit C. On retranche de A les arêtes du circuit C. Si le graphe résultant a des arêtes, chacune de ses composantes connexes a tous ses sommets pairs donc, par hypothèse de récurrence, possède un circuit eulerien C_i ; et chacune de ces composantes connexes possède de plus un sommet s_i commun avec C. On obtient alors un circuit eulerien sur G comme suit : on parcourt C ; chaque fois que l'on passe par un s_i, on fait une

excursion le long de C_i à partir de s_i (un circuit peut être parcouru en prenant n'importe lequel de ses points comme départ et arrivée).

Voici par exemple comment tracer une enveloppe sans doubler les traits et sans lever le crayon : c'est un problème de circuit eulerien ! Les seuls sommets impairs sont d et e. Tout chemin eulerien devra donc partir de l'un et arriver à l'autre. Des chemins possibles sont : $dbacedcbe$ et $dcebcabde$ (il y en a d'autres).



Un chemin (ou un circuit) est dit *hamiltonien* s'il passe par chaque sommet une fois et une seule (naturellement, G doit être connexe !). Contrairement au problème précédent, le problème de l'existence d'un circuit hamiltonien (« problème du voyageur de commerce ») est *très difficile*.

2.1.3 Arbres libres

Plusieurs définitions modélisent l'idée d'arborescence. Nous allons maintenant introduire celle que les géomètres utilisent sous le nom d'arbre, et que l'on nomme en informatique arbre libre. Dans la section 2.3, nous en étudierons des variantes.

Théorème et définition 16. Un *cycle* dans le graphe non orienté G est un circuit élémentaire de longueur au moins égale à 3. Le graphe G est dit *acyclique* s'il n'admet aucun cycle. Les propriétés suivantes sont équivalentes :
(i) Le graphe G est connexe et acyclique.
(ii) Le graphe G est connexe et $|A| = |S| - 1$.
(iii) Le graphe G est acyclique et $|A| = |S| - 1$.
(iv) Quels que soient $s, t \in S$, il existe un unique chemin simple reliant s à t.
On appelle *arbre libre* un graphe vérifiant ces propriétés.

Démonstration. L'équivalence de (i), (ii) et (iii) découle des faits élémentaires énoncés ci-dessous, dont la vérification est un exercice amusant, qui sera en partie laissé au plaisir du lecteur.
Soit tout d'abord G un graphe connexe. Si G comporte un cycle, on peut lui retrancher l'une quelconque des arêtes de ce cycle et il restera connexe. Par suppressions consécutives, on arrive ainsi à un graphe *connexe minimal* (*i.e.* que toute suppression d'arête déconnecte), qui est acyclique. Si l'on retranche une arête d'un graphe connexe acyclique, on le déconnecte (il est donc minimal). Toujours parce que G est connexe, il a au moins $|S| - 1$ arêtes. En effet, on peut supposer G minimal. En lui retranchant une arête $\{s, t\}$, on obtient deux graphes connexes $G_1 := (S_1, A_1)$ et $G_2 := (S_2, A_2)$ et l'on a $|S| = |S_1| + |S_2|$ et $|A| = |A_1| + |A_2| - 1$. On conclut par récurrence sur $|S|$, le cas d'un seul sommet étant trivial. Enfin, de manière duale : tout graphe acyclique peut être complété en un graphe *acyclique maximal*. Un graphe acyclique est maximal si, et seulement s'il est connexe. Un graphe acyclique a au plus $|S| - 1$ arêtes. Prouvons maintenant l'équivalence de (i) et (iv). Supposons G connexe acyclique. Entre s et t, il y a un chemin simple (connexité). S'il y en avait deux, en supprimant leurs arêtes communes, on produirait un ou plusieurs cycles de G. La réciproque se montre de manière analogue.

Proposition 17. (i) Soit $S := [\![1, n]\!]$. Tout ensemble $E \subset \mathfrak{S}_n$ formé de transpositions permet de définir un graphe $G := (S, A)$, où $A := \big\{\{i, j\} \in \mathcal{P}_2(S) \mid \tau_{i,j} \in E\big\}$. Pour que E engendre le groupe $\mathfrak{S}_n$, il faut, et il suffit, que G soit connexe.
(ii) Il faut au moins $(n-1)$ transpositions pour engendrer $\mathfrak{S}_n$.

Démonstration. Supposons G connexe et montrons que toute transposition $\tau_{i,j}$ de $\mathfrak{S}_n$ est un produit de transpositions de E, ce qui suffira d'après le module II.2 (section 2.3) de [L1]. On va le démontrer par récurrence sur la longueur ℓ d'un chemin qui relie i à j dans G. Si $\ell = 0$ ou 1, donc $j = i$ ou j est voisin de i, c'est immédiat. Si $\ell \geqslant 2$, j est voisin de k qui peut être relié à i par un chemin de longueur $\ell - 1$. Alors $\tau_{i,k}$ est un produit T de transpositions de E (hypothèse de récurrence), $\tau_{k,j}$ est dans E par définition, et $\tau_{i,j} = \tau_{k,j}\tau_{i,k}\tau_{k,j}$ est bien produit d'éléments de E. Supposons réciproquement que E engendre $\mathfrak{S}_n$. Pour tout $\tau \in E$ et tout $i \in S$, $\tau(i)$ est égal à i ou est voisin de i. L'image de i par un produit d'éléments de E est donc connectée à i. Comme, pour tout j, la transposition $\tau_{i,j}$ est produit d'éléments de E, tout j est connecté à i.
L'assertion (ii) résulte alors du théorème 16 de la page précédente. ∎

2.2 Graphes orientés

2.2.1 Définitions

Définition 9. Un *graphe orienté* est un un couple (S, A), où S est un ensemble fini non vide (ses éléments sont les *sommets*) et A une partie de l'ensemble $S \times S$ des couples d'éléments de S (les éléments de A sont les *arcs*). L'arc $a := (s, t) \in A$ est dit *incident extérieurement* au sommet s (ou *sortant* ou *partant* en s) et celui-ci est l'*origine* de a ; il est dit *incident intérieurement* au sommet t (ou *entrant* ou *arrivant* en t) et celui-ci est l'*extrémité* de a. Le nombre d'arêtes incidentes intérieurement au sommet s est le *degré entrant* $\deg_e(s)$ de s. Le nombre d'arêtes incidentes extérieurement au sommet s est le *degré sortant* $\deg_s(s)$ de s. Le *degré total* de s est $\deg(s) := \deg_e(s) + \deg_s(s)$.

Cette définition autorise donc les arcs reliant un sommet à lui-même, ou *boucles* ; elle exclut la possibilité d'avoir plusieurs arcs de même origine et même extrémité (voir cependant plus loin la notion de *graphe étiqueté*) ; mais, bien sûr, des arcs opposés (s, t) et (t, s) sont possibles. En analogie avec la proposition 12 de la page 363, on trouve que, pour tout graphe orienté $G := (S, A)$, on a l'égalité : $\sum_{s \in S} \deg_s(s) = \sum_{s \in S} \deg_e(s)$. On appelle *graphe complet* sur S le graphe $G := (S, S \times S)$ et *graphe induit* par $G := (S, A)$ sur $S' \subset S$ le graphe $G' := (S', A')$, où $A' := A \cap (S' \times S')$. La représentation dans $\mathbb{R}^2$ ou $\mathbb{R}^3$ est possible aux mêmes conditions que pour les graphes non orientés, le sens d'un arc étant figuré par une flèche ; bien entendu, la représentation d'une boucle est nécessairement courbe, ainsi que celle de l'un des deux arcs opposés (s, t) et (t, s) si tous deux sont présents.

Définition 10. Si $(s,t) \in A$, on dit que t est un *successeur* de s et que s est un *prédécesseur* de t. Un *chemin de longueur* n de s à t est une suite $s_0, \ldots, s_n$ de sommets tels que $s_0 = s$, $s_n = t$ et s_{i+1} est un successeur de s_i pour $0 \leqslant i \leqslant n-1$. Ce chemin est dit *élémentaire* ou *simple* si les s_i sont deux à deux distincts. Un *circuit* est un chemin d'un sommet à lui-même. Le sommet t est dit *accessible* à partir du sommet s s'il existe un chemin de s à t. On dit que t est *accessible depuis* $E \subset S$ s'il est accessible depuis un élément de E. L'ensemble des sommets accessibles depuis s (resp. depuis E) est noté $\mathrm{Acc}(s)$ $\big($resp. $\mathrm{Acc}(E)\big)$.

Le graphe de la relation binaire « successeur » sur S est donc $A \subset S \times S$. (*Attention : le mot graphe est ici employé dans le sens du module « Fondements » de [L1], section 5.*) La relation d'accessibilité n'est rien d'autre que la clôture réflexive transitive de cette relation ([L1], module « Fondements »). En analogie avec le lemme 13 de la page 364, on trouve que, de tout chemin reliant deux sommets distincts, on peut extraire un chemin élémentaire. La longueur d'un chemin élémentaire est inférieure ou égale à $|S| - 1$. Pour décrire algorithmiquement un graphe orienté $G := (S, A)$, on peut supposer que $S := \{1, \ldots, n\}$. La *matrice d'adjacence* de G est la matrice $M_G := (m_{i,j}) \in M_n(\mathbb{R})$ telle que $m_{i,j} := 1$ si j est un successeur de i, $m_{i,j} := 0$ sinon. En analogie avec la proposition 14 de la page 364, on trouve que le nombre de chemins de longueur ℓ de i à j est le coefficient d'indices (i,j) de M_G^ℓ. On introduit également la matrice de booléens $\overline{M}_G$ obtenue en remplaçant les 1 par des **vrai** et les 0 par des **faux**. Le coefficient d'indices (i,j) de $\overline{M}_G^\ell$ vaut **vrai** s'il y a un chemin de longueur ℓ de i à j, et **faux** sinon ; de même, le coefficient d'indices (i,j) de $\sum_{k=0}^{\ell} \overline{M}_G^k$ vaut **vrai** s'il y a un chemin de longueur inférieure ou égale à ℓ de i à j, et **faux** sinon. On en déduit un algorithme pour décider si j est accessible à partir de i : on pose $M_0 := \overline{I}_n$, puis $M_{\ell+1} := \overline{I}_n + M_\ell \overline{M}_G$; la matrice $M_{n-1} = (\overline{I}_n + M_\ell \overline{M}_G)^{n-1}$ (avec $n := |S|$) contient la réponse. On l'appelle *matrice d'accessibilité* du graphe.

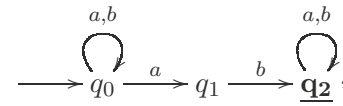
2.2.2 Automates

Généralités

Abstraitement, une machine est une entité qui, partant d'un état d'attente initial, réagit à des stimuli externes en changeant d'état interne et en rendant des résultats. Nous allons considérer une machine qui lit des mots, de gauche à droite et une lettre à la fois ; qui n'a qu'un nombre fini d'états (et, en particulier, une mémoire finie) ; et qui rend un résultat susceptible de deux valeurs : succès (le mot est *reconnu*, ou *accepté*) ou bien échec.

Définition 11.
Un *automate fini sur l'alphabet* X est un quadruplet $\mathcal{A} := (Q, I, F, \Delta)$, où Q est un ensemble fini non vide (les *états*), I et F des sous-ensembles de Q (les états *initiaux* et *finaux*) et Δ un sous-ensemble de $Q \times X \times Q$, la *relation de transition*.

Appelons *graphe orienté étiqueté par un ensemble* X un couple $G := (S, A)$, où S est l'ensemble fini des sommets et $A \subset S \times X \times S$ est l'ensemble des *arcs étiquetés*. Si $(s, x, t) \in A$, on dit que *l'arc* (s, t) *est étiqueté par* x. On peut donc associer à l'automate $\mathcal{A}$ le graphe étiqueté (Q, Δ). Nous noterons $q \xrightarrow{x} q'$ un arc étiqueté $(q, x, q') \in \Delta$, et nous l'appellerons *transition*.
Notre définition autorise expressément que, dans certaines situations, aucune transition ne soit prévue : la machine se bloque, il y a échec de la reconnaissance ; il peut même n'y avoir aucun état initial. . On dit que l'automate $\mathcal{A}$ est *complet* si I est non vide et si : $\forall (q, x) \in Q \times X \ , \ \exists q' \in Q \ : \ q \xrightarrow{x} q'$ (existence d'une transition pour tout symbole lu). Notre définition autorise également l'indéterminisme : dans certains cas, plusieurs transitions sont possibles et l'on ne peut prévoir laquelle aura lieu ; il peut même y avoir plusieurs états initiaux. On dit que l'automate $\mathcal{A}$ est *déterministe*, si $|I| \leqslant 1$ et si, pour tout $\forall (q, x)$, il y a au plus un état q' tel que $q \xrightarrow{x} q'$ (unicité de la transition). Si $|I| = 1$ et si $\forall (q, x) \in Q \times X \ , \ \exists! q' \in Q \ : \ q \xrightarrow{x} q'$ (existence et unicité de la transition), on dit que $\mathcal{A}$ est un *automate fini déterministe complet*, ce que nous abrégerons en AFDC.
Nous ajouterons une *flèche entrante* $\to q$ pour tout état initial $q \in I$, et conviendrons de figurer en **gras souligné** les états finaux $\underline{\mathbf{q'}} \in F$. La dynamique est la suivante : au départ, l'automate $\mathcal{A}$ est dans un état $q_0 \in I$, et il va lire un mot $w \in X^*$. Chaque fois qu'il lit une lettre x et qu'il est dans un état q, il effectue l'une des transitions $(q, x, q') \in \Delta$. S'il n'y en a aucune (c'est-à-dire si $\delta(q, x) = \varnothing$, ce qui ne peut se produire que pour un automate non complet), il y a échec ; sinon, $\mathcal{A}$ *transite* vers l'état q', et l'on continue. C'est, bien entendu, un processus indéterministe, sauf si $\mathcal{A}$ est … déterministe ! À la fin de la lecture du mot, si l'on a atteint un état final, il y a succès (le mot est reconnu), sinon, il y a échec.

Exemple. Soit $\mathcal{A}$ l'automate 

, c'est-à-dire, plus formellement, $X := \{a, b\}$ et $\mathcal{A} := (Q, I, F, \Delta)$, où $Q := \{q_0, q_1, q_2\}$, $I := \{q_0\}$, $F := \{q_2\}$ et $\Delta := \{(q_0, a, q_0), (q_0, b, q_0), (q_0, a, q_1), (q_1, b, q_2), (q_2, a, q_2), (q_2, b, q_2)\}$. Cet automate n'est ni déterministe ni complet. Sa dynamique est décrite plus loin.

Définition 12. Un *calcul* de longueur n dans $\mathcal{A}$, d'origine q et d'extrémité q' est une suite $q_0 := q \xrightarrow{x_0} q_1 \xrightarrow{x_1} \cdots \xrightarrow{x_{n-1}} q_n := q'$ de transitions, que l'on abrège en $q \xrightarrow[\mathcal{A}]{w} q'$, où $w := x_0 \cdots x_{n-1} \in X^*$ est la *trace* de ce calcul (c'est un mot de longueur n). Un calcul *réussi* est un calcul dont l'origine est dans I et l'extrémité dans F. Le *langage reconnu par* $\mathcal{A}$, noté $L(\mathcal{A})$, est l'ensemble des traces de calculs réussis de $\mathcal{A}$:

$$L(\mathcal{A}) := \{w \in X^* \mid \exists q \in I, q' \in F \ : \ q \xrightarrow[\mathcal{A}]{w} q'\}.$$

Un langage L sur X est dit *reconnaissable* s'il existe $\mathcal{A}$ tel que $L = L(\mathcal{A})$. La classe des langages reconnaissables sur X est notée $\mathrm{Rec}(X) \subset \mathcal{P}(X^*)$.

Comme on peut toujours renommer les états d'un automate pour avoir $Q := [\![1, n]\!]$, il est

clair que l'ensemble des automates est (à renommage près) dénombrable, et $\mathrm{Rec}(X)$ est lui-même dénombrable : la plupart des langages ne sont pas reconnaissables.
Il sera commode d'introduire, pour tout $q \in Q$, le langage des « mots reconnaissables depuis q » : $L_q(\mathcal{A}) := \{w \in X^* \mid \exists q' \in F \ : \ q \xrightarrow[\mathcal{A}]{w} q'\}$. On a donc : $L(\mathcal{A}) = \bigcup_{q \in I} L_q(\mathcal{A})$.

Exemple. Dans l'automate de l'exemple de la page 369, la lecture de aa peut provoquer les suites de transitions suivantes : $q_0 \xrightarrow{a} q_0 \xrightarrow{a} q_0$ ou $q_0 \xrightarrow{a} q_0 \xrightarrow{a} q_1$ (et, dans les deux cas, échec, car arrêt sur état non final), ou encore $q_0 \xrightarrow{a} q_1 \to$ ECHEC (car aucune transition sur a n'est alors possible). La lecture de aba peut provoquer les suites de transitions suivantes : $q_0 \xrightarrow{a} q_0 \xrightarrow{b} q_0 \xrightarrow{a} q_0$, ou $q_0 \xrightarrow{a} q_0 \xrightarrow{b} q_0 \xrightarrow{a} q_1$, ou $q_0 \xrightarrow{a} q_1 \xrightarrow{b} q_2 \xrightarrow{a} q_2$, qui est un succès, car arrêt sur un état final. En examinant quels chemins sont possibles, on vérifie les égalités : $L(\mathcal{A}) = L_{q_0}(\mathcal{A}) = X^*abX^*$, $L_{q_1}(\mathcal{A}) = bX^*$ et $L_{q_2}(\mathcal{A}) = X^*$.

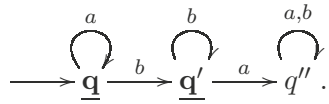
Exercice 4.

À quelle condition $L_q(\mathcal{A})$ $\big($resp. $L(\mathcal{A})\big)$ est-il propre ?

Solution. Seul un calcul de longueur 0 peut avoir pour trace le mot vide, et un calcul de longueur 0 de q à q' n'est possible que si $q = q'$. Ainsi, $L_q(\mathcal{A})$ est propre si, et seulement si, $q \notin F$ et $L(\mathcal{A})$ est propre si, et seulement si, $I \cap F = \varnothing$.

Automate des résiduels

Soit $L \in \mathrm{Rat}(X)$. Notons $L_1 := L, L_2, \ldots, L_n$ ses résiduels (théorème 10 de la page 362). On construit un AFDC en posant $Q := \{1, \ldots, n\}$, $I := \{1\}$, $F := \{i \mid \varepsilon \in L_i\}$ et en prenant pour image de l'état i par la transition de lettre x l'unique j tel que $x^{-1}L_i = L_j$. Cet automate est appelé *automate des résiduels* de L. On prouve par récurrence que l'existence d'un calcul de i à j de trace w équivaut à $w^{-1}L_i = L_j$. Comme $w \in L$ équivaut à $\varepsilon \in w^{-1}L$, on voit que *le langage reconnu par l'automate des résiduels est* L. En particulier, *tout langage rationnel peut être reconnu par un AFDC* : c'est une partie du théorème de Kleene (théorème 19 de la page ci-contre). Soient par exemple $X := \{a, b\}$ et $L := a^*b^*$. On vérifie que $a^{-1}L = L$ et $b^{-1}L = L' := b^*$, puis $b^{-1}L' = L'$ et $a^{-1}L' = L'' := \varnothing$, enfin, que $a^{-1}L'' = b^{-1}L'' = L''$. Le langage L est donc reconnu par l'automate :

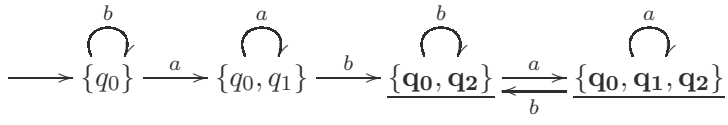


Automate des parties

Soit $\mathcal{A} := (Q, I, F, \Delta)$ un automate arbitraire sur l'alphabet X. L'*automate des parties* de $\mathcal{A}$ est l'automate $\mathcal{A}' := (Q', I', F', \Delta')$ défini par : $Q' := \mathcal{P}(Q)$; $I' := \{I\}$; $F' := \{A \subset Q \mid A \cap F \neq \varnothing\}$; et $\Delta' := \{(A, x, B) \mid \exists a \in A, b \in B \ : \ (a, x, b) \in \Delta\}$. On vérifie aisément que c'est un AFDC : par exemple, l'image de $A \in Q'$ par la transition de lettre x est $B := \{b \in Q \mid \exists b \in B \ : \ (a, x, b) \in \Delta\}$. On prouve par récurrence sur $|w|$ que $A \xrightarrow[\mathcal{A}']{w} B$ équivaut à l'existence de $q \in A$ et $q' \in B$ tels que $q \xrightarrow[\mathcal{A}]{w} q'$. On conclut alors que $L(\mathcal{A}) = L(\mathcal{A}')$:

Théorème 18. Tout langage reconnaissable l'est par un AFDC.

En pratique, on utilise l'AFDC obtenu en restreignant $\mathcal{A}'$ au sous-ensemble $\mathrm{Acc}(I') \subset Q'$, qui reconnaît le même langage : il est en effet clair que seuls les états accessibles apparaissent dans des calculs réussis. C'est ce dernier automate que l'on appelle *déterminisé de* $\mathcal{A}$. Voici par exemple le déterminisé de l'automate de l'exemple de la page 369 :



Théorème 19 (Théorème de Kleene).
Un langage est reconnaissable si, et seulement s'il est rationnel : $\mathrm{Rat}(X) = \mathrm{Rec}(X)$.

Démonstration. Soit $L := L(\mathcal{A})$. Si $w \in L_q$ n'est pas vide, soit x sa première lettre. Alors $w = xw'$, où $w' \in L_{q'}$ pour un q' tel que $q \xrightarrow{x} q'$. Avec la notation de l'équation (1) page 358, on a donc l'équation : $L_q = \sum_{(q,x,q')\in\Delta} xL_{q'} + \chi_F(q)$ (comparer à la proposition 8 de la page 360). Par résolution de ce système (proposition 7 de la page 359), on voit que les L_q sont rationnels, donc également $L = \sum_{q\in I} L_q$.

Si L est rationnel, l'exemple de l'automate des résiduels montre qu'il est reconnaissable. ∎

C'est ce théorème qui justifie notre affirmation que les langages rationnels sont ceux que l'on peut reconnaître avec une mémoire bornée. On peut le préciser de la manière suivante. Soit L un langage, reconnu par un AFDC $\mathcal{A}$, d'état initial q_0. Notons $q_0.w$ l'image de q_0 par la suite de transitions de trace w. Alors $L_{q_0w} = w^{-1}L_{q_0}$. En particulier, $|Q|$ est supérieur ou égal au nombre de résiduels de L. L'automate des résiduels est donc minimal parmi les $AFDC$ reconnaissant $\mathcal{A}$.

Les méthodes décrites ici sont théoriques. Pratiquement, la construction d'un automate reconnaissant le langage décrit par une expression régulière donnée et, réciproquement, le calcul d'une expression régulière dénotant le langage reconnu par un automate donné, se font par d'autres méthodes.

Théorème 20 (Lemme de l'étoile, ou lemme d'Ogden).
Soit L un langage reconnaissable. Il existe alors un entier N tel que tout mot $w \in L$ vérifiant $|w| \geqslant N$ admette un *facteur itérant*, *i.e.* une décomposition $w = u_1vu_2$ telle que $v \neq \varepsilon$ et $u_1v^*u_2 \subset L$.

La réciproque est fausse (exercice I.8.40 de la page 388).

Démonstration. Soit $\mathcal{A} := (Q, I, F, \Delta)$ un automate qui reconnaît L et soit $N := |Q|$. Pour tout mot $w \in L$ de longueur $n \geqslant N$, il y a un calcul réussi :

$$q_0 := q \xrightarrow{x_0} q_1 \xrightarrow{x_1} \cdots \xrightarrow{x_{n-1}} q_n := q'$$

de trace $x_0 \cdots x_{n-1} = w$. Les $n+1 > N$ états q_i ne peuvent être tous distincts ; il existe donc $i < j$ tels que $q_i = q_j$. La partie $q_i \overset{x_i}{\to} \cdots \overset{x_{j-1}}{\to} q_j$ de ce calcul, qui a pour trace le mot $v := x_i \cdots x_{j-1}$, peut être supprimée, ou, au contraire, répétée un nombre indéfini de fois, et l'on aura encore un calcul réussi. Notant $u_1 := x_0 \cdots x_{i-1}$ et $u_2 := x_j \cdots x_{n-1}$, on voit que tous les mots $u_1 v^k u_2$ sont des traces de calculs réussis : on a bien un facteur itérant. ∎

Exemple. Aucun mot du langage $\{a^n b^n \mid n \in \mathbb{N}\}$ n'admet de facteur itérant : ce langage n'est donc pas reconnaissable, ni rationnel.

2.3 Arbres

2.3.1 Arbres enracinés

Nous allons successivement étudier différentes notions d'arbres (« enracinés », « libres », « généraux » et « binaires ») qui apparaissent en informatique.

Proposition et définition 21. Un sommet s_0 du graphe orienté $G := (S, A)$ est une *racine* si, pour tout $s \in S$, il existe un unique chemin de s_0 à s. Une telle racine est alors unique. De plus, chaque sommet $s \neq s_0$ admet un unique prédécesseur et s_0 n'admet pas de prédécesseur. Réciproquement, ces propriétés impliquent que s_0 est racine. Un *arbre enraciné* est un graphe orienté qui a une racine.

Démonstration. Notons, pour abréger, $s_0 \to s$ l'unique chemin de s_0 à s. Pour $s \neq s_0$, en prolongeant $s_0 \to s$ par un chemin de s à s_0, on obtiendrait un chemin de s_0 à lui-même qui ne peut être que le chemin de longueur 0, ce qui est impossible. Ainsi, la racine est unique et n'a pas de prédécesseur. Le sommet qui précède s sur $s_0 \to s$ en est un prédécesseur t ; et, pour tout prédécesseur t' de s, $s_0 \to t'$ prolongé par l'arc (t', s) est $s_0 \to s$, d'où $t' = t$. Réciproquement, si les propriétés sur les prédécesseurs sont vraies, $s_0 \to s$ se construit à l'envers en prenant les prédécesseurs itérés de s. ∎

Nous allons préciser le lien entre arbres libres et arbres enracinés. Soit tout d'abord $G := (S, A)$ un arbre enraciné, de racine s_0. En *oubliant* l'orientation de G, c'est-à-dire en posant $\overline{A} := \big\{\{x, y\} \mid (x, y) \in A\big\}$, on définit un graphe non orienté $\overline{G} := (S, \overline{A})$, qui est un arbre libre. Soient réciproquement $\overline{G} := (S, \overline{A})$ un arbre libre et s_0 un sommet arbitraire. Pour tout sommet $s \in S$, il y a un unique chemin élémentaire de s_0 à s, que nous noterons $s_0 \to s$. Les sens de parcours $s_0 \to s$ sont compatibles entre eux et permettent de définir un graphe orienté $G := (S, A)$, où A est formé des (s, t) tels que $\{s, t\} \in \overline{A}$ et s est sur le trajet $s_0 \to t$. Le graphe G est un arbre enraciné de racine s_0. Une méthode physique pour le réaliser est d'attraper $\overline{G}$ par le sommet s_0 et de laisser pendre le tout ! Ainsi *les deux correspondances ci-dessus sont réciproques l'une de l'autre.*

2.3.2 Arbres infinis

On note $\text{Pere}(s)$ l'unique prédécesseur d'un sommet qui n'est pas la racine et $\text{EnsFils}(s)$ l'ensemble des successeurs de s. On peut donc caractériser un arbre enraciné comme un ensemble fini S muni d'un élément s_0 et d'une application Pere de $S \setminus \{s_0\}$ dans S, vérifiant la propriété d'absence de cycle : il n'existe pas de sommet s et d'entier $k \geqslant 1$ tel que $\text{Pere}^k(s) = s$. Cette description permet de parler d'arbres infinis. Un *arbre (enraciné) infini* est un ensemble infini S muni d'un sommet privilégié (appelé *racine*) $s_0 \in S$ et d'une application $\text{Pere} : S \setminus \{s_0\} \to S$ vérifiant la propriété d'absence de cycle et la propriété de descendance : $\forall s \in S, \exists k \in \mathbb{N} \ : \ \text{Pere}^k(s) = s_0$. On dit que l'arbre est *finitaire* si, pour tout $s \in S$, l'ensemble $\text{EnsFils}(s) := \text{Pere}^{-1}(s)$ est fini. À titre d'exemple d'application du résultat suivant, voir le théorème 24 de la page 381.

Théorème 22 (Lemme de König). Tout arbre infini finitaire admet une *branche infinie* $s_0, s_1, \ldots$ telle que $\forall i \in \mathbb{N}\ ,\ s_{i+1} \in \text{EnsFils}(s_i)$.

Démonstration. Elle repose sur l'axiome du choix ([L1], module « Fondements »).

Appelons *descendance* de $s \in S$, et notons $\text{Desc}(s)$, l'ensemble obtenu par clôture réflexive transitive de la relation de filiation ; ainsi, $\text{Desc}(s_0) = S$, qui est infini. Supposons construits $s_0, \ldots, s_k$ tels que chacun est père du suivant et que chaque $\text{Desc}(s_i)$ est infini. De l'égalité :

$$\text{Desc}(s_k) = s_k \cup \bigcup_{s \in \text{EnsFils}(s_k)} \text{Desc}(s),$$

et de la finitude de $\text{EnsFils}(s_k)$, on déduit qu'il existe $s_{k+1} \in \text{EnsFils}(s_k)$ tel que $\text{Desc}(s_{k+1})$ est infini, ce qui achève la construction. Noter qu'il a fallu prendre l'hypothèse de récurrence de descendance infinie, alors que nous n'avions besoin que de la non vacuité. ∎

2.3.3 Arbres généraux et forêts

Pour comprendre la définition suivante, il faut visualiser un arbre comme une racine s_0 reliée à des sous-arbres. On ordonne le multiensemble $\text{EnsFils}(s_0)$ de ceux-ci ce qui en fait une *liste* $\text{ListeFils}(s_0)$. En effet, en algorithmique, il est naturel d'ordonner les structures de données pour les parcourir.

Définition 13. On appelle *arbre (général)* un couple formé d'une *racine* et d'une forêt. On appelle *forêt* une liste d'arbres (généraux).

Contrairement aux apparences, la définition inductive ci-dessus ne tourne pas en rond. Si l'on convient de noter $R(f)$ l'arbre dont la racine est reliée à la forêt f, on peut commencer par l'arbre le plus simple de tous, pour lequel f est la liste vide $[]$. Pour simplifier, on notera cet arbre R au lieu de $R([])$. On peut maintenant construire des arbres plus compliqués. L'usage des informaticiens est de représenter les filiations de haut en bas et les éléments d'une liste gauche à droite. Le choix d'appeler tous les nœuds R (on appelle en général *nœuds* les sommets d'un arbre) signifie que l'on ne va s'intéresser qu'à la forme

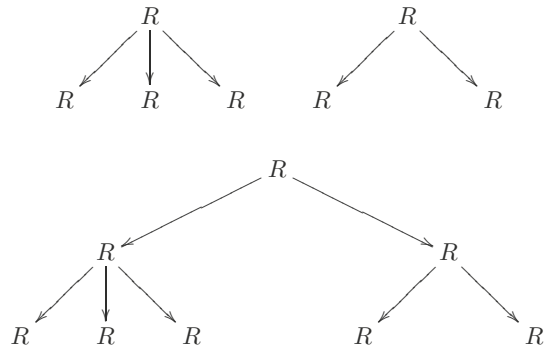
de l'arbre, pas son contenu. Le mot « général » exprime alors que l'on n'impose, pour le moment, aucune restriction aux formes possibles. Par conséquent, la *liste des sous arbres de* $t := R(f)$, que l'on note $\text{ListeSousArbres}(t) := f$ a plus de signification que la liste des fils de sa racine.

Un nœud est dit *interne* s'il admet des successeurs et *externe* s'il n'en admet pas. L'*arité* d'un nœud est le nombre de ses successeurs. Le *niveau* d'un nœud est la longueur du chemin de la racine à ce nœud. La *profondeur* ou *hauteur* $\eta(t)$ d'un arbre t est le plus grand des niveaux de ses nœuds (ou de ses nœuds externes, ce qui revient au même). La *taille* $|t|$ d'un arbre t est le nombre de ses nœuds.

Exemple.

La forêt $l := [R([R;R;R]), R([R,R])]$ représentée ci-contre est composée d'un arbre de taille 4 et de profondeur 1 et d'un arbre de taille 3 et de profondeur 1.

L'arbre $t := R([R([R;R;R]), R([R,R])])$ représenté ci-contre a été obtenu en reliant la racine R à la forêt l. Il a 2 nœuds internes (de niveaux 0 et 1) et 4 nœuds externes (de niveaux 1 et 2). Sa taille est $|t| = 6$ et sa profondeur est $\eta(t) = 2$.



2.3.4 Arbres binaires

Un arbre binaire est un arbre dont tous les nœuds ont pour arité 0 ou 2. Les nœuds internes (d'arité 2) sont appelés *nœuds* et notés N ; les nœuds externes (d'arité 0) sont appelés *feuilles* et notés F. Il est d'usage de ne pas considérer les arbres binaires comme un cas particulier des arbres généraux, mais comme un nouveau type, susceptible d'une définition inductive :

Définition 14.
Un *arbre binaire* est soit une feuille F, soit un arbre t ayant une racine N et deux *sous-arbres* t_1 et t_2 ; on note alors $t := N(t_1, t_2)$, $t_1 = \text{SousArbreGauche}(t)$ et $t_2 = \text{SousArbreDroit}(t)$. La *taille* $|t|$ d'un arbre binaire t est le nombre de ses nœuds.

On peut calculer *inductivement* la hauteur et la taille selon les formules suivantes :

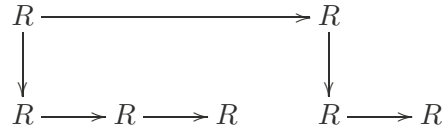
$$|F| = 0 \quad \text{et} \qquad |N(t_1,t_2)| = 1 + |t_1| + |t_2|,$$
$$\eta(F) = 0 \quad \text{et} \quad \eta\big(N(t_1,t_2)\big) = 1 + \max\big(\eta(t_1), \eta(t_2)\big).$$

On a également un *principe d'induction structurelle pour les arbres binaires* : si un prédicat P est vrai pour F et si $\big(P(t_1) \quad \text{et} \quad P(t_2)\big) \Rightarrow P\big(N(t_1,t_2)\big)$, alors P est vrai pour tous les arbres binaires. En voici une illustration.

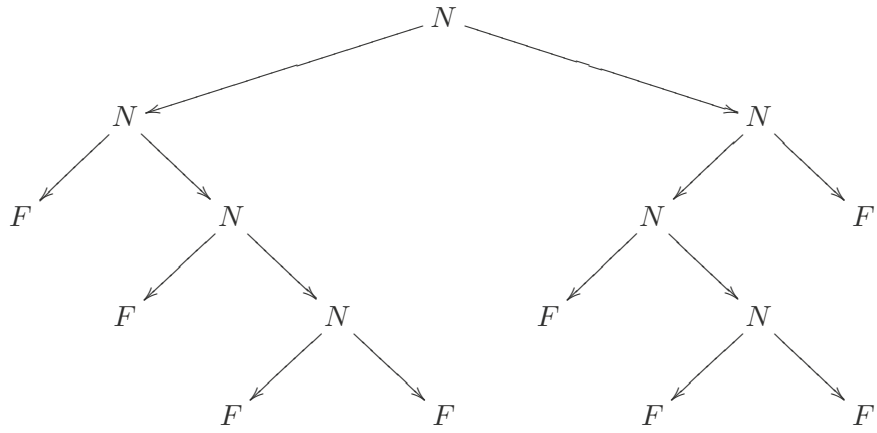
Proposition 23. Pour tout arbre binaire t, le nombre de feuilles est égal à $|t| + 1$.

Démonstration. C'est vrai pour F (taille 0 et 1 feuille) ; si c'est vrai pour t_1 et t_2, c'est vrai pour $N(t_1, t_2)$ en vertu de la formule inductive pour la taille et du fait (évident) que : nombre de feuilles de $N(t_1, t_2) =$ nombre de feuilles de $t_1 +$ nombre de feuilles de t_2. ∎

Si l'on avait défini la taille en spécialisant la définition donnée pour les arbres généraux, on n'obtiendrait que des entiers impairs (d'après la proposition). Il est donc raisonnable de considérer que le paramètre déterminant est le nombre de nœuds (internes).
Il y a un autre lien entre les arbres généraux et les arbres binaires. On peut transformer une forêt en arbre binaire selon la *correspondance de rotation*, également appelée *fils aîné-frère droit*. On considère une forêt comme une fratrie et l'on ne représente que la liaison (verticale) d'un père à son aîné, et les liaisons (horizontales) de chaque frère vers son puîné. La forêt de l'exemple de la page 374 devient alors :



Si l'on fait maintenant tourner la figure obtenue de 45 degrés, que l'on remplace les R par des N et que l'on ajoute les F nécessaires, on reconnaît un arbre binaire :



Le programme récursif qui transforme la forêt $l := [R([R; R; R]), R([R, R])]$ en l'arbre binaire ci-dessus et le programme qui opère la transformation inverse sont écrits page 382.

3 Récursion

Il y a, bien entendu, une théorie des fonctions récursives, mais ce n'est pas notre objet. Nous voulons, par des exemples, montrer comment on écrit de telles fonctions, et comment on raisonne dessus. À l'occasion, des illustrations en Pascal et en CAML seront proposées.

3.1 Exemples expliqués

L'exemple archétypal est la fonction *factorielle* de $\mathbb{N}$ dans $\mathbb{N}$, définie en arithmétique par la formule : $\forall n \in \mathbb{N}\ ,\ n! := \prod_{i=1}^{n} i$; La convention usuelle sur les produits vides donne la valeur $0! := 1$. La relation de récurrence $n! := n(n-1)!$ suggère une itération :

```
fact-it(n)              (* version impérative itérative    *)
    res := 1 ;
    pour i de 1 a n faire res := i * res ;
    rendre(res) ;;
```

Voici une autre méthode :

```
fact(n)                 (* version fonctionnelle récursive *)
si n = 0
    alors rendre(1)
    sinon rendre(n * fact(n-1)) ;;
```

L'idée se résume ainsi : pour calculer $n!$, si $n = 0$, il n'y a rien à faire. Si $n \geqslant 1$, il faut d'abord calculer $(n-1)!$ puis multiplier par n. C'est donc une sorte de régression temporelle, et nous verrons comment un ordinateur la gère. Les différences notables entre les deux versions sont les suivantes :
1) Dans la première version, une variable **res** est déclarée (donc une zone mémoire est allouée) puis transformée par des affectations successives : on agit par *effets de bord*. Dans la deuxième version, les transferts de valeurs se font uniquement par appels de fonctions (et résultats rendus par celles-ci). C'est en cela que la première version est *impérative* et la deuxième *fonctionnelle*.
2) Dans la première version, le contrôle est explicite : c'est la boucle **pour** qui assure la succession des opérations. Dans la deuxième version, c'est l'appel *récursif* de la fonction avec un argument « antérieur » qui implique que $n-1$ est utilisé avant n, etc. C'est en cela que la première version est *itérative* et la deuxième *récursive*.
Le déroulement de la fonction récursive **fact** est le suivant :

```
fact(2) := 2 * fact(1)
        := 2 * (1 * fact(0))
        := 2 * (1 * 1) = 2 * 1 = 2
                    (* calcul final sans appel de fonction *)
```

À « l'aller », il n'y a que des appels de fonctions ; au « retour », les résultats remontent et les calculs sont effectués. On peut imaginer que la fonction **fact** travaille avec des variables locales **n** et **x**. Lorsque l'on appelle **fact(val)**, il se passe les choses suivantes :

```
n := val ; si n = 0 alors rendre(1) ;
x := fact(n-1) ;                        (* ici, n > 1 *)
rendre (n * x) ;;
```

Le malheur est que l'appel de **fact(n-1)** commence par détruire le contenu de **n**, qui ne sera plus le bon au retour. Par exemple, voici ce que donnerait l'appel de **fact(2)** :

```
n := 2 ; si n = 0 rendre 1 (* n'a pas lieu *) ;
         appel de fact(1) ;
n := 1 ; si n = 0 rendre 1 (* n'a pas lieu *) ;
         appel de fact(0) ;
n := 0 ; si n = 0 rendre 1
            (* a lieu, retour a l'appelant fact(1) *) ;
x := 1 ;  rendre(n * x)
            (* c'est-à-dire 0 car n = 0 à ce stade *) ;
x := 0 ; etc
```

Les valeurs nécessaires au calcul sont irrémédiablement perdues. Avant d'appeler **fact(n-1)**, l'incarnation **fact(n)** doit sauvegarder sa zone de travail, c'est-à-dire le contenu de **n** et **x**. La structure de données qui le permet est la pile : c'est comme une liste ou un tableau, mais on y range toujours en tête (ou en haut), et l'on y récupère toujours en tête (ou en haut). Les dernières données rangées sont donc les premières récupérées, comme dans une pile d'assiettes ou de devoirs en retard. Les anglo-saxons disent « LIFO » (last in, first out, que l'on peut traduire par « premier arrivé, premier servi »). Nous conviendrons donc qu'une pile **P** a été définie, que l'instruction **empiler(n,P)** y empile **n** (au sommet, par principe), et que l'instruction **n := depiler(P)** enlève le sommet de la pile **P** et le range dans **n**. Voici maintenant une programmation correcte du déroulement interne de **fact(val)** :

```
n := val ; si n = 0 rendre(1) ;
empiler(n,P) ; x := fact(n-1) ;
n := depiler(P) ; rendre(n * x) ;;
```

Nous allons représenter la dynamique d'exécution de **fact(2)** ; pour simplifier, la pile sera représentée comme une liste, sa tête correspondant au sommet de pile. Au départ, la pile est dans un état indéterminé : **P = [---]**, qui devra être restitué intact à la fin. À droite des instructions exécutées, nous signalons l'état des choses. Pour symboliser l'entrée dans l'exécution d'une fonction imbriquée, nous décalons à droite le texte des instructions.

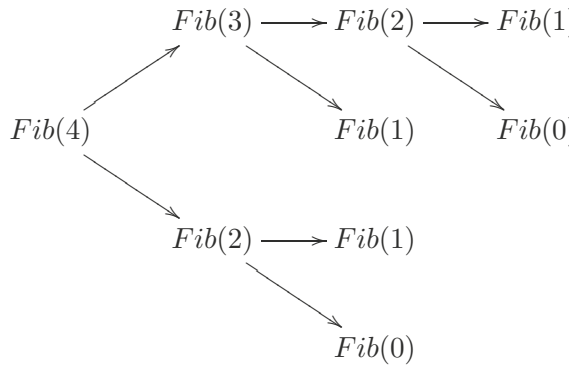
```
n := 2 ; empiler(2,P) ;           (* n = 2, P = [2 ;--] *)
fact(1): n := 1 ; empiler(1,P) ;  (* n = 1, P = [1;2;--]*)
         fact(0): n := 0 ;
                  rendre 1 ;      (* n = 0, P inchangé  *)
         x := fact(0) = 1 ;       (* n et P inchangés   *)
         n := depiler(P) = 1 ;    (* n = 1, P = [2;--]  *)
         rendre n * x = 1 ;       (* n et P inchangés   *)
x := fact(1) = 1 ;                (* n et P inchangés   *)
n := depiler(P) = 2 ;             (* n = 2, P = [--]    *)
rendre n*x = 2 ;;
```

Le résultat rendu est correct, la pile a été restaurée … Nous allons maintenant explorer (mais moins en détail) divers exemples instructifs.

Suite de Fibonacci.
Rappelons qu'elle est définie par $F_0 := 0$, $F_1 := 1$ et $F_n := F_{n-1} + F_{n-2}$, d'où le programme suivant :

```
Fib(n)
si n = 0 alors rendre 0 sinon si n = 1 alors rendre 1
sinon rendre Fib(n-1) + Fib(n-2) ;;
```

On observe alors l'*arbre des appels* figuré à droite. Il y a donc deux appels de **Fib(2)**, trois appels de **Fib(1)** et deux appels de **Fib(0)**, d'où une *redondance* des calculs, qui s'aggrave exponentiellement pour de plus grandes valeurs de l'argument d'appel (exercice I.8.47 de la page 389).



Le problème est résolu par *vectorisation* : on demande à la fonction de calculer récursivement le couple (F_n, F_{n-1}). Le nombre d'appels est maintenant n :

```
Fib2(n)                              (* non défini pour n = 0 *)
si n = 1 alors rendre (1,0) sinon
   ((a,b) := Fib2(n-1) ; rendre (a+b,a)) ;;
```

Coefficients binomiaux. Il s'agit de calculer les coefficients binomiaux par la formule de Pascal : $\binom{n}{p} = \binom{n-1}{p} + \binom{n-1}{p-1}$. Le lecteur vérifiera que la situation est pire que dans l'exemple précédent (exercice I.8.48 de la page 389).

```
Pasc(n,p)
si p = 0 alors rendre 1 sinon si n = 0 alors rendre 0
sinon rendre Pasc(n-1,p) + Pasc(n-1,p-1) ;;
```

Parité par récursion mutuelle. On part de l'observation que n est pair si, et seulement si, $n-1$ est impair, etc. On fait appel à une *récursion mutuelle*, qui n'est pas plus mystérieuse que les suites définies par récurrence double, comme $u_{n+1} := u_n + v_n$ et $v_{n+1} := u_n - v_n$ (mais voir l'exercice I.8.49 de la page 389).

```
Pair(n)
si n = 0 rendre vrai sinon rendre Impair(n-1) ;;
Impair(n)
si n = 0 rendre faux sinon rendre Pair(n-1) ;;
```

Algorithme de Cholesky. Outre qu'il illustre le module d'algèbre bilinéaire, l'exemple suivant montre comment traiter des données encombrantes (ici, des matrices) : il n'est pas question de les passer en argument de fonctions récursives, ce qui reviendrait à les recopier en de nombreux exemplaires dans la pile ; on lit donc les données dans des variables globales et les résultats sont produits par effet de bord sur des variables globales.
La matrice symétrique définie positive A sera représentée par un tableau à deux indices **`A[1..n,1..n]`**, ainsi que la matrice triangulaire supérieure T. On veut calculer T à coefficients diagonaux positifs telle que ${}^tTT = A$. Pour cela, on considère A et T comme écrits par blocs de côtés $n-1$ et 1 : $A = \begin{pmatrix} A' & a \\ {}^ta & \alpha \end{pmatrix}$ et $T = \begin{pmatrix} T' & t \\ 0 & \tau \end{pmatrix}$, où $A', T' \in M_{n-1}(\mathbb{R})$, $a, t \in M_{1,n-1}(\mathbb{R})$ et $\alpha, \tau \in \mathbb{R}$. On doit successivement résoudre les équations : ${}^tT'T' = A'$, ${}^tT't = a$ et ${}^ttt + \tau^2 = \alpha$. On commence donc par calculer la sous-matrice **`T[1..n-1,1..n-1]`** en appliquant récursivement la *procédure* Cholesky à la sous-matrice **`A[1..n-1,1..n-1]`**. Le programme suivant suppose connus les tableaux **`A[1..n,1..n]`** et **`T[1..n,1..n]`** comme *variables globales* et ne passe en argument que la taille k de la partie à traiter. On suppose que T a été initialisée à 0. On fera appel à des fonctions auxiliaires : la fonction **`resoudre-systeme(B,C,X)`**, qui range dans **`X`** la solution du système linéaire (ici triangulaire) ${}^tBX = C$; et les fonctions carré scalaire **`carscal`** et de racine carrée **`sqrt`**.

```
Cholesky(k)   (* calcule T[1..k,1..k] d'après A[1..k,1..k] *)
si k = 0 alors rien sinon (Cholesky(k-1) ;
  resoudre-systeme(T[1..k-1,1..k-1],A[k,1..k-1],T[k,1..k-1]) ;
  T[k,k] := sqrt(A[k,k]) - carscal(X));;
```

Les tours de Hanoï. Voici enfin une application *non numérique* de la récursion. On suppose que l'on a empilé 7 disques de tailles décroissantes de bas en haut. Ces disques sont posés sur une position S (pour « source ») et doivent être transportés sur une position B (pour « but ») selon les règles suivantes : on ne peut déplacer qu'un disque à la fois ; on ne peut poser un grand disque sur un petit disque. Il est évidemment *nécessaire* de disposer d'au moins une position intermédiaire G (pour « garage »), et l'on va voir que c'est suffisant. Avant d'examiner la solution qui suit, le lecteur est chaudement encouragé à trouver tout seul comment déplacer 2 disques, puis 3 disques selon ces règles.
La fonction suivante prend en argument le nombre de disques et rend la liste des mouvements (S,G), (S,B) … à effectuer. L'idée est la suivante : pour transporter 7 disques de S en B, on transporte les 6 disques supérieurs de S en G, puis le disque inférieur de S en B puis les 6 disques supérieurs de G en B. Pendant la première et la dernière phase, on n'a pas à se préoccuper de la présence du disque inférieur en S ou en B, car il est plus grand que tous les autres et la deuxième règle ne sera pas violée. La notation « [] » représentera la liste vide. La juxtaposition de listes et éléments sera notée « : : ».

```
Hanoi(n,S,B,G)
si n = 0 alors rendre [] sinon
   rendre Hanoi(n-1,S,G,B) :: (S,B) :: Hanoi(n-1,G,B,S) ;;
```

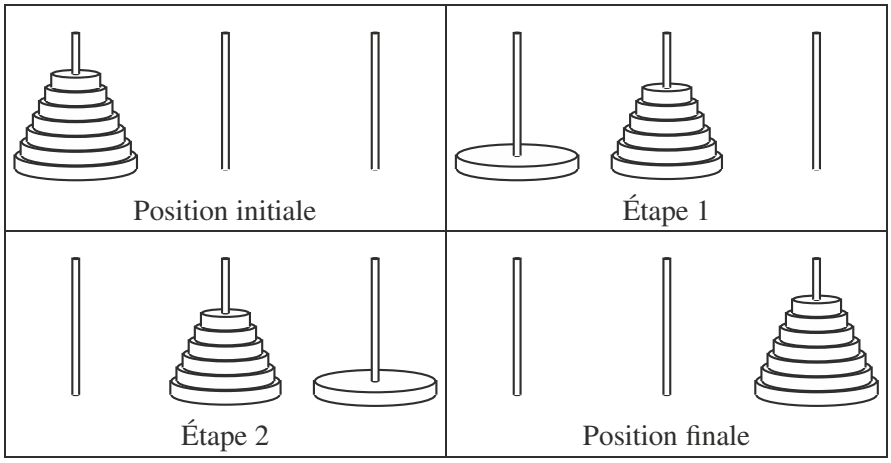


FIGURE I.8.1 – Le problème des tours de Hanoï

3.2 Analyse d'algorithmes récursifs

Comme pour les algorithmes itératifs ([L1], module d'algorithmique), l'analyse des algorithmes récursifs repose sur l'induction, mais les formulations sont différentes. Les méthodes que nous exposons ici ne s'appliquent qu'à certains types de programmes récursifs ; d'ailleurs, un résultat fondamental en logique (ou en informatique théorique) est qu'*il n'existe pas de procédure générale pour décider si un algorithme termine pour toute valeur de ses arguments*. (Voir également l'exercice I.8.52 de la page 390.)

Nous appellerons (informellement) *programme récursif élémentaire* une fonction **`F(w)`** dont l'argument w varie dans un domaine $\mathcal{A}$ et dont le déroulement est le suivant :

1. À partir de son argument w, elle « extrait » une liste finie $\arg_F(w)$ d'arguments $w_i \in \mathcal{A}$. Cette liste peut être vide (cela donnera le « cas de sortie » de la récursion) ou de taille arbitraire (mais finie). Le point essentiel est que la détermination des w_i est « directe », elle ne fait pas elle-même appel à un algorithme récursif.
2. Pour chacun des $w_i \in \arg_F(w)$, il y a un appel récursif à $F(w_i)$.
3. À partir des résultats $F(w_i)$, la fonction fait un calcul « direct » et rend un résultat.

Le mot *fonction* n'a pas ici le sens « statique » qui est le sien en mathématiques : une fonction est un programme qui rend un résultat, par opposition à une procédure qui agit par effet de bord ; une fonction a un déroulement, c'est un objet dynamique. Tous les exemples de fonctions que nous avons donnés entrent dans le cadre de la définition précédente. Mais les fonctions d'Ackermann-Peter et de Morris (exercices I.8.53 et I.8.54 de la page 390) n'y entrent pas.

L'induction sur la taille des arguments requiert une structure supplémentaire. Nous supposerons $\mathcal{A}$ *gradué*, c'est-à-dire muni d'une application « taille » $w \mapsto |w|$ à valeurs dans un ensemble ordonné artinien E ; typiquement, $E = \mathbb{N}$, mais ce peut être $\mathbb{N} \times \mathbb{N}$ muni de l'ordre lexicographique, etc. Nous supposerons de plus que nos programmes F sont *bien*

fondés dans le sens suivant :

$$\forall w \in \mathcal{A}\ ,\ \forall w' \in \arg_F(w)\ ,\ |w'| < |w|.$$

Noter que cette propriété est tautologiquement satisfaite lorsque la liste $\arg_F(w)$ est vide (cas de base de la récursion).

Théorème 24 (Terminaison et correction).
(i) Un programme bien fondé termine pour toute valeur de son argument.
(ii) Soit P_F un prédicat sur $\mathcal{A}$ énonçant une propriété du résultat de l'appel de `F(w)`. Supposons de plus que ce prédicat est *héréditaire* dans le sens suivant :

$$\forall w \in \mathcal{A}\ ,\ \big(\forall w' \in \arg_F(w)\ ,\ P_F(w')\big) \Longrightarrow P_F(w). \tag{4}$$

Alors, $\forall w \in \mathcal{A}\ ,\ P_F(w)$.

Démonstration. (i) Il suffit d'appliquer le lemme de König (théorème 22 de la page 373) à l'*arbre des appels* de `F(w)`. Ce dernier est défini ainsi. Ses nœuds sont des listes d'éléments du domaine $\mathcal{A}$: en fait, les contenus possibles de la pile de récursion. La racine est la liste $[w]$. Les fils d'un nœud $[w_0; \cdots; w_n]$ sont tous les $[w_0; \cdots; w_n; w']$, où $w' \in \arg_F(w)$. Par hypothèse, cet arbre est finitaire. S'il admettait une branche infinie, on aurait une suite $w_0 := w, w_1 \in \arg_F(w_0), \ldots$ d'éléments de $\mathcal{A}$ de tailles strictement décroissantes, ce qui est impossible. L'arbre des appels est donc fini, et le temps d'exécution aussi.
(ii) On raisonne par induction noetherienne (section 6.1 du module « Fondements » de [L1]). S'il existe un argument $w \in \mathcal{A}$ pour lequel $P_F(w)$ est faux, il en existe un de taille minimum. Puisque les $w' \in \arg_F(w)$ sont de tailles strictement plus petites, ils satisfont à $P_F(w')$. D'après la propriété d'hérédité (4), cela entraîne que $P_F(w)$ est vrai, ce qui est une contradiction. ∎

Exemples.

1. Si $F := fact$, on a $\mathcal{A} := \mathbb{N}$, $E := \mathbb{N}$, $|n| = n$ et $\arg_F(0) := []$, $\arg_F(n) := [n-1]$ si $n \geqslant 1$. La fonction est bien fondée, donc l'algorithme termine à coup sûr. Pour prouver la correction, on considère le prédicat $P_F(n)$: « le calcul de `fact(n)` rend $n!$ ». Montrons qu'il est héréditaire c'est-à-dire (4).

   Si $n = 0$, on a $\forall n' \in \arg_F(n)\ ,\ P_F(n')$, puisque $\arg_F(n)$ est vide ; il faut donc montrer que l'appel de `fact(0)` rend $0!$. Mais, dans ce cas, la fonction rend bien 1 (sans appel récursif). Si $n \geqslant 1$, on a $\arg_F(n) := [n-1]$ et il s'agit de montrer que $P_F(n-1) \Rightarrow P_F(n)$. Supposons donc que l'appel de `fact(n-1)` rend $(n-1)!$. L'appel de `fact(n)` provoque l'appel récursif de `fact(n-1)`, qui rend $(n-1)!$, et l'on rend alors `n * fact(n-1)`, c'est-à-dire $n(n-1)! = n!$.

2. Si $F := Fib$, on a $\mathcal{A} := \mathbb{N}$, $E := \mathbb{N}$, $|n| = n$ et $\arg_F(0) := []$, $\arg_F(1) := []$ et $\arg_F(n) := [n-1, n-2]$ si $n \geqslant 2$. Comme $n-1, n-2 < n$, la fonction est bien fondée et l'algorithme termine à coup sûr.

3. Si $F := Pasc$, on a $\mathcal{A} := \mathbb{N}\times\mathbb{N}$, $E := \mathbb{N}$, $|(n,p)| = n$ et $\arg_F(n,p) := []$ si $n = 0$ ou $p = 0$, $\arg_F(n,p) := [(n-1,p),(n-1,p-1)]$ si $n \geqslant 1$ et $p \geqslant 1$. La fonction est bien fondée, donc l'algorithme termine à coup sûr.
4. Si $F := Hanoi$, on supposera que S, B, G sont $0, 1, 2$ dans un certain ordre ; on a donc $\mathcal{A} := \mathbb{N} \times \mathfrak{S}_3$, $\arg_F(0,\sigma) := []$ et :

$$\arg_F(n,\sigma) := [(n-1,\sigma\circ\tau_{2,3}),(n-1,\sigma\circ\tau_{1,3})] \quad \text{pour } n \geqslant 1.$$

Prenant $E := \mathbb{N}$ et $|(n,a,b,c)| = n$ et, on voit que la fonction est bien fondée, donc l'algorithme termine à coup sûr.

3.3 Récursion et induction structurelle

Un type important de récursion intervient lorsque les éléments de $\mathcal{A}$ admettent une *décomposition naturelle* et que la liste $\arg_F(w)$ est calculée à partir de cette décomposition. C'est le cas des objets qui ont été définis par induction, comme les expressions régulières (introduites après la définition 6 de la page 361) et les arbres binaires (définition 14 de la page 374).

Récursion sur les arbres binaires. Puisque tout arbre binaire est soit égal à F soit de la forme $N(t_1,t_2)$, et ce, de manière déterminée, un style de programmation proche de l'induction structurelle est possible. Montrons par exemple comment calculer la taille de t :

```
Taille(t)
| F -> 0
| N(t1,t2) -> 1 + Taille(t1) + Taille(t2) ;;
```

Ce qui est sous-entendu s'exprime comme suit :

```
si t = F alors rendre 0 sinon
    (mettre t sous la forme N(t1,t2);
    rendre 1 + Taille(t1) + Taille(t2)) ;;
```

On *filtre* **t** selon les *profils* possibles **F** et **N(t1,t2)** ; si c'est le second filtrage qui réussit, on nomme **t1** et **t2** les composantes correspondantes. La barre verticale **|** de séparation joue le rôle d'un « ou ». Le filtrage par profils est appelé, en anglais, *pattern matching*. La preuve de terminaison d'algorithmes construits sur ce modèle est presque tautologique. La preuve de correction se fait par induction structurelle. À titre d'exemple, nous allons établir une bijection entre les arbres binaires de taille n et les forêts de taille n (la taille d'une forêt est la somme des tailles de ses éléments). Pour cela, nous définirons deux fonctions de conversion et prouverons qu'elles sont réciproques l'une de l'autre. Ces fonctions ne font que traduire la transformation décrite page 375.

```
ArbreBinaireEnForet(t)
| F -> []
| N(t1,t2) ->
    R(ArbreBinaireEnForet(t1)) :: ArbreBinaireEnForet(t2) ;;
```

Pour la fonction réciproque, nous devrons filtrer les profils possibles d'une forêt. Soit celle-ci est vide et c'est la liste **[]**. Soit elle est de la forme **x :: q**, c'est-à-dire une liste dont le premier élément est **x** et la suite est **q**. Dans ce dernier cas, **x** est un arbre général, donc il est lui même de la forme **R(f)**, où **f** est une forêt ; et, bien entendu, **q** est une liste d'arbres généraux, donc une forêt.

```
ForetEnArbreBinaire(f)
| [] -> F
| R(f1) :: f2 ->
    N(ForetEnArbreBinaire(f1),ForetEnArbreBinaire(f2)) ;;
```

Prouvons, par induction structurelle, que la conversion de t en f suivie de la conversion dans le sens contraire, redonne bien t. Si $t = F$, il est transformé en $[]$, puis en F, ce qui convient. Supposons que cela marche pour t_1 (qui donne f_1, puis t_1) et pour t_2 (qui donne f_2, puis t_2), et soit $t := N(t_1, t_2)$. Le premier programme transforme alors t en $f := R(f_1) :: f_2$; et le deuxième programme transforme f en $N(t_1, t_2) = t$. Pour le reste de la preuve (conversion de forêt en arbre et retour, et taille) voir l'exercice I.8.56 de la page 390.

Récursion sur les expressions régulières. Voici une définition inductive des expressions régulières bien adaptée au filtrage : une expression régulière est **Vide** (dénotant le langage $\varnothing$) ; ou bien de la forme **Epsilon** (dénotant le langage réduit au mot vide ε) ; ou bien de la forme **Sing(a)**, où a est un caractère (dénotant le langage réduit au mot d'une lettre **a**) ; ou bien de l'une des formes **Plus(e1,e2)**, **Conc(e1,e2)**, **Star(e1)**, où **e1** et **e2** sont des expressions régulières (dénotant respectivement une somme, une concaténation ou une étoile). Voici le calcul de la propriété d'être propre, basé sur le lemme 5 de la page 358.

```
Propre(e)
| Vide -> vrai | Epsilon -> faux | Sing(a) -> vrai
| Plus(e1,e2) -> Propre(e1) et Propre(e2)
| Conc(e1,e2) -> Propre(e1) ou Propre(e2)
| Star(e1) -> faux ;;
```

Exercice type corrigé

I.8.0 On définit inductivement une *expression algébrique* comme suit : les variables x, y, z u, v, w sont des expressions algébriques ; si e_1 et e_2 sont des expressions algébriques, alors $e_1 + e_2$ et $e_1 * e_2$ sont des expressions algébriques. Ainsi, avec les règles usuelles de parenthésage, $e_0 := (x + y) * (u + v)$ est une expression algébrique.

(i) Associer à une expression algébrique un arbre binaire qui n'en retient que la structure (les variables correspondent aux feuilles, les opérateurs aux nœuds). Ainsi, e_0 a pour image **N(N(F,F),N(F,F))**. On donnera une définition inductive, puis un algorithme récursif de conversion (avec filtrage). Combien d'expressions ont pour image un arbre donné ?

(ii) On associe à une expression algébrique son *écriture postfixe*, obtenue en écrivant toujours les

signes d'opération après les deux opérandes et en omettant les parenthèses. Ainsi, e_0 a pour image **''x y + u v + *''**. Écrire la définition inductive puis le programme de conversion. Les symboles $x, y, z, u, v, w, +, *$ seront représentés par les caractères **'x'**, **'y'**, **'z'**, **'u'**, **'v'**, **'w'** et l'on notera **c1^c2** la concaténation de deux mots.
(iii) Un algorithme (efficace) d'évaluation d'une expression e à partir de son écriture postfixe **p** utilise une pile susceptible de stocker les valeurs numériques des variables. La chaîne de caractères **p** est lue de gauche à droite. Chaque fois que l'on lit un caractère représentant une variable, on empile sa valeur. Chaque fois que l'on lit un caractère représentant une opération, on dépile deux valeurs, on leur applique l'opération et l'on empile le résultat. Décrire les états successifs de la pile lorsque l'on applique l'algorithme à l'écriture postfixe de e_0 (on supposera que toutes les variables valent 1). Démontrer par induction structurelle que, si la pile est vide avant l'exécution de l'algorithme, à la fin, elle contient la valeur correcte. Exprimer inductivement en fonction e la hauteur maximale atteinte par la pile.
Solution. (i) Définition inductive : à x, y, z u, v, w, on associe F ; à $e_1 + e_2$ et $e_1 * e_2$, on associe $N(t_1, t_2)$, où t_1, t_2 ont été associés à e_1, e_2. Pour reconstituer e à partir de l'arbre associé t, il faut choisir un opérateur (parmi deux) pour chacun des $|t|$ nœuds et une variable (parmi six) pour chacune des $(|t| + 1)$ feuilles : il y a $2^{|t|}6^{|t|+1}$ possibilités.

```
ExpalgEnArbre(e)
| X -> F | Y -> F | Z -> F | U -> F | V -> F | W -> F
| Plus(e1,e2) -> N(ExpalgEnArbre(e1),ExpalgEnArbre(e2))
| Mult(e1,e2) -> N(ExpalgEnArbre(e1),ExpalgEnArbre(e2)) ;;
```

(ii) Définition inductive : à x, y, z u, v, w, on associe les caractères **'x'**, **'y'**, **'z'**, **'u'**, **'v'**, **'w'** ; à $e_1 + e_2$ et $e_1 * e_2$, on associe respectivement **p1^p2^'+'** et **p1^p2^'*'**, où **p1**, **p2** ont été associés à e_1, e_2.

```
Postfixe(e)
| X -> 'x' | Y -> 'y' | Z -> 'z' | U -> 'u' | V -> 'v' | W -> 'w'
| Plus(e1,e2) -> Postfixe(e1)^Postfixe(e2)^'+'
| Mult(e1,e2) -> Postfixe(e1)^Postfixe(e2)^'*' ;;
```

(iii) Voici les états successifs de la pile (représentée sommet à gauche)lors de la lecture de l'écriture postfixe **''xy+uv+*''** : **[]**, puis **[1]**, puis **[1;1]**, puis **[2]**, puis **[1;2]**, puis **[1;1;2]**, puis **[2;2]**, puis enfin **[4]** (valeur finale).
Nous démontrerons par induction sur e que, si la pile est dans l'état **[---]** avant l'exécution de l'algorithme, elle est dans l'état **[v;---]** après, avec la valeur correcte de **v**. Nous calculerons en même temps le nombre total de positions utilisées au-delà de la partie de départ en fonction de l'arbre associé. Si e est une variable, son écriture postfixe **p** est une lettre et l'algorithme ne fait qu'empiler sa valeur : la propriété est vérifiée. Le nombre de positions utilisées est 1.
Supposons la propriété vraie pour e_1 (écriture postfixe **p1**, valeur **v1**, nombre de positions utilisées n_1) et pour e_2 (écriture postfixe **p2**, valeur **v2**, nombre de positions utilisées n_2) et soit $e := e_1 + e_2$ (le raisonnement est le même pour $e_1 * e_2$). Son écriture postfixe est **p1^p2^'+'**. L'algorithme lit ce mot de gauche à droite et traite d'abord **p1** ; à la fin ce cette lecture, la pile est dans l'état **[v1;---]** (c'est l'hypothèse d'induction pour e_1) ; le nombre maximum de positions utilisées pendant cette phase a été n_1 ; et il reste à lire **p2^'+'** Pendant la deuxième phase, on lit **p2** ; à la fin, la pile est dans l'état **[v2;v1;---]** (c'est l'hypothèse d'induction pour e_2, utilisée avec l'état de départ **[v1;---]**) ; le nombre maximum de positions utilisées pendant cette phase a été n_2, mais au-delà de **v1** : le nombre maximum de positions utilisées pendant la lecture de **p1^p2** a donc été $\max(n_1, 1 + n_2)$. Il reste à lire **'+'**, qui suscite le dépilage de **v2** et **v1** et l'empilage de

v1+v2, qui est la valeur **v** de e : la pile est bien dans l'état prévu **[v;---]**.
On trouve de plus le calcul inductif suivant du nombre $\chi(e)$ de positions utilisées : si e est une variable, $\chi(e) := 1$; si $e := e_1 + e_2$ ou $e_1 * e_2$, $\chi(e) := \max\big(\chi(e_1), 1 + \chi(e_2)\big)$.

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.8.1 Définir inductivement les éléments du langage L_1 de l'exercice 1 de la page 354 et du langage D de la page 354.

I.8.2 Montrer que $\tilde{\tilde{u}} = u$ et que $\widetilde{uv} = \tilde{v}\tilde{u}$. En déduire que $u\tilde{u}$ est un *palindrome* (un mot v égal à $\tilde{v}$). Obtient-on ainsi tous les palindromes ?

I.8.3 On dit que le mot w' est une *recombinaison* de w, ce que l'on notera $w \mathcal{R} w'$, s'il existe u_1, v, u_2 tels que $w = u_1 v u_2$ et $w' = u_1 \tilde{v} u_2$ (penser aux aventures de l'ADN). Est-ce une relation d'équivalence ? Montrer que, pour toute lettre x, on a les équivalences : $w \mathcal{R} w' \Leftrightarrow xw \mathcal{R} xw' \Leftrightarrow wx \mathcal{R} w'x$. En déduire un algorithme qui teste si $w \mathcal{R} w'$.

I.8.4 Comment obtenir $u \mapsto |u|$ par application du théorème 3 de la page 356 ?

I.8.5 Démontrer que les relations « facteur », « préfixe » et « suffixe » sont des ordres artiniens ([L1] , module « Fondements », section 6.1). La restriction de la relation « facteur » à l'ensemble des facteurs de v est-elle un ordre total ?

I.8.6 $*$ Démontrer les équivalences suivantes : pour que $uv = vu$, il faut, et il suffit, qu'il existe w tel que $u, v \in w^*$ (*i.e.* u et v sont puissances de w) ; pour qu'il existe des entiers k, l non tous deux nuls tels que $u^k = v^l$, il faut, et il suffit, qu'il existe w tel que $u, v \in w^*$.

I.8.7 $*$ (i) Soient $w, w' \in X^*$. Montrer que, pour qu'il existe des mots u et v tels que $w = uv$ et $w' = vu$, il faut, et il suffit, qu'il existe un mot v tel que $vw = w'v$. On dit alors que les mots w et w' sont *conjugués*.
(ii) Vérifier que le groupe cyclique $\mathbb{Z}/n\mathbb{Z}$ agit sur l'ensemble X^n des mots de longueur n par translation des indices et que la la classe de conjugaison de w est une orbite sous l'action de $\mathbb{Z}/n\mathbb{Z}$. Démontrer que le nombre de conjugués de w est égal à la *période* de w, c'est-à-dire au plus petit diviseur p de $n := |w|$ tel que $w(i) = w(i + p)$ pour tout $i \in [\![0, n - p - 1]\!]$.

I.8.8 $*$ Démontrer que le sous-monoïde de X^* engendré par $L \subset X^*$ est L^*.

I.8.9 Démontrer les égalités : $(ab^*)^*a^* = a^*(b^*a)^* = (a + b)^*$ et $(ab)^*a = a(ba)^*$.

I.8.10 Décrire le langage des écritures décimales des nombres décimaux positifs.

I.8.11 On définit l'*étoile propre* d'un langage L comme $L^+ := \bigcup_{n \geqslant 1} L^n$. Démontrer les égalités : $L^+ = LL^* = L^*L$ et $L^* = \varepsilon + L^+$. Que vaut $(\varepsilon + L)^+$?

I.8.12 Montrer que, si $\varepsilon \in L_1$, tout langage de la forme $L := L_1^*(L_2 + L_3)$ est solution de l'équation d'Arden $L = L_1L + L_2$.

I.8.13 $*$ L'équation d'Arden $L = L_1L+L_2$ peut être vue comme équation au point fixe et résolue par itération : on pose $A_0 := \varnothing$ puis $A_{i+1} := L_1A_i+L_2$; montrer que la suite des A_i est croissante et que $\bigcup_{i\geqslant 0} A_i$ est le plus petit point fixe.

I.8.14 $**$ Montrer que l'on peut écrire le système de la proposition 7 de la page 359 sous forme matricielle et le résoudre par itération.

I.8.15 Comparer $u(u^{-1}L)$, $u^{-1}(uL)$ et L.

I.8.16 $*$ Démontrer toutes les règles de calcul relatives aux résiduels.

I.8.17
Calculer tous les résiduels de $\{u \in \{a,b\}^* \mid |u|_a, |u|_b \text{ pairs}\}$ et ceux de $\{u \in \{a,b\}^* \mid |u|_a = |u|_b\}$. Ces langages sont-ils rationnels ?

I.8.18 Notant $P := \{u \in \{a,b\}^* \mid |u|_a, |u|_b \text{ pairs}\}$ et $I := \{u \in \{a,b\}^* \mid |u|_a, |u|_b \text{ impairs}\}$, vérifier que $P = (aa + bb)P + (ab + ba)I + \varepsilon$ et $I = (aa + bb)I + (ab + ba)P$, et en déduire que P et I sont rationnels.

I.8.19 $*$ Soit $X := \{0,1\}$. Notons, pour $i = 0, 1, 2$, $L_i \subset X^*$ le langage des écritures en base 2 des entiers congrus à i modulo 3. On conviendra, pour simplifier, que l'écriture en base 2 de 0 est ε. Montrer les égalités suivantes : $L_0 = L_0.0 + L_1.1 + \varepsilon$, $L_1 = L_2.0 + L_0.1$ et $L_2 = L_1.0 + L_2.1$ En déduire que L_0, L_1 et L_2 sont rationnels.

I.8.20 $*$ Montrer que $u \in L \Leftrightarrow \varepsilon \in u^{-1}L$. En déduire un algorithme récursif qui teste si u appartient au langage dénoté par une expression régulière.

I.8.21 $**$ Soit D le langage de Dyck (page 354). Déterminer $D \cap a^*b^*$ et en déduire que D n'est pas rationnel. Pour tout $w \in D$, montrer que $|w|_a = |w|_b$ et que, pour tout préfixe propre w' de w, on a $|w'|_a \geqslant |w'|_b$. Démontrer la réciproque.
Soit D_n l'ensemble des mots de D de longueur $2n$. Montrer que l'application $(w_1, w_2) \mapsto aw_1bw_2$ est une bijection de l'union disjointe des $D_i \times D_j$ tels que $i+j = n-1$ sur D_n (pour $n \geqslant 1$). En déduire une relation de récurrence définissant $|D_n|$ et un lien avec les nombres de Catalan (page 42).

I.8.22 $*$ Soient $X := \{a,b\}$ et $L := \{u \in X^* \mid u = \tilde{u}\}$ (langage des palindromes). Calculer $L' := L \cap a^*ba^*$. Démontrer qu'aucun mot de L' n'admet de facteur itérant (*cf.* le théorème 20 de la page 371). Démontrer que L' a une infinité de résiduels. Le langage L est-il rationnel ?

Exercices sur la section 2

I.8.23 Combien y a-t-il de graphes non orientés, resp. orientés (S, A) sur S ?

I.8.24 La *liste d'adjacence* d'un graphe non orienté G (resp. d'un graphe orienté) est un tableau T_G de n listes, où $T_G[i]$ est la liste des voisins (resp. des successeurs) de i. Écrire les algorithmes de conversion entre matrice et liste d'adjacence.

I.8.25 $*$ (i) Soit $G := (S, A)$ un graphe orienté. On veut calculer l'ensemble $\mathrm{Acc}(E)$ des sommets accessibles par un chemin depuis $E \subset S$. On note, pour tout $X \subset S$, $\Psi(X)$ l'ensemble des successeurs des éléments de X. Montrer que Ψ est croissante pour l'inclusion et que $\Psi(X \cup Y) = \Psi(X) \cup \Psi(Y)$. On pose $X_0 := E$ et $X_{k+1} := X_k \cup \Psi(X_k)$. Montrer que $(X_k)_{k \geqslant 0}$ est une suite croissante de parties de E et que donc elle stationne en un entier m. Montrer que $\mathrm{Acc}(E) = X_m$. Peut-on majorer m *a priori* ?
(ii) On cherche à optimiser le calcul précédent. On pose $Y_0 := X_0$ et, pour $k \geqslant 1$, $Y_k := X_k \backslash X_{k-1}$. Montrer que les couples (X_k, Y_k) se calculent inductivement par les formules : $Y_{k+1} := \Psi(Y_k) \backslash X_k$ et $X_{k+1} := X_k \cup Y_{k+1}$.
(iii) Adapter ces résultats au cas d'un graphe non orienté.

I.8.26 $**$ L'algorithme de Warshall améliore la méthode de calcul de la matrice de connexion d'un graphe non orienté qui suit la proposition 14 de la page 364 et de son analogue pour la matrice d'accessibilité d'un graphe orienté après la définition 10 de la page 368. On introduit une matrice auxiliaire X, dans laquelle on a recopié la matrice d'adjacence du graphe.

```
Pour k de 1 a n faire pour i de 1 a n faire
pour j de 1 a n faire X[i,j] := X[i,j] ou (X[i,k] et X[k,j]) ;;
```

À la fin, X est la matrice de connexion (ou d'accessibilité). Justifier cet algorithme ; il s'agit de comprendre ce que représente X après k exécutions de la boucle externe.

I.8.27 $**$ Soient $S_1 := \{s_1, \ldots, s_n\}$, $S_2 := \{t_1, \ldots, t_n\}$ (que l'on suppose disjoints), et $S = S_1 \cup S_2$ et $G := (S, A)$ un graphe (non orienté) biparti. On introduit la matrice de polynômes en n^2 indéterminées $M \in \mathbb{R}[X_{1,1}, \ldots, X_{n,n}]$ dont le coefficient d'indices (i, j) vaut $X_{i,j}$ si s_i et t_j sont adjacents, et 0 sinon. Un *mariage parfait* sur G est une bijection μ de S_1 sur S_2 telle que $\mu(s)$ est adjacent à s. Montrer que l'existence d'un mariage parfait sur G équivaut à $\det M \neq 0$.

I.8.28 À l'aide de la proposition 17 de la page 367, montrer que les $\tau_{1,i}$ $(2 \leqslant i \leqslant n)$ et les $\tau_{i-1,i}$ $(2 \leqslant i \leqslant n)$ forment deux familles minimales de générateurs de $\mathfrak{S}_n$. Quel nombre maximum de facteurs pris dans l'une ou l'autre famille est *a priori* nécessaire pour obtenir une transposition quelconque ? Ce maximum est-il atteint dans le cas où $n := 3$ ou 4 ?

I.8.29 Définir la matrice d'adjacence pour un multigraphe de telle sorte que la proposition 14 de la page 364 reste valide.

I.8.30 On dit que le graphe orienté $G := (S, A)$ est *fortement connexe* si $\forall s \in S$, $\mathrm{Acc}(s) = S$. Démontrer G est fortement connexe si, et seulement s'il existe N tel que $\overline{M}_G^N$ ne comporte que des coefficients **vrai**. Peut-on choisir N ?

I.8.31 Combien y a-t-il d'automates finis sur X donné ayant un ensemble d'états Q donné ? Combien sont déterministes complets ?

I.8.32 Soit $\mathcal{A} := (Q, I, F, \Delta)$ un AFDC. Quel langage est reconnu par $(Q, I, Q \setminus F, \Delta)$?

I.8.33 Soient $\mathcal{A}_1 := (Q_1, I_1, F_1, \Delta_1)$ et $\mathcal{A}_2 := (Q_2, I_2, F_2, \Delta_2)$ deux automates finis sur X. Quitte à renommer les états de $\mathcal{A}_2$, on suppose que Q_1 et Q_2 sont disjoints, de sorte que l'on peut former l'union disjointe $Q_1 \vee Q_2$. On définit l'automate somme : $\mathcal{A}_1 \oplus \mathcal{A}_2 := (Q_1 \vee Q_2, I_1 \vee I_2, F_1 \vee F_2, \Delta_1 \vee \Delta_2)$ (on rappelle que $\vee$ dénote l'union disjointe). Quel langage est reconnu par $\mathcal{A}_1 \vee \mathcal{A}_2$?

I.8.34 Soient $\mathcal{A}_1 := (Q_1, I_1, F_1, \Delta_1)$ et $\mathcal{A}_2 := (Q_2, I_2, F_2, \Delta_2)$ deux automates finis sur X. Déterminer le langage reconnu par $\mathcal{A}_1 \times \mathcal{A}_2 := (Q_1 \times Q_2, I_1 \times I_2, F_1 \times F_2, \Delta)$, où $\Delta := \{\big((q_1, q_2), x, (q_1', q_2')\big) \mid (q_1, x, q_1') \in \Delta_1 \quad \text{et} \quad (q_2, x, q_2') \in \Delta_2\}$.

I.8.35 Quel langage est reconnu par l'automate :

$$\longrightarrow q_0 \xrightarrow{a} q_1 \xrightarrow{b} q_2 \xrightarrow{b} q_3 \xrightarrow{a} \underline{\mathbf{q_4}} \ ?$$

I.8.36 Pour tout mot w, décrire explicitement un automate qui reconnaît le langage $\{w\}$. Comment construire un automate qui reconnaît un langage fini ?

I.8.37 Décrire l'automate des résiduels de a^*b^*, puis celui de $X^* \setminus a^*b^*$.

I.8.38 $**$ (i) À partir d'automates reconnaissant L_1, L_2, L, décrire explicitement des automates qui reconnaissent $L_1 + L_2$, L_1L_2, L^*.
(ii) Soit $\mathcal{A} := (Q, I, F, \Delta)$ un automate. Pour $s, t \in Q$, on note $L_{s,t}$ l'ensemble des traces des calculs d'origine s et d'extrémité t. Exprimer $L(\mathcal{A})$ en fonction des $L_{s,t}$. Pour tout $A \subset Q$, on note $L_{s,t}(A)$ le sous-ensemble de $L_{s,t}$ correspondant à des calculs dont tous les états intermédiaires sont dans A. Que vaut $L_{s,t}(\varnothing)$? Démontrer, pour $s, t \in Q$ et $u \in A \subset Q$, l'égalité : $L_{s,t}(A) = L_{s,t}(A\setminus\{u\}) + L_{s,u}(A\setminus\{u\})\big(L_{u,u}(A\setminus\{u\})\big)^* L_{u,t}(A\setminus\{u\})$. En déduire que les $L_{s,t}(A)$ sont rationnels, ainsi que $L(\mathcal{A})$ (algorithme de Mc Naughton-Yamada).

I.8.39 Soit L un langage rationnel inclus dans $\{a^n b^n \mid n \in \mathbb{N}\}$. Montrer que L est fini.

I.8.40 Soit $X := \{a, b\}$. Montrer que tout mot non vide du langage $\{u \in X^* \mid |u|_a = |u|_b\}$ admet un facteur itérant. Qu'en déduire au sujet de la réciproque du lemme d'itération ?

I.8.41 $**$ On pose $S := [\![1, n]\!]$ et l'on considère un graphe non orienté $G := (S, A)$. Soit $(e_i)_{1 \leqslant i \leqslant n}$ la base canonique de $\mathbb{R}^n$. Démontrer que la famille $(e_i - e_j)_{\{i,j\} \in A}$ est libre si, et seulement si, G est acyclique. Démontrer que cette famille engendre l'hyperplan $H \subset \mathbb{R}^n$ d'équation $\sum_{i=1}^{n} x_i = 0$ si, et seulement si, G est connexe. Retrouver ainsi le théorème 16 de la page 366.

I.8.42 $*$ Démontrer la caractérisation suivante d'un arbre enraciné fini ou infini : c'est un ensemble ordonné S admettant un unique minimum et tel que, pour tout $s \in S$, l'ensemble des minorants de s soit une chaîne ([L1], module « Fondements ») finie.

I.8.43 Dessiner tous les arbres généraux de tailles $1, 2, 3, 4$.
Dessiner tous les arbres binaires de tailles $0, 1, 2, 3$.

I.8.44
Démontrer par induction structurelle que, pour tout arbre binaire t, on a $\eta(t) \leqslant |t| \leqslant 2^{\eta(t)} - 1$.
Décrire les arbres qui réalisent les égalités extrêmes.
Combien y a-t-il d'arbres binaires de taille n et de hauteur n ?

I.8.45 $*$ (i) Notons a_n le nombre d'arbres généraux de taille n et b_n le nombre de forêts de taille n. Démontrer les égalités : $a_n = b_{n-1}$ et $b_n = \sum_{k \geqslant 0} \sum_{i_1 + \cdots + i_k = n} a_{i_1} \cdots a_{i_k}$.

(ii) Notons c_n le nombre d'arbres binaires de taille n. Démontrer les égalités : $c_0 = 1$ et $c_n = \sum_{i+j=n-1} c_i c_j$. En déduire que les c_n sont les nombres de Catalan (page 42).

(iii) Soient E un ensemble muni d'une loi non associative $\star$ et $a \in E$. À tout produit correctement parenthésé obtenu à partir de a seul, on associe un arbre binaire selon les règles suivantes : $a \mapsto F$; si $p_1 \mapsto t_1$ et $p_2 \mapsto t_2$, alors $p_1 \star p_2 \mapsto N(t_1, t_2)$. Par exemple, $(a \star a) \star a \mapsto N\big(N(F,F),F\big)$ et $a \star (a \star a) \mapsto N\big(F, N(F,F)\big)$. Montrer que l'on obtient ainsi une bijection. Quel est le rapport avec la question (ii) ?
(iv) Formuler et prouver une relation avec le langage de Dyck.

Exercices sur la section 3

I.8.46 Un réel **a** étant fixé, reconnaître (avec preuve !) la fonction récursive suivante :

```
F(n)
si n = 0 alors rendre(1) ;
(q,r) := diveucl(n,2) ; x := F(q) ; y := x * x ;
si r = 0 alors rendre(y) sinon rendre(a * y) ;;
```

I.8.47 $*$ Combien l'exécution de **Fib(n)** provoque-t-elle d'appels de **Fib(k)** ? .

I.8.48 $**$ Combien l'exécution de **Pasc(n,p)** provoque-t-elle d'appels et de calculs redondants ? Comment éviter ce problème ?

I.8.49 En posant $C(n) := \big(Pair(n), Impair(n)\big)$ transformer la récursion mutuelle de la page 378 en une récursion simple.

I.8.50 $*$ Écrire récursivement l'algorithme de Karatsuba de multiplication dichotomique de polynômes ([L1], module II.8, section 3.3).

I.8.51 Vérifier que **Hanoi(2,S,B,G) = [(S,G),(S,B),(G,B)]** et calculer **Hanoi(3,S,B,G)**. Combien y a-t-il de déplacements dans **Hanoi(n,S,B,G)** ?

I.8.52 $*$ Supposons donnée une fonction **termine** telle que l'appel de **termine(f)**, où **f** désigne une *fonction*, rende **vrai** si la fonction **f** termine pour toute valeur de ses arguments et **faux** sinon. Dire ce que rend l'appel de la fonction ci-dessous ; qu'en déduire ?
test(x) : si termine(test) alors rendre test(x) sinon rendre vrai.

I.8.53 $*$ Démontrer que la fonction suivante termine sur tous arguments $n, p \in \mathbb{N}$ et calculer **AP(1,p)**, **AP(2,p)**, **AP(3,p)**, **AP(n,1)** :

```
AP(n,p)                          (* fonction d'Ackermann-Peter *)
si n = 0 alors rendre p+1 sinon si p = 0 alors rendre AP(n-1,1)
sinon rendre AP(n-1,AP(n,p-1)) ;;
```

I.8.54 Avec la définition ci-dessous, calculer **Morris(1,1)** et **Morris(2,2)** :

```
Morris(n,p)                       (* fonction de Morris *)
si n = 0 alors rendre 1 sinon rendre Morris(n-1,Morris(n,p)) ;;
```

I.8.55 Programmer le calcul récursif de la hauteur d'un arbre binaire par filtrage.

I.8.56 Prouver par induction structurelle que la composition des fonctions

ForetEnArbreBinaire et **ArbreBinaireEnForet**

dans cet ordre donne l'identité. Prouver de même qu'elle conserve la taille.

I.8.57 Programmer le calcul des résiduels par une lettre du langage dénoté par une expression régulière, basé sur les formules de la page 360.

I.8.58 Dans l'exercice I.8.0 de la page 383, montrer que $\chi(e)$ ne dépend que de l'arbre binaire associé à e et en donner un calcul inductif. Pour quelle forme d'arbre ce coût est-il le pire (parmi les arbres de taille donnée n) ?

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 X^{e} siècle.

L'âge d'or de l'analyse classique débute donc à la fin du XVII^{e} siècle, justifié par la cinématique. Au siècle suivant, les Bernoulli et Euler, calculateurs 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. Ce volume consacre justement un important module à une initiation à l'analyse de Fourier.

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 XIX^{e} 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 et sont développés dans ce volume).

Vinrent ensuite de fécondes extensions hors des fonctions réelles ou complexes d'une ou plusieurs variables : *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, espaces vectoriels topologiques ou normés (Hilbert, Banach), espaces de Sobolev, distributions (L. Schwartz), 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.

Espaces vectoriels normés

II.1

Dans ce module, K désigne l'un des corps $\mathbb{R}$ ou $\mathbb{C}$. Si $x \in K$, $|x|$ est la valeur absolue de x si $K := \mathbb{R}$ et le module de x si $K := \mathbb{C}$. Par « espace vectoriel », on entend un K-espace vectoriel, de dimension quelconque (finie ou non), sauf mention du contraire.

Ce module peut être vu comme une « boîte à outils » : dans ce volume y seront puisés définitions et résultats de nature « topologique », au fur et à mesure des besoins.

Le cours de [L1] a déjà introduit diverses notions « topologiques » (ce terme ne sera pas défini formellement) : limites, continuité, parties ouvertes, fermées, compactes, etc., concernant $\mathbb{R}$ puis $\mathbb{R}^n$. Nous reprenons ici l'étude de ces notions (et de quelques notions nouvelles), dans le cadre plus général des espaces vectoriels normés. Une telle généralisation est nécessaire pour pouvoir travailler dans des espaces de fonctions (ils ne sont plus de dimension finie), cela apparaîtra clairement dans les modules II.2, II.4 et dans le module de [L2-L3] concernant les polynômes orthognaux, puis dans [MPA].

Lorsqu'on manie des « vecteurs » qui n'ont plus deux ou trois composantes, comme dans le plan ou l'espace, mais peuvent être par exemple des fonctions, il est utile de représenter la situation donnée par un *dessin* afin de faire appel à l'intuition. Bien entendu, de tels dessins peuvent *suggérer* mais pas *remplacer* une démonstration en bonne et due forme.

Il est bon d'indiquer dès maintenant les trois notions essentielles abordées dans ce module : la complétude (espaces complets), la compacité (espaces compacts), et la convexité.

1 Espaces vectoriels normés, espaces métriques

Le lecteur a déjà rencontré à deux reprises la notion de norme sur un espace vectoriel : d'abord dans $\mathbb{R}^n$ (*cf.* le cours de [L1]) puis dans un espace préhilbertien (*cf.* le module I.4). Nous présentons cette notion dans un cadre plus général, adapté aux besoins de l'analyse.

1.1 Normes

Définition 1. On appelle *norme* sur un espace vectoriel E toute application $\|\cdot\|$ de E dans $\mathbb{R}_+ := [0, +\infty[$, notée $x \mapsto \|x\|$ et possédant les propriétés suivantes :
(i) Pour tout $x \in E$, $\|x\| = 0$ implique $x = 0$ (séparation).
(ii) Pour tout $(\alpha, x) \in K \times E$, $\|\alpha x\| = |\alpha|\,\|x\|$ (homogénéité).
(iii) Pour tous $x, y \in E$, $\|x + y\| \leqslant \|x\| + \|y\|$ (inégalité triangulaire).
On appelle *espace vectoriel normé* (en abrégé e.v. n.) tout couple $\big(E, \|\cdot\|\big)$ formé d'un espace vectoriel E et d'une norme $\|\cdot\|$ sur E.

S'il y a plusieurs normes à considérer sur un même espace vectoriel, on peut les noter $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|'$, etc. En pratique, lorsque le contexte indique clairement la norme $\|\cdot\|$ choisie sur E, nous parlerons de « l'espace vectoriel normé E », au lieu de « l'espace vectoriel normé $\big(E, \|\cdot\|\big)$ ». Si $K := \mathbb{R}$ (resp. $K := \mathbb{C}$), on dit aussi « e.v. n. réel (resp. complexe) ».
Voici quelques conséquences immédiates de la définition d'une norme $\|\cdot\|$ sur un espace vectoriel E. D'abord, en prenant $\alpha := -1$ dans la propriété (ii), il vient $\|-x\| = \|x\|$ pour tout $x \in E$. L'inégalité triangulaire se généralise à n vecteurs (récurrence évidente) :

$$\|x_1 + x_2 + \cdots + x_n\| \leqslant \|x_1\| + \|x_2\| + \cdots + \|x_n\|, \quad \text{soit} \quad \left\| \sum_{i=1}^{n} x_i \right\| \leqslant \sum_{i=1}^{n} \|x_i\|.$$

Voici une inégalité utile, valable pour tous $x, y \in E$:

$$\|x - y\| \;\geqslant\; \big|\|x\| - \|y\|\big|. \tag{1}$$

En effet, l'égalité $x = (x-y)+y$ entraîne $\|x\| \leqslant \|x-y\|+\|y\|$, soit $\|x-y\| \geqslant \|x\|-\|y\|$. De même $\|x - y\| \geqslant \|y\| - \|x\|$, d'où la conclusion.
Nous allons donner de nombreux exemples de normes, sur des espaces vectoriels variés. Nous conseillons vivement au lecteur d'avoir plusieurs de ces exemples à l'esprit chaque fois qu'une notion nouvelle est introduite.

Exemples. 1) Sur K, $x \mapsto |x|$ est évidemment une norme, et K sera toujours implicitement muni de cette norme.
2) Si F est un sous-espace vectoriel d'un e.v. n. $\big(E, \|\cdot\|\big)$, la restriction de $\|\cdot\|$ à F est une norme sur F. Lorsqu'on parle de l'e.v. n. F, c'est toujours de cette norme qu'il s'agit.
3) Soit E un espace préhilbertien sur K. Nous savons que $x \mapsto \|x\| := (\,x \mid x\,)^{1/2}$ est une norme sur E, appelée norme euclidienne si $K := \mathbb{R}$ et norme hermitienne si $K := \mathbb{C}$ (*cf.* le module I.4). Ainsi, sur le $\mathbb{R}$-espace vectoriel $\mathbb{C}$, $z \mapsto |z|$ est la norme euclidienne associée au produit scalaire $(z, z') \mapsto (z\bar{z}' + \bar{z}z')/2$. Rappelons trois cas particuliers importants.
a) Le produit scalaire usuel sur K^n conduit à la norme (euclidienne ou hermitienne) usuelle $\|\cdot\|$ sur K^n : $\|\mathbf{x}\| := \Big(|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2\Big)^{1/2}$ si $\mathbf{x} := (x_1, \ldots, x_n)$. Quand nous parlons de « l'e.v. n. K^n », c'est implicitement de cette norme qu'il s'agit.
b) Soient $I := [a, b]$ un intervalle fermé borné ($a < b$) et $\mathcal{C}(I, K)$ l'espace des fonctions

continues de I dans K. Pour toutes fonctions $f, g \in \mathcal{C}(I,K)$, posons :

$$(f \mid g) := \int_a^b \overline{f(x)} g(x)\,\mathrm{d}x \quad \text{et} \quad \|f\|_2 := \left(\int_a^b \left|f(x)\right|^2 \mathrm{d}x\right)^{1/2}. \tag{2}$$

La première formule définit un produit scalaire sur $\mathcal{C}(I,K)$, et la norme associée est $\|\cdot\|_2$.
c) Soit $\ell^2(K)$ l'espace des suites $u := (u_n)$ de K *de carré sommable*, *i.e.* telles que la série $\sum\limits_{n\geqslant 0} |u_n|^2$ converge. Si $u := (u_n)$ et $v := (v_n)$ appartiennent à $\ell^2(K)$, posons :

$$(u \mid v) := \sum_{n=0}^{+\infty} \bar{u}_n v_n \quad \text{et} \quad \|u\|_2 := \left(\sum_{n=0}^{+\infty} |u_n|^2\right)^{1/2}. \tag{3}$$

La première formule définit un produit scalaire sur $\ell^2(K)$, et la norme associée est $\|\cdot\|_2$.
4) L'espace vectoriel K^n peut être muni d'autres normes (utiles) que celle qui provient du produit scalaire canonique. Par exemple, pour tout $\mathbf{x} := (x_1, \ldots, x_n) \in K^n$, posons :

$$\|\mathbf{x}\|_\infty := \max\bigl(|x_1|, \ldots, |x_n|\bigr) \quad \text{et} \quad \|\mathbf{x}\|_1 := |x_1| + |x_2| + \cdots + |x_n|. \tag{4}$$

Nous laissons le lecteur vérifier que $\|\cdot\|_\infty$ et $\|\cdot\|_1$ sont bien des normes. Par exemple, si $\mathbf{x} := (x_1, \ldots, x_n)$ et $\mathbf{y} := (y_1, \ldots, y_n)$, on a $|x_j + y_j| \leqslant |x_j| + |y_j| \leqslant \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty$ pour tout $j \in [\![1, n]\!]$, d'où , en passant au maximum, $\|\mathbf{x} + \mathbf{y}\|_\infty \leqslant \|\mathbf{x}\|_\infty + \|\mathbf{y}\|_\infty$. Ces différentes normes seront « comparées » dans la section 1.7.
5) Soient $E_1, \ldots, E_n$ des espaces vectoriels munis de normes $\|\cdot\|_1, \ldots, \|\cdot\|_n$. On peut munir leur produit $E := E_1 \times E_2 \times \cdots \times E_n$ d'une norme, par exemple le lecteur vérifiera que chacune des formules ci-dessous définit une norme sur E :

$$\begin{cases} \|x\| := \max(\|x_1\|_1, \ldots, \|x_n\|_n) \\ \|x\|' := \bigl(\|x_1\|_1^2 + \cdots + \|x_n\|_n^2\bigr)^{1/2} \end{cases} \qquad \text{pour tout} \quad x := (x_1, \ldots, x_n) \in E. \tag{5}$$

Lorsque nous parlerons de l'e.v. n. E, nous supposerons implicitement que E est muni de la norme $\|\cdot\|$. Tout ce que nous dirons sur cet e.v. n. dans ce module reste vrai si l'on remplace la norme $\|\cdot\|$ par la norme $\|\cdot\|'$, le lecteur pourra aisément le vérifier.
6) Soient X un ensemble non vide et $\mathcal{B}(X,K)$ l'ensemble des applications *bornées* de X dans K. C'est un sous-espace vectoriel de $\mathcal{F}(X,K)$, espace vectoriel formé des applications de X dans K. Si $f \in \mathcal{B}(X,K)$, posons $\|f\|_\infty := \sup\limits_{x\in X} |f(x)|$. Puisque f est bornée, $\|f\|_\infty \in \mathbb{R}_+$. Nous obtenons une norme $\|\cdot\|_\infty$ sur $\mathcal{B}(X,K)$. Vérifions par exemple l'inégalité triangulaire. Soient $f, g \in \mathcal{B}(X,K)$. Pour tout $x \in X$, l'inégalité triangulaire dans K donne : $|(f+g)(x)| = |f(x) + g(x)| \leqslant |f(x)| + |g(x)| \leqslant \|f\|_\infty + \|g\|_\infty$. D'où $\|f+g\|_\infty \leqslant \|f\|_\infty + \|g\|_\infty$, en passant à la borne supérieure.
Cette norme, appelée *norme de la convergence uniforme*, est extrêmement importante et sera étudiée en détail dans le module suivant. En prenant $X := [\![1, n]\!]$ $(n \geqslant 1)$, on a $\mathcal{B}(X,K) = K^n$, et l'on retrouve la norme $\|\cdot\|_\infty$ de l'exemple 4. Voyons deux autres cas.
Prenons $X := \mathbb{N}$. Une application de $\mathbb{N}$ dans K est une suite de K. Ainsi $\mathcal{B}(\mathbb{N},K)$ est l'espace vectoriel des suites *bornées* de K, qui sera noté $\ell^\infty(K)$. Pour toute suite bornée $u := (u_n) \in \ell^\infty(K)$, on a donc : $\|u\|_\infty := \sup\limits_n |u_n|$.

Supposons que $X := [a,b]$, $a < b$. L'ensemble $\mathcal{C}\big([a,b],K\big)$ des fonctions continues (donc bornées) de $[a,b]$ dans K est un sous-espace vectoriel de $\mathcal{B}\big([a,b],K\big)$. Nous disposons déjà de deux normes sur $\mathcal{C}\big([a,b],K\big)$: $\|\cdot\|_2$ et $\|\cdot\|_\infty$. Le lecteur pourra vérifier que la formule suivante définit une autre norme sur $\mathcal{C}\big([a,b],K\big)$, nous la noterons $\|\cdot\|_1$:

$$\|f\|_1 \;:=\; \int_a^b |f(t)|\,\mathrm{d}t. \tag{6}$$

7) Sur un espace vectoriel E de dimension finie $n \geqslant 1$, il existe toujours des normes. Partons d'une norme $\|\cdot\|$ sur K^n (la norme usuelle ou une autre) et d'une base $\mathcal{B}$ de E. Si $x \in E$, soit $X \in K^n$ le vecteur formé des composantes de x dans la base $\mathcal{B}$, et posons $\|x\|_E := \|X\|$. Nous obtenons une norme $\|\cdot\|_E$ sur E. Pour l'inégalité triangulaire, par exemple, soient $x, y \in E$ et $X, Y \in K^n$ les vecteurs correspondants. Le vecteur formé des composantes de $x+y$ étant $X+Y$, l'inégalité $\|X+Y\| \leqslant \|X\| + \|Y\|$ entraîne $\|x+y\|_E \leqslant \|x\|_E + \|y\|_E$. Noter que $\|\cdot\|_E$ dépend du choix de la base $\mathcal{B}$. Par exemple, partant de la norme $\|\cdot\|_\infty$ sur K^n, on a $\|x\|_E := \max\big(|x_1|,\dots,|x_n|\big)$.

Dans l'espace usuel $\mathbb{R}^n$, le calcul de la norme d'un vecteur est aisé : c'est essentiellement le théorème de Pythagore. Dans un e.v. n. quelconque, ce n'est pas toujours aussi facile.

Exercice 1.
Soit $f : \mathbb{R} \to \mathbb{R}$ la fonction $x \mapsto xe^{-x^2}$. Montrer que f est bornée, *i.e.* $f \in \mathcal{B}(\mathbb{R},\mathbb{R})$, et calculer $\|f\|_\infty$ (*cf.* l'exemple 6).

Solution. Pour tout $x \in \mathbb{R}$, $f'(x) = (1-2x^2)e^{-x^2}$, donc f' ne s'annule qu'en $\pm 1/\sqrt{2}$. Ainsi f croît sur $\big[0, 1/\sqrt{2}\big]$ puis décroît sur $\big[1/\sqrt{2}, +\infty\big[$, de plus $\lim\limits_{+\infty} f = 0$.

La fonction f étant impaire, elle atteint son maximum absolu en $1/\sqrt{2}$ et son minimum absolu en $-1/\sqrt{2}$. D'où $\|f\|_\infty = \sup\limits_{x\in\mathbb{R}} |f(x)| = f(1/\sqrt{2}) = 1/\sqrt{2e}$.

1.2 Espaces métriques

Soient $\big(E, \|\cdot\|\big)$ un e.v. n. et $x, y \in E$. Comme dans le cas d'un espace préhilbertien (par exemple $\mathbb{R}^n$), la *distance* $d(x,y)$ de x à y est définie ainsi à partir de la norme :

$$d(x,y) \;:=\; \|x-y\|. \tag{7}$$

D'où une *fonction distance* $d : E \times E \to \mathbb{R}_+$, définie par $(x,y) \mapsto d(x,y)$. Les propriétés de la norme font que la fonction distance possède les propriétés (D1), (D2), (D3) ci-dessous. Par exemple, $d(x,z) = \|x-z\| = \|(x-y)+(y-z)\| \leqslant \|x-y\|+\|y-z\| = d(x,y)+d(y,z)$ pour tous $x, y, z \in E$. Cela suggère de donner la définition générale suivante :

Définition 2. On appelle *distance* sur un ensemble X toute application d de $X \times X$ dans $\mathbb{R}_+$ ayant les propriétés suivantes :
(D1) Pour tous $x, y \in X$, $d(y,x) = d(x,y)$ (symétrie).
(D2) Pour tous $x, y \in X$, $d(x,y) = 0$ si, et seulement si, $x = y$ (séparation).
(D3) Pour tous $x, y, z \in X$, $d(x,z) \leqslant d(x,y) + d(y,z)$ (inégalité triangulaire).
On appelle *espace métrique* tout couple (X,d) formé d'un ensemble X et d'une distance d sur X.

Voici une conséquence de ces propriétés $\big($*cf.* la preuve de la minoration (1)$\big)$:

$$d(x,y) \geqslant |d(x,z) - d(y,z)| \quad \text{pour tous} \quad x, y, z \in X. \tag{8}$$

En pratique, nous parlerons souvent de l'espace métrique X (ou même de l'espace X si le contexte le permet), sous-entendant la fonction distance, qui sera implicitement notée d.
Soit E un e.v. n.. Si l'on définit la fonction d par la formule (7), (E,d) est un espace métrique. Lorsque nous parlerons de l'e.v. n. E en tant qu'espace métrique, c'est toujours de (E,d) qu'il s'agira. Noter qu'on a : $d(x+a, y+a) = d(x,y)$ pour tous $a, x, y \in E$; on dit que d est *invariante par translation*. Cette propriété d'invariance n'a pas de sens dans un espace métrique quelconque (qui ne possède pas d'addition).
Soient (X,d) un espace métrique et A une partie de X. La restriction d' de d à $A \times A$ est une distance sur A. On dit que (A,d') est un *sous-espace métrique*, ou simplement un sous-espace, de (X,d). Lorsque nous parlerons de l'espace métrique A, il s'agira toujours de (A,d'). En pratique, d' sera aussi (abusivement) notée d, ou simplement omise, et nous dirons que A est un *sous-espace* de X. Cette notion de sous-espace est fondamentale. Par exemple, pour définir la continuité d'une fonction f définie sur une certaine sphère S de l'espace usuel, nous n'aurons pas besoin de faire référence à l'espace $\mathbb{R}^3$, il suffira de considérer S comme un espace métrique à part entière.
Toute partie d'un e.v. n. est ainsi un espace métrique, de façon naturelle. Par exemple, toute partie de $\mathbb{R}^n$ est un espace métrique, étant sous-entendu que la distance est définie par $d(x,y) := \|x - y\|$, où $\|\cdot\|$ est la norme euclidienne standard.
En pratique, tous les espaces métriques seront des parties d'espaces vectoriels normés, à part quelques exemples ou exercices. Mais alors, pourquoi parler d'espaces métriques plus généraux ? D'abord, cela ne complique pas la présentation des notions importantes (au contraire) et rend l'utilisation de ces notions beaucoup plus souple. Ensuite, c'est au fond naturel car, même si la Terre est (en gros) une sphère de $\mathbb{R}^3$, nos ancêtres ont mesuré des distances sur la Terre (supposée plane ou ronde, suivant les époques) bien avant de pouvoir en mesurer dans l'espace !

Exemples.

1. Sur un ensemble quelconque X, le lecteur vérifiera que les formules $d(x,y) := 0$ si $x = y$ et $d(x,y) := 1$ si $x \neq y$ définissent une distance d, appelée *distance discrète*.
2. Si $d_1, \ldots, d_n$ sont des distances sur X et si $\lambda_1, \ldots, \lambda_n$ sont des réels strictement positifs, il est clair que $\sum_{i=1}^{n} \lambda_i d_i$ est une distance sur X.

3. Soit (X,d) un espace métrique. Définissons $d' : X \times X \to \mathbb{R}$ par la formule $d'(x,y) := \min\big(1, d(x,y)\big)$. Alors d' est aussi une distance sur X. Seule l'inégalité triangulaire $d'(x,z) \leqslant d'(x,y) + d'(y,z)$ vaut d'être démontrée. C'est évident si $d(x,y) \geqslant 1$ ou $d(y,z) \geqslant 1$, car $d'(x,z) \leqslant 1$. Si $d(x,y) < 1$ et $d(y,z) < 1$, on a $d'(x,z) \leqslant d(x,z) \leqslant d(x,y) + d(y,z) = d'(x,y) + d'(y,z)$, d'où le résultat.
4. Soit S la sphère unité de l'espace euclidien usuel $\mathbb{R}^n$, $n \geqslant 1$. Notons $\mathbf{P}$ l'ensemble des droites vectorielles de $\mathbb{R}^n$, *i.e.* l'espace projectif $\mathbf{P}^{n-1}(\mathbb{R})$ (*cf.* le module I.6). Toute droite vectorielle $D \in \mathbf{P}$ contient exactement deux vecteurs unitaires, qui sont opposés. Soient $D, D' \in \mathbf{P}$. Choisissons $x, x' \in S$ tels que $D = \mathbb{R}x$ et $D' = \mathbb{R}x'$, puis posons $\delta(D,D') = \min\big(\|x - x'\|, \|x + x'\|\big)$. C'est licite, car cette expression ne change pas si l'on remplace x (ou x') par son opposé. Montrons que δ est une distance sur $\mathbf{P}$. La symétrie est évidente. La séparation l'est aussi, car $\delta(D,D') = 0$ si, et seulement si, $x' = \pm x$, ce qui équivaut à $D = D'$. Pour l'inégalité triangulaire, soient $D, D', D'' \in \mathbf{P}$. On peut choisir $x \in S \cap D$, puis $x' \in S \cap D'$ tel que $\delta(D,D') = \|x - x'\|$, et enfin $x'' \in S \cap D''$ tel que $\delta(D',D'') = \|x' - x''\|$. Alors :

$$\begin{aligned}\delta(D,D'') = \min\big(\|x - x''\|, \|x + x''\|\big) &\leqslant \|x - x''\| = \|(x - x') + (x' - x'')\| \\ &\leqslant \|x - x'\| + \|x' - x''\| = \delta(D,D') + \delta(D',D'').\end{aligned}$$

Proposition 1. Pour $i = 1, \dots, n$, soit (X_i, d_i) un espace métrique. Posons $X := X_1 \times X_2 \times \cdots \times X_n$. La formule suivante définit une distance d sur X :

$$d(x,y) := \max\big(d_1(x_1,y_1), \dots, d_n(x_n,y_n)\big) \tag{9}$$

pour tous $x := (x_1, \dots, x_n)$ et $y := (y_1, \dots, y_n)$ appartenant à X.

Démonstration. Soient $x := (x_1, \dots, x_n)$, $y := (y_1, \dots, y_n)$, $z := (z_1, \dots, z_n)$ trois éléments de X. Pour tout $i \in [\![1, n]\!]$, $d_i(x_i, z_i) \leqslant d_i(x_i, y_i) + d_i(y_i, z_i) \leqslant d(x,y) + d(y,z)$. D'où, en passant au maximum, $d(x,z) \leqslant d(x,y) + d(y,z)$. L'inégalité triangulaire est donc établie dans X. La vérification des deux autres propriétés (séparation et symétrie) est immédiate et laissée au lecteur. ∎

Nous supposerons toujours que X est muni de la distance d ainsi définie. Montrons comment généraliser au cas d'un produit *dénombrable* d'espaces métriques.

Soient $\big((X_n), d_n\big)_{n\in\mathbb{N}}$ une famille dénombrable d'espaces métriques et $X := \prod_{n\in\mathbb{N}} X_n$. Rappelons qu'un élément de X est une famille $x := (x_n)_{n\in\mathbb{N}}$ telle que chaque x_n appartient à X_n. Si $x := (x_n)$ et $y := (y_n)_n$ sont deux points de X, posons :

$$d(x,y) := \sum_{n=0}^{+\infty} 2^{-n} \min\big(1, d_n(x_n,y_n)\big). \tag{10}$$

Cette définition est licite, car le terme général de la série est majoré par 2^{-n}. On obtient ainsi une distance d sur X. La vérification de l'inégalité triangulaire, par exemple, est analogue à la précédente et laissée au lecteur. Dans la définition de d, on peut bien sûr remplacer 2 par n'importe quel réel $a > 1$ fixé, ce qui définit sur X une nouvelle distance.

Lorsque les (X_n, d_n) sont tous égaux à un même espace métrique (E, δ), X n'est autre que l'ensemble des suites de E (applications de $\mathbb{N}$ dans E), noté $\mathcal{F}(\mathbb{N}, E)$ ou $E^{\mathbb{N}}$.

- Si par exemple $E = [0,1]$, avec sa distance usuelle, l'espace métrique $\big([0,1]^{\mathbb{N}}, d\big)$ est appelé *cube de Hilbert*.
- Si $E = \{0,1\}$ muni de la distance discrète, l'espace métrique $\big(\{0,1\}^{\mathbb{N}}, d\big)$ est appelé *ensemble de Cantor*. L'ensemble de Cantor possède d'intéressantes propriétés, certaines d'entre elles seront étudiées dans la section 3 (voir aussi les exercices II.1.27 et II.1.43).

Disposant de distances, nous pouvons définir les boules et les sphères, comme dans $\mathbb{R}^n$ ou plus généralement dans un espace préhilbertien. Voici les définitions générales :

Définition 3. Soit a un point de X. Pour tout nombre réel $r > 0$, on pose :

$$B(a,r) := \big\{x \in X \mid d(a,x) < r\big\} \quad \text{(boule ouverte de centre } a \text{ et de rayon } r\text{)},$$

$$\overline{B}(a,r) := \big\{x \in X \mid d(a,x) \leqslant r\big\} \quad \text{(boule fermée de centre } a \text{ et de rayon } r\text{)},$$

$$S(a,r) := \big\{x \in X \mid d(a,x) = r\big\} \quad \text{(sphère de centre } a \text{ et de rayon } r\text{)}.$$

Dans un e.v. n. E, ces définitions se traduisent ainsi : $B(a,r) := \big\{x \in E \mid \|x-a\| < r\big\}$, $\overline{B}(a,r) := \big\{x \in E \mid \|x-a\| \leqslant r\big\}$ et $S(a,r) := \big\{x \in E \mid \|x-a\| = r\big\}$. La boule unité ouverte (resp. fermée) de E est $B(0,1)$ $\big($resp. $\overline{B}(0,1)\big)$, la sphère unité de E est $S(0,1) := \big\{x \in E \mid \|x\| = 1\big\}$, c'est l'ensemble des vecteurs *unitaires* de E. Tout vecteur non nul x s'écrit de façon unique $x := \lambda u$, où u est unitaire et $\lambda \in \mathbb{R}_+^*$ $\big($en fait $\lambda := \|x\|\big)$.

Noter à cet égard la particularité de la dimension 1. Supposons que $\dim E := 1$, et soit u un vecteur unitaire. Si $K := \mathbb{R}$, $S(0,1) = \{-u, u\}$ est formée de deux points alors que, si $K := \mathbb{C}$, on a $S(0,1) = \{\lambda u \mid \lambda \in \mathbb{U}\}$, où $\mathbb{U}$ est le cercle unité de $\mathbb{C}$.

La notion de boule est *relative*. Par exemple, $B(0,2) =]-2,2[$ dans $\mathbb{R}$, alors que $B(0,2) = [-1,2[$ dans $[-1,3[$. En général, soit A un sous-espace de X. Si $a \in A$ et $r > 0$, $B_A(a,r) = A \cap B_X(a,r)$, avec des notations évidentes.

Exercice 2.

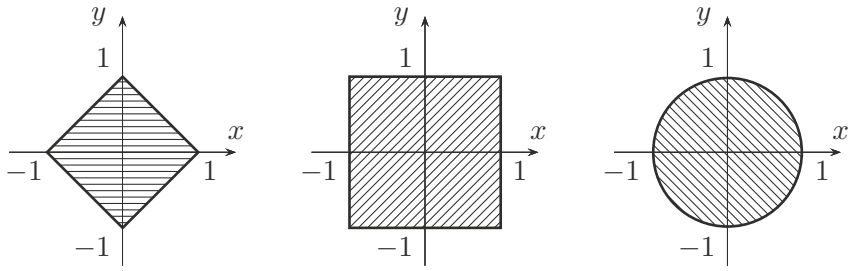
Si deux boules fermées $\overline{B}(a,r)$ et $\overline{B}(a',r')$ d'un e.v. n. $E \neq \{0\}$ sont égales, montrer que $a = a'$ et $r = r'$.

Solution. Supposons que $a = a'$ et, par exemple, $r' > r$. Soit $u \in E$ un vecteur unitaire. Alors $x := a + \big((r+r')/2\big)u$ appartient à $\overline{B}(a,r')$ mais pas à $\overline{B}(a,r)$, ce qui est absurde. Supposons ensuite que $a \neq a'$ et, par exemple, $r' \geqslant r$. Posons ici (faire un dessin) :

$$x := a' + \frac{r+r'}{2\|a-a'\|}\,(a'-a).$$

Alors $d(a',x) = (r+r')/2$ et $d(a,x) = d(a,a') + (r+r')/2$ (le vérifier). On en déduit que $x \in \overline{B}(a',r')$ mais $x \notin \overline{B}(a,r)$, car $d(a,x) \geqslant d(a,a') + r > r$. C'est encore absurde, et il ne reste que le cas où $a = a'$ et $r = r'$. Le résultat de cet exercice peut être faux pour des boules d'un espace métrique, *cf.* l'exercice II.1.5.

Les boules et sphères dépendent de la distance (ou la norme) choisie. Dans $\mathbb{R}^2$, représentons la boule unité fermée pour chacune des normes $\|\cdot\|_1$, $\|\cdot\|_\infty$ et $\|\cdot\|$ (de gauche à droite) :



$|x| + |y| \leqslant 1$ $\max(|x|, |y|) \leqslant 1$ $x^2 + y^2 \leqslant 1$

FIGURE II.1.1 – La boule unité dépend de la norme choisie.

Ces trois boules ne se ressemblent pas, mais nous verrons qu'elles possèdent les mêmes propriétés topologiques (compacité et convexité, notamment).

Définition 4. Soit A une partie non vide d'un espace métrique X. On appelle *diamètre* de A la borne *supérieure* $\mathrm{Diam}(A) := \sup\{d(a,b) \mid (a,b) \in A \times A\}$. Par convention, $\mathrm{Diam}(\varnothing) := 0$. Si ce diamètre est fini, on dit que A est *bornée*.

Le diamètre d'un singleton est 0. Dans $\mathbb{R}^2$, le diamètre d'un cercle de rayon R est $2R$ (d'où la terminologie), mais il est bon que le lecteur le prouve. Soit c un point de X quelconque. Pour que la partie A soit bornée, il faut, et il suffit, qu'elle soit contenue dans une boule de centre c. Si X est un e.v.n., on retrouve la définition habituelle : A est bornée si, et seulement s'il existe un réel $M > 0$ tel que $\|x\| \leqslant M$ pour tout $x \in A$.
Une application f d'un ensemble Z dans un espace métrique X est dite *bornée* si son image $f(Z)$ est une partie bornée de X.

Définition 5. Soient A et B deux parties non vides d'un espace métrique (X,d). On appelle *distance* de A et B le nombre $d(A,B)$ défini ainsi (c'est ici une borne *inférieure*) :

$$d(A,B) := \inf\{d(a,b) \mid (a,b) \in A \times B\}.$$

Si $A := \{a\}$, on écrit $d(a,B)$ au lieu de $d(\{a\},B)$. Si $A \cap B$ n'est pas vide, il est clair que $d(A,B) = 0$. La réciproque est fausse : dans $\mathbb{R}$, $d(0,\mathbb{R}^*) = 0$, mais $0 \notin \mathbb{R}^*$. La borne inférieure définissant $d(A,B)$ n'est donc pas toujours atteinte. En outre, si $a, b \in X$ sont distincts, $d(\{a\},\{a,b\}) = 0$, donc l'application $(A,B) \mapsto d(A,B)$ n'est pas une distance sur l'ensemble des parties non vides de X. La notation $d(A,B)$ est cependant utile.

1.3 Limites de suites dans un espace métrique

Dans la définition de la limite d'une suite réelle, remplaçons la valeur absolue par une distance. Nous obtenons tout naturellement la définition suivante :

Définition 6. Soient X un espace métrique et ℓ un point de X. On dit qu'une suite (u_n) à valeurs dans X *admet ℓ comme limite* ou *converge vers ℓ* si la suite réelle de terme général $d(u_n, \ell)$ converge vers zéro, ce qui se traduit ainsi :

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \geqslant N, \quad d(u_n, \ell) \leqslant \varepsilon. \tag{11}$$

Si $\varepsilon > 0$ est donné arbitrairement, les u_n doivent donc tous appartenir, à partir d'un certain rang (dépendant de ε), à la boule fermée $\overline{B}(\ell, \varepsilon)$. Dans un e.v. n. E, cela s'écrit :

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \geqslant N, \quad \|u_n - \ell\| \leqslant \varepsilon. \tag{12}$$

À partir de cette définition, ce qui a été dit pour les suites réelles se transpose sans difficulté à un espace métrique. Par exemple, si (u_n) converge vers $\ell \in X$ et aussi vers $\ell' \in X$, on a $\ell = \ell'$ (unicité de la limite) ; la preuve est la même que dans $\mathbb{R}$ (remplacer les valeurs absolues par des distances). On dit que (u_n) est *convergente* si elle possède une limite et *divergente* si elle n'en possède pas.

Une suite réelle peut avoir une limite infinie, par exemple $+\infty$. Dans un espace métrique, il n'y a pas d'analogue utile. Ainsi, dans un e.v. n., dire qu'une suite (u_n) a une limite infinie n'a pas de sens (penser à $\mathbb{R}^2$), mais la suite *réelle* $\big(\|u_n\|\big)$ peut avoir pour limite $+\infty$.

Dans un espace métrique, on ne peut pas faire la somme de deux suites. Dans un e.v. n., par contre, nous disposons du résultat suivant, dont la preuve est facile et laissée au lecteur (utiliser convenablement les suites réelles et ne pas mettre d'epsilon) :

Proposition 2. Dans un e.v. n., soient (u_n) et (v_n) deux suites convergentes, de limites respectives u et v. Pour tous $\lambda, \mu \in K$, la suite $(\lambda u_n + \mu v_n)$ converge vers $\lambda u + \mu v$.

Exemples.

1. Dans un espace métrique produit $X \times Y$, considérons une suite $\big((x_n, y_n)\big)$ et un point $(a, b) \in X \times Y$. Alors $\big((x_n, y_n)\big)$ converge vers (a, b) si, et seulement si, (u_n) converge vers a (dans X) et (y_n) converge vers b (dans Y). C'est évident, vu la formule $d\big((x_n, y_n), (a, b)\big) = \max\big(d(x_n, a), d(y_n, b)\big)$. La généralisation au cas d'un produit fini d'espaces métriques est facile et laissée au lecteur.
2. Pour tout $n \in \mathbb{N}$, notons f_n la fonction $x \mapsto x^n$. La suite (f_n) converge-t-elle ? Cette question est évidemment mal posée : tout dépend de l'espace vectoriel E dans lequel on opère, et aussi de la norme choisie. Prenons d'abord $E := \mathcal{C}\big([0, b], \mathbb{R}\big)$, avec $b > 0$ fixé, muni de la norme $\|\cdot\|_\infty$. Dans ce cas, $\|f_n\| = b^n$ pour tout n. Si $b < 1$, (f_n) converge vers 0 (fonction nulle). Si $b > 1$, la suite $\big(\|f_n\|_\infty\big)$ a pour limite $+\infty$ (f_n « s'éloigne indéfiniment »). Pour le cas $b := 1$, voir le module suivant. Changeons de norme : au lieu de $\|\cdot\|_\infty$, prenons la norme euclidienne $\|\cdot\|_2$

(exemple 3b) de la page 394). Pour tout n, $\|f_n\|_2^2 = \displaystyle\int_0^b x^{2n}\,\mathrm{d}x = b^{2n+1}/(2n+1)$. Il en résulte que (f_n) converge vers 0 si $b \leqslant 1$, et que $\big(\|f_n\|_2\big)$ a pour limite $+\infty$ si $b > 1$.

Une sous-suite de (u_n) est une suite $(u_{n_k})_{k\geqslant 0}$, où $(n_k)_{k\geqslant 0}$ est une suite strictement croissante de $\mathbb{N}$. Comme dans le cas réel, si une suite (u_n) converge, chacune de ses sous-suites converge et a même limite que (u_n). La notion de sous-suite est utile (penser au théorème de Bolzano-Weierstraß dans le cours de [L1]), elle suggère une définition :

Définition 7. Dans un espace métrique X, un point $a \in X$ est appelé *valeur d'adhérence* d'une suite (u_n) si la propriété suivante est vérifiée :

$$\forall \varepsilon > 0,\ \forall N \in \mathbb{N},\ \exists n \geqslant N, \quad d(u_n, a) \leqslant \varepsilon. \tag{13}$$

Il est essentiel de comparer avec la propriété (11). Nous avons juste interverti deux symboles $\forall$ et $\exists$, ce qui change tout ! Pour le comprendre, formulons autrement la définition :

PROPRIÉTÉ. *Le point a est valeur d'adhérence de (u_n) si, et seulement si, pour tout $\varepsilon > 0$, l'ensemble $\{n \in \mathbb{N} \mid d(u_n, a) \leqslant \varepsilon\}$ est infini.*

L'équivalence des deux formulations est évidente, car une partie T de $\mathbb{N}$ est infinie si, et seulement si, pour tout $N \in \mathbb{N}$, T contient (au moins) un entier $n \geqslant N$.

Proposition 3. Pour que a soit valeur d'adhérence d'une suite (u_n), il faut, et il suffit, que a soit limite d'une sous-suite de (u_n).

Démonstration. Supposons que a soit limite d'une sous-suite (u_{n_k}) de (u_n). Soient $\varepsilon > 0$ et $N \in \mathbb{N}$. Il existe $P \in \mathbb{N}$ tel que $(k \geqslant P) \implies \big(d(u_{n_k}, a) \leqslant \varepsilon\big)$. Puisque (n_k) a pour limite $+\infty$, il existe $k \in \mathbb{N}$ tel que $k \geqslant P$ et $n_k \geqslant N$. Alors $n := n_k$ convient : $n \geqslant N$ et $d(u_n, a) \leqslant \varepsilon$. Cela prouve que a est valeur d'adhérence de (u_n).
Supposons inversement que a soit valeur d'adhérence de (u_n). Définissons une suite (n_k) de $\mathbb{N}$ en stipulant que $n_0 := 0$ et que, pour tout $k \in \mathbb{N}$, n_{k+1} est le plus petit des entiers p strictement supérieurs à n_k et tels que $d(u_p, a) \leqslant 1/(k+1)$; il y a de tels entiers p, vu la propriété ci-dessus. Il est clair que $(u_{n_k})_{k\geqslant 0}$ converge vers a. ∎

Dans $\mathbb{R}$, la suite (n) n'a aucune valeur d'adhérence, car chacune de ses sous-suites a pour limite $+\infty$. La suite $\big((-1)^n\big)$ a deux valeurs d'adhérence : -1 et 1. Plus généralement, si r est un nombre rationnel, mis sous la forme $r := p/q$, où $p \in \mathbb{Z}$, $q \in \mathbb{N}^*$ et $p \wedge q = 1$, la suite complexe $\big(\exp(2\pi i n r)\big)$ a exactement q valeurs d'adhérence, à savoir ses q premiers termes (le vérifier). La suite $(0, 1, 0, 2, 0, 3, 0, 4, \ldots)$ n'a qu'une seule valeur d'adhérence, à savoir 0, et pourtant cette suite diverge. On a cependant le résultat suivant :

Remarque. Une suite convergente a une seule valeur d'adhérence, à savoir sa limite. Supposons en effet que (u_n) converge vers ℓ, et soit a une valeur d'adhérence de (u_n). D'après la proposition précédente, a est limite d'une sous-suite de (u_n), et nous savons que cette sous-suite converge alors vers ℓ. Par unicité de la limite, $a = \ell$.

Exercice 3.
Soit s une bijection de $\mathbb{N}$ sur $\mathbb{Q}$ (il en existe, car $\mathbb{Q}$ est dénombrable). Montrer que tout réel est valeur d'adhérence de la suite $\big(s(n)\big)$.
Solution. Soit $a \in \mathbb{R}$. Donnons-nous un réel $\varepsilon > 0$ et un entier $N \geqslant 0$.
L'intervalle $]a, a+\varepsilon]$ contient une infinité de nombres rationnels (*cf.* le cours de [L1]). Il existe donc $q \in \mathbb{Q}$ tel que $a < q \leqslant a+\varepsilon$ et $q \neq s(0), \ldots, s(N)$. Si $n := s^{-1}(q)$, on a donc $n \geqslant N+1$ et $\big|s(n)-a\big| = |q-a| \leqslant \varepsilon$. Ainsi, a est valeur d'adhérence de $\big(s(n)\big)$.

1.4 Parties ouvertes, parties fermées

Nous allons enrichir notre vocabulaire « topologique ». La définition d'une partie ouverte (resp. fermée) de $\mathbb{R}^n$ a été donnée dans le cours de [L1]. Sa généralisation est facile :

Définition 8. Soit X un espace métrique. Une partie U de X est dite *ouverte* si, pour tout $a \in U$, U contient une boule de centre a. Une partie F de X est dite *fermée* si son complémentaire $X \setminus F$ est une partie ouverte de X.

Le fait que U soit ouverte s'exprime donc ainsi : $\forall a \in U,\ \exists r > 0,\ B(a,r) \subset U$. Si une partie de X est ouverte (resp.fermée), on dira aussi que c'est un ouvert (resp. un fermé) de X. Voici quelques propriétés simples mais importantes.

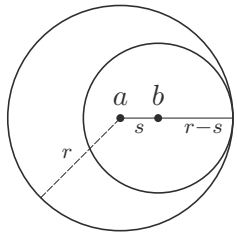
Proposition 4. Soit X un espace métrique. Alors :
(O1) La partie vide et X sont des ouverts.
(O2) Toute réunion d'ouverts est ouverte.
(O3) Toute intersection *finie* d'ouverts est ouverte.
Pour les fermés, on a des propriétés analogues ou « duales » :
(F1) La partie vide et X sont des fermés.
(F2) Toute intersection de fermés est fermée.
(F3) Toute réunion *finie* de fermés est fermée.

Démonstration. Soit $(A_i)_{i\in I}$ une famille de parties de X, de réunion R et d'intersection S. Supposons que tous les A_i soient ouverts, et soit $a \in R$. Il existe $i \in I$ tel que $a \in A_i$. Puisque A_i est ouverte, il existe $r > 0$ tel que $B(a,r) \subset A_i$. Alors $B(a,r) \subset R$, donc R est ouverte. Supposons ensuite que $I := \{i_1, \ldots, i_n\}$ soit fini, et soit $b \in S$. Si $k \in [\![1, n]\!]$, il existe $r_k > 0$ tel que $B(b, r_k) \subset A_{i_k}$. Alors $r := \min(r_1, \ldots, r_n) > 0$, et $B(b,r) \subset B(b,r_k) \subset A_{i_k}$ pour tout k, donc $B \subset S$. Cela montre que S est ouverte.
Supposons que les A_i soient fermés. Pour tout $i \in I$, $B_i := X \setminus A_i$ est donc un ouvert. La « dualité » évoquée utilise les formules suivantes : $\bigcup_{i\in I} (X \setminus A_i) = X \setminus \bigcap_{i\in I} A_i$ et $\bigcap_{i\in I} (X \setminus A_i) = X \setminus \bigcup_{i\in I} A_i$ (*cf.* le cours de [L1]). Les B_i étant ouverts, leur réunion $X \setminus S$ est ouverte (on vient de le voir), donc S est fermée. De même, si I est finie, l'intersection $X \setminus R$ des B_i est ouverte, donc R est fermée. Les propriétés (O1) et (F1) sont évidentes. ∎

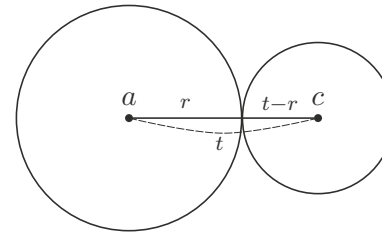
Soit X un espace métrique. Les points (singletons) sont fermés. En effet, si $a \in X$ et $b \in X \setminus \{a\} := U$, on a $r := d(a,b) > 0$, et $a \notin B(b,r)$, donc $B(b,r) \subset U$. Ainsi U est ouvert, *i.e.* $\{a\}$ est fermé.
Justifions maintenant la terminologie de « boule ouverte » et de « boule fermée » :

Proposition 5. Soient X un espace métrique et $a \in X$. Pour tout $r > 0$, la boule $B(a,r)$ est ouverte, alors que la boule $\overline{B}(a,r)$ et la sphère $S(a,r)$ sont fermées.

Démonstration. Soient $b \in B(a,r)$ et $s := d(a,b)$, donc $0 \leqslant s < r$. Si $x \in B(b,r-s)$, on a $d(a,x) \leqslant d(a,b) + d(b,x) = s + d(b,x) < r$, donc $x \in B(a,r)$. Ainsi $B(a,r)$ contient $B(b,r-s)$, ce qui montre que $B(a,r)$ est ouverte.
Soient $c \in X \setminus \overline{B}(a,r)$ et $t := d(a,c)$, de sorte que $t > r$. Si $y \in B(c,t-r)$, on a $t = d(a,c) \leqslant d(a,y) + d(y,c) < d(a,y) + (t-r)$, donc $d(a,y) > r$, *i.e.* $y \in X \setminus \overline{B}(a,r)$. Ainsi $B(c,t-r) \subset X \setminus \overline{B}(a,r)$, ce qui montre que $X \setminus \overline{B}(a,r)$ est ouvert, c'est-à-dire que $\overline{B}(a,r)$ est fermée. La sphère $S(a,r) = \overline{B}(a,r) \cap \big(X \setminus B(a,r)\big)$ est donc fermée. ∎



$B(a,r)$ est ouverte



$\overline{B}(a,r)$ est fermée

Ainsi $\mathbb{Z}$ est fermé dans $\mathbb{R}$, parce que $\mathbb{R} \setminus \mathbb{Z}$ est réunion des intervalles ouverts $]n, n+1[$, n parcourant $\mathbb{Z}$. Noter par ailleurs que, dans un e.v. n. non nul E, $\{0\}$ est l'intersection des boules ouvertes $B(0,r)$ pour $r > 0$, bien que $\{0\}$ ne soit pas ouvert (*cf.* la proposition 4 de la page précédente). Bien entendu, une partie n'est pas toujours ouverte ou fermée, par exemple $[0,1[$ n'est ni fermée ni ouverte dans $\mathbb{R}$.
Voici une importante caractérisation « séquentielle » (*i.e.* à l'aide de suites) des fermés, évitant le recours aux complémentaires, dans un espace métrique X :

Théorème 6. Une partie F de X est fermée si, et seulement si, elle contient la limite de toute suite *de* F convergeant *dans* X (*i.e.* si elle est « stable par passage à la limite »).

Démonstration. Supposons que F soit fermée, et soit (u_n) une suite de F convergeant vers $\ell \in X$. Montrons que $\ell \in F$. Supposons que $\ell \in X \setminus F$. Puisque $X \setminus F$ est ouvert, il existe $r > 0$ tel que $B(\ell,r) \subset X \setminus F$. Il existe ensuite un rang N à partir duquel tous les u_n sont dans $B(\ell,r)$. Mais alors $u_N \in B(\ell,r) \subset X \setminus F$, et c'est absurde car $u_N \in F$.
Supposons maintenant que F ne soit pas fermée. Il existe donc un point $a \in X \setminus F$ tel qu'aucune boule de centre a ne soit contenue dans $X \setminus F$. Autrement dit, une telle boule rencontre toujours F. Pour tout $n \in \mathbb{N}^*$, soit donc $u_n \in F \cap B(a,1/n)$. La suite (u_n) converge évidemment vers a, donc F n'est pas stable par passage à la limite, car $a \notin F$. ∎

Ce théorème montre, par exemple, que la diagonale $\Delta(X) := \{(x,x) \mid x \in X\}$ de X est fermée dans $X \times X$, vu l'exemple 1 de la page 401. Soient A, B deux parties de l'espace métrique X. Si B est fermée, $A \cap B$ est fermée *dans* A (cela résulte aussitôt du théorème), et l'on en déduit que si B est ouverte, $A \cap B$ est ouverte *dans* A (remplacer B par $X \setminus B$). Pour une réciproque, voir l'exercice II.1.8.

Exemple. Soit A une partie fermée non vide de $\mathbb{R}$. Si A est majorée (resp. minorée), sa borne supérieure (resp. inférieure) appartient à A. En effet, si par exemple A est majorée, sa borne supérieure M est finie. Pour tout $n \in \mathbb{N}^*$, $M - 1/n$ ne majore pas A. Il existe donc $a_n \in A$ tel que $M - 1/n \leqslant a_n \leqslant M$. La suite (a_n) de A converge alors vers M, d'où $M \in A$ puisque A est fermée.

Dans ce qui suit, A désigne une partie d'un espace X. Voici deux définitions « duales » :

Théorème et définition 7. Un point x de X est dit *intérieur* à A si A contient une boule de centre a. L'ensemble des points de X intérieurs à A est appelé *intérieur* de A et noté $\mathrm{Int}(A)$, c'est (au sens de l'inclusion) le plus grand ouvert de X contenu dans A.

Théorème et définition 8. Un point x de X est dit *adhérent* à A s'il est limite d'une suite de A. L'ensemble des points de X adhérents à A est appelé *adhérence* de A et noté $\overline{A}$, c'est (au sens de l'inclusion) le plus petit fermé de X contenant A.

Ainsi, A est fermée (resp. ouverte) si, et seulement si, $\overline{A} = A$ $\big($resp. $\mathrm{Int}(A) = A\big)$.

Démonstration. Soit W la réunion des ouverts de X contenus dans A. C'est bien, au sens de l'inclusion, le plus grand ouvert de X contenu dans A. Soit $x \in X$. Si $x \in W$, il existe un ouvert U contenu dans A et contenant x, puis un $r > 0$ tel que $B(x,r) \subset U$. Alors $B(x,r) \subset A$, donc x est intérieur à A. Inversement, si x est intérieur à A, A contient une boule ouverte de centre x. Cette boule est un ouvert contenu dans A, donc dans W, et par suite $x \in W$. Ainsi $W = \mathrm{Int}(A)$, ce qui établit le premier théorème.
Soit Φ le complémentaire de $\mathrm{Int}(X \setminus A)$. C'est donc un fermé, qui contient A puisque $\mathrm{Int}(X \setminus A) \subset X \setminus A$. Si F est un fermé contenant A, $X \setminus F$ est un ouvert contenu dans $X \setminus A$, donc dans $\mathrm{Int}(X \setminus A)$. Ainsi $X \setminus F \subset \mathrm{Int}(X \setminus A) = X \setminus \Phi$, d'où $F \supset \Phi$.
Soit enfin $x \in X$. Si x est adhérent à A, x est limite d'une suite de A (donc de Φ), ce qui montre que $x \in \Phi$, car Φ est fermé. Inversement, supposons que $x \in \Phi$, c'est-à-dire que $x \notin \mathrm{Int}(X \setminus A)$. Pour tout $n \in \mathbb{N}^*$, la boule $B(x, 1/n)$ n'est donc pas contenue dans $X \setminus A$, *i.e.* il existe un point $u_n \in A \cap B(x, 1/n)$. La suite (u_n) est une suite de A convergeant vers x, donc x est adhérent à A. Ainsi $\Phi = \overline{A}$, ce qui établit le second théorème.
Cette preuve donne en prime la formule $\overline{A} = X \setminus \mathrm{Int}(X \setminus A)$, soit $\mathrm{Int}(X \setminus A) = X \setminus \overline{A}$ puis, en remplaçant A par $X \setminus A$: $\overline{(X \setminus A)} = X \setminus \mathrm{Int}(A)$. ∎

Si A n'est pas vide, $\overline{A} = \big\{x \in X \mid d(x,A) = 0\big\}$. Soit en effet $x \in X$. Puisque $d(x,A) = \inf\limits_{a \in A} d(x,a)$, $d(x,A) = 0$ si, et seulement s'il existe une suite (a_n) de A telle que $\lim\limits_{n \to +\infty} d(x, a_n) = 0$, *i.e.* $\lim\limits_{n \to +\infty} a_n = x$, ce qui revient à dire que x est adhérent à A.

Exercice 4.
Soit E un e.v. n. non nul et $a \in E$. Pour tout $r > 0$, montrer que l'adhérence de $B(a,r)$ est $\overline{B}(a,r)$, que l'intérieur de $\overline{B}(a,r)$ est $B(a,r)$, et que $S(a,r)$ est d'intérieur vide.

Solution. Puisque $\overline{B}(a,r)$ est fermée et contient $B(a,r)$, elle contient $\overline{B(a,r)}$. Soit $x \in \overline{B}(a,r)$. Si $x \in B(a,r)$, *a fortiori* $x \in \overline{B(a,r)}$. Sinon, $\|x-a\| = r$. Pour tout $n \in \mathbb{N}^*$, soit u_n l'image de a par l'homothétie de centre x et de rapport $1/n$. On vérifie que $\|x-u_n\| = r/n$ et $\|a-u_n\| = r(n-1)/n < r$ (faire un dessin). Alors (u_n) est une suite de $B(a,r)$ convergeant vers x, donc x est adhérent à $B(a,r)$. Cela prouve la formule $\overline{B(a,r)} = \overline{B}(a,r)$, ce qui, en passant, justifie la notation $\overline{B}(a,r)$. La preuve de l'égalité $\mathrm{Int}\big(\overline{B}(a,r)\big) = B(a,r)$ est analogue, nous la laissons au lecteur.
Soit enfin $x \in S(a,r)$. Pour tout $n \in \mathbb{N}^*$, définissons u_n comme ci-dessus. Puisque $\|x-u_n\| = r/n$ et $\|u_n - a\| < r$, la boule $\overline{B}(x, r/n)$ n'est pas contenue dans $S(a,r)$. On en déduit que x n'est pas intérieur à $S(a,r)$, et par suite l'intérieur de $S(a,r)$ est vide. Dans un espace métrique, une sphère peut avoir un intérieur non vide, *cf.* l'exercice II.1.5.

Remarque. Soit E un e.v. n.. L'adhérence d'un sous-espace vectoriel V de E est un sous-espace vectoriel de E. En effet, soient $x, y \in \overline{V}$ et $\lambda, \mu \in K$. Il existe deux suites (x_n) et (y_n) de V convergeant vers x et y respectivement. La suite $(\lambda x_n + \mu y_n)$ converge vers $\lambda x + \mu y$, et $\lambda x_n + \mu y_n \in V$ pour tout n, donc $\lambda x + \mu y \in \overline{V}$.

Définition 9. On dit que A est *dense* (ou partout dense) dans X si $\overline{A} := X$, autrement dit si tout point de X est limite d'une suite de points de A.

Les modules II.2, II.4 et le module « Polynômes orthogonaux » de [L2-L3] montreront clairement l'importance de cette définition. Si une partie A de X est dense, tout point x de X est limite d'une suite de A, *i.e.* x peut être approché, d'aussi près qu'on veut, par un point de A.
L'équivalence des deux formes de la définition vient du théorème précédent. Par exemple, $\mathbb{Q}$ est une partie dense de $\mathbb{R}$, car tout réel x est limite d'une suite de nombres rationnels (par exemple la suite (u_n) formée des approximations décimales de x à 10^{-n} près par défaut). De même $\mathbb{R} \setminus \mathbb{Q}$ est dense dans $\mathbb{R}$, car tout $x \in \mathbb{Q}$ est limite d'une suite de nombres irrationnels, par exemple $\big(x + 2^{-n}\sqrt{2}\big)$. Ainsi $\overline{\mathbb{R} \setminus \mathbb{Q}}$ contient $\mathbb{Q}$, et aussi $\mathbb{R} \setminus \mathbb{Q}$.

Proposition 9. Soit V un sous-groupe additif de $\mathbb{R}$. Ou bien V est dense dans $\mathbb{R}$, ou bien V est de la forme $a\mathbb{Z}$, où a est un réel positif.

Démonstration. On peut supposer V non nul. Alors $A := V \cap]0, +\infty[$ n'est pas vide, car, si $x \in V$ et $x < 0$, $-x$ appartient aussi à V. Soit $m := \inf A$. Distinguons deux cas.
Supposons d'abord m strictement positif et montrons que $V = m\mathbb{Z}$. Par définition d'une borne inférieure, il existe $x \in A$ tel que $m \leqslant x < 2m$. Alors $x - m \in V$ (car V est un

sous-groupe de $\mathbb{R}$) et $0 \leqslant x - m < m$. Mais m minore A, donc $x - m \notin A$ et par suite $x = m$. Ainsi $m \in V$, d'où $m\mathbb{Z} \subset V$. Inversement, soit $y \in V$. Posons $k := [y/m]$, de sorte que $y/m = k + u$, avec $u \in [0, 1[$, donc $y = km + um$, avec $um \in [0, m[$. Maintenant, $y - km \in V \cap [0, m[= \{0\}$, d'où $y = km \in m\mathbb{Z}$.

Supposons que $m = 0$, et montrons que V est dense dans $\mathbb{R}$. Soit $t \in \mathbb{R}$. Étant donné $\varepsilon > 0$, il s'agit de montrer que $V \cap]t - \varepsilon, t + \varepsilon[$ n'est pas vide. Puisque $\inf A = 0$, il existe $m' \in V$ tel que $0 < m' \leqslant \varepsilon$. En posant $k := [t/m']$, on a comme ci-dessus $0 \leqslant t - km' < m'$, d'où $|t - km'| < m' \leqslant \varepsilon$, donc $km' \in V \cap]t - \varepsilon, t + \varepsilon[$. ∎

Définition 10. Un espace métrique est dit *séparable* s'il contient une partie *dense* finie ou dénombrable.

Ainsi $\mathbb{R}$ est séparable puisque $\mathbb{Q}$ est dénombrable et dense dans $\mathbb{R}$.
Voici (encore !) trois définitions utiles.

Définition 11.
On appelle *frontière* de A l'ensemble des points de $\overline{A}$ n'appartenant pas à $\mathrm{Int}(A)$.
On note $\partial(A)$ cette frontière, de sorte que $\partial(A) := \overline{A} \setminus \mathrm{Int}(A) = \overline{A} \cap \overline{(X \setminus A)}$.

La deuxième égalité a été prouvée plus haut, elle montre que $\partial(A)$ est fermée. Noter que $\mathrm{Int}(A)$, $\mathrm{Int}\big(X \setminus A\big)$ et ∂A forment une partition de X, puisque $\mathrm{Int}\big(X \setminus A\big) = X \setminus \overline{A}$ et $\partial A = \overline{A} \setminus \mathrm{Int}(A)$.
Dans un e.v. n., la frontière d'une boule ouverte $B(a, r)$ est la sphère $S(a, r)$, puisque $\overline{B(a, r)} = \overline{B}(a, r)$ et $\mathrm{Int}\big(B(a, r)\big) = B(a, r)$. De même, la boule fermée $\overline{B}(a, r)$ et la sphère $S(a, r)$ ont pour frontière $S(a, r)$. Dans $\mathbb{R}$, les frontières respectives de $]0, 1[$, $\{0, 1\}$ et $\mathbb{Q}$ sont $\{0, 1\}$, $\{0, 1\}$ et $\mathbb{R}$ (le vérifier).

Définition 12. On appelle *voisinage* d'un point a de X toute partie V de X contenant une boule de centre a (autrement dit, V contient un ouvert contenant a).

Noter que toute intersection *finie* $V := V_1 \cap V_2 \cap \cdots \cap V_p$ de voisinages de a est un voisinage de a. En effet, pour chaque i, il existe $r_i > 0$ tel que V_i contienne $B(a, r_i)$, et alors V contient $B(a, r)$, où $r := \min(r_1, \ldots, r_p)$.
Cette notion de voisinage simplifie parfois le langage. Par exemple, un point $x \in X$ est adhérent à A si, et seulement si, tout voisinage de x rencontre A, chacune de ces deux conditions se traduisant ainsi : $\forall r > 0,\ B(x, r) \cap A \neq \varnothing$ (le vérifier). De même une partie de X est ouverte si, et seulement si, c'est un voisinage de chacun de ses points. Autre exemple, déjà connu dans $\mathbb{R}$: par définition, une fonction $f : X \to \mathbb{R}$ présente un maximum relatif en a si, et seulement si, $f \leqslant f(a)$ au voisinage de a. C'est plus court que la phrase suivante, qui a le même sens : $\exists r > 0,\ \forall x \in X, \big(d(a, x) < r\big) \Longrightarrow \big(f(x) \leqslant f(a)\big)$!

Définition 13. Une partie A de X est dite *discrète* si tout point $a \in A$ est le centre d'une boule B telle que $A \cap B = \{a\}$.

Ainsi, $\mathbb{Z}$ est une partie discrète de $\mathbb{R}$ car, pour tout $n \in \mathbb{Z}$, $\{n\} = \mathbb{Z} \cap]n-1, n+1[$. L'espace X lui-même est discret si, et seulement si, tous ses points sont ouverts.

Toutes les notions introduites dans cette section sont *relatives*. Par exemple, $]0,1[$ n'est pas fermé dans $\mathbb{R}$, mais c'est un fermé de l'espace $]0,1[$. Ainsi, « fermé » ne veut rien dire dans l'absolu. Il faut dire « fermé de… » ou « fermé dans… », sauf si le contexte ne laisse aucune ambiguïté.

Exercice 5.
Soient S, T deux parties de X, $S \subset T$. Si T est fermée dans X et S est fermée dans T, montrer que S est fermée dans X. Remplacer ensuite « fermée » par « ouverte ».

Solution. Traitons le cas « fermé », à l'aide du théorème 6 de la page 404. Soit (u_n) une suite de S, convergeant vers un point $a \in X$. C'est aussi une suite de T ; comme T est fermée dans X, a appartient déjà à T. Maintenant (u_n) est une suite de S convergeant vers $a \in T$, donc $a \in S$ puisque S est fermée dans T. Ainsi S est fermée dans X.
Passons au cas « ouvert ». Soit $x \in S$. Puisque S est ouverte dans T, il existe $r > 0$ tel que S contienne la boule ouverte *de* T de centre x et de rayon r, c'est-à-dire $T \cap B(x,r)$. Puisque T est ouverte dans X, il existe $r' > 0$ tel que $B(x,r') \subset T$. Il est alors clair que S contient $B\big(x, \min(r,r')\big)$ (le vérifier). Ainsi S est ouverte dans X.

1.5 Applications continues

Notre objectif est de *définir* la continuité d'une application d'un espace métrique dans un autre, et de forger des outils permettant de travailler avec cette notion. Comme dans $\mathbb{R}$, il faut d'abord parler de limites d'applications (terme préféré à « fonction » dans ce cadre).
Partons d'une situation simple, une fonction $f : I := [0,1[\to \mathbb{R}$. Nous savons ce qu'est la limite (éventuelle) de f en 1, mais parler de limite de f en 5 semble déraisonnable : il y a beaucoup de points de I « voisins » de 1, mais c'est faux pour 5. Avec le vocabulaire de la section précédente, le point 1 est *adhérent* à I, mais pas le point 5. À partir de là, on devine comment généraliser la définition d'une limite.
Voici le matériel qui est donné : deux espaces métriques (X,d) et (Y,δ), une partie A de X, une application $f : A \to Y$, un point a de X *adhérent à* A ($a \in \overline{A}$), et enfin un point ℓ de Y (c'est la « future limite »).
Avec ces hypothèses, voici la généralisation évoquée :

Définition 14. On dit que f a pour limite ℓ en a, ou que $f(x)$ tend vers ℓ quand x tend vers a, ce que l'on écrit $\ell = \lim\limits_{x \to a} f(x)$, si la propriété suivante est vérifiée :

$$\forall \varepsilon > 0,\ \exists \alpha > 0,\ \forall x \in A, \quad \big(d(x,a) \leqslant \alpha\big) \Longrightarrow \big(\delta\big(f(x),\ell\big) \leqslant \varepsilon\big). \qquad (14)$$

Noter que, si $a \in A$, f ne peut avoir d'autre limite en a que $f(a)$. On peut traduire la propriété (14) ainsi : $f(x)$ est aussi voisin de ℓ que l'on veut, pourvu que $x \in A$ soit assez voisin de a. En effet, soit $W \subset Y$ un voisinage de ℓ donné. Par définition d'un voisinage, il existe $\varepsilon > 0$ tel que $B(\ell,\varepsilon) \subset W$. La propriété (14) affirme qu'il existe un $\alpha > 0$ tel que :

$$\forall x \in A, \quad \big(d(x,a) \leqslant \alpha\big) \Longrightarrow \big(\delta\big(f(x),\ell\big) \leqslant \varepsilon\big).$$

Un tel α étant choisi, $V := B(a,\varepsilon) \subset X$ est un voisinage de a et, pour tout $x \in A \cap V$, $f(x) \in B(\ell,\varepsilon) \subset W$. Ainsi, $(x \in A \cap V) \Longrightarrow \big(f(x) \in W\big)$.
En fait, on peut souvent éviter les $\varepsilon, \alpha, \ldots$, grâce à l'important critère séquentiel suivant :

Théorème 10.
L'application f a pour limite ℓ en a si, et seulement si, pour toute suite (u_n) de A convergeant vers a (une telle suite existe), la suite $\big(f(u_n)\big)$ converge vers ℓ.

Démonstration. Il existe une suite de A convergeant vers a car $a \in \overline{A}$. Supposons que f ait pour limite ℓ en a, et soit (u_n) une suite de A convergeant vers a.
Soit $\varepsilon > 0$; prenons $\alpha > 0$ tel que :

$$\forall x \in A, \quad \big(d(x,a) \leqslant \alpha\big) \Longrightarrow \big(\delta\big(f(x),\ell\big) \leqslant \varepsilon\big).$$

Puisque (u_n) converge vers a, il existe $N \in \mathbb{N}$ tel que $n \geqslant N$ implique $d(u_n,a) \leqslant \alpha$ et donc $\delta\big(f(u_n),\ell\big) \leqslant \varepsilon$. Ainsi, $\big(f(u_n)\big)$ converge vers ℓ.
Supposons ensuite que f n'ait pas pour limite ℓ en a. En prenant la négation de la propriété (14), on a donc : $\exists \varepsilon > 0,\ \forall \alpha > 0,\ \exists x \in A,\ \big(d(x,a) \leqslant \alpha \text{ et } \delta\big(f(x),\ell\big) > \varepsilon\big)$. Considérons un $\varepsilon > 0$ dont l'existence est ainsi assurée. Soit $n \in \mathbb{N}^*$. Le choix $\alpha := 1/n$ montre qu'il existe $a_n \in A$ tel que $d(a_n,a) \leqslant 1/n$ et $\delta\big(f(a_n),\ell\big) > \varepsilon$. Alors (a_n) converge vers a, mais $\big(f(u_n)\big)$ ne converge pas vers ℓ puisque $\delta\big(f(a_n),\ell\big) > \varepsilon$ pour tout $n \geqslant 1$. Cela prouve, par contraposition, la suffisance de la propriété (14). ∎

Pour la continuité en un point, on applique ce qui précède avec $A := X$. On obtient ceci :

Définition 15.
Soit $f : X \to Y$ une application entre deux espaces métriques. On dit que f est continue en un point $a \in X$ si $f(x)$ tend vers $f(a)$ lorsque $x \in X$ tend vers a. Si f est continue en tout point de X, on dit que f est *continue* (ou continue sur X).

Théorème 11.
L'application $f : X \to Y$ est continue au point $a \in X$ si, et seulement si, pour toute suite (u_n) de X convergeant vers a, la suite $\big(f(u_n)\big)$ converge vers $f(a)$.

Voici des conséquences immédiates. Pour que $f : X \to Y$ soit continue il faut, et il suffit, qu'elle transforme toute suite convergente de X en une suite convergente de Y. Si f est continue, sa restriction à $A \subset X$ l'est aussi (toute suite de A convergeant dans A converge dans X).
Si $f : X \to Y$ est continue, son graphe $\Gamma := \big\{(x,y) \mid x \in X\big\}$ est une partie fermée de $X \times Y$. Soit en effet $\big(x_n,y_n)\big)_n$ une suite de Γ convergeant vers $(x,y) \in X \times Y$, montrons que $(x,y) \in \Gamma$. Par hypothèse, $(x_n)_n$ converge vers x et $(y_n)_n$ converge vers y. Pour tout n, $y_n = f(x_n)$, donc $\big(f(x_n)\big)_n$ converge vers y. D'où $y = f(x)$, c'est-à-dire $(x,y) \in \Gamma$, puisque $\big(f(x_n)\big)_n$ converge vers $f(x)$, vu la continuité de f. La réciproque est

fausse. Ainsi, la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(0) = 0$ et $f(x) = 1/x$ si $x \neq 0$ n'est pas continue en 0. Cependant son graphe, qui est la réunion de $\{(0,0)\}$ et de l'hyperbole d'équation $xy = 1$, est un fermé de $\mathbb{R}^2$ (le vérifier).
Dans ce qui suit, $X, Y, Z, X', \ldots$ désignent des espaces métriques.

Proposition 12. Deux applications continues $f, g : X \to Y$ coïncidant sur une partie *dense* A de X sont égales.

Démonstration. En effet, soit $x \in X$. Il existe une suite (u_n) de A convergeant vers x. Puisque f et g sont continues en x, $\big(f(u_n)\big)$ et $\big(g(u_n)\big)$ convergent vers $f(x)$ et $g(x)$ respectivement. Par hypothèse, $f(u_n) = g(u_n)$ pour tout n, donc $f(x) = g(x)$. ∎

Proposition 13. Si $f : X \to Y$ est continue en $a \in X$ et si $g : Y \to Z$ est continue en $f(a)$, $g \circ f$ est continue en a. La composée de deux applications continues est continue.

Démonstration. C'est une simple application du théorème. Soit en effet (u_n) une suite de X convergeant vers a. Puisque f est continue en a, $\big(f(u_n)\big)$ converge vers $f(a)$; puisque g est continue en $f(a)$, $\big(g\big(f(u_n)\big)\big)$ converge vers $g\big(f(a)\big)$. ∎

Enfin, en utilisant la proposition 2 de la page 401, on obtient le résultat suivant :

Proposition 14. Soient X un espace métrique et E un e.v. n.. Si $f, g : X \to E$ sont continues et si $\lambda, \mu \in K$, alors $\lambda f + \mu g$ est continue.

Les applications continues de X dans E forment donc un sous-espace vectoriel de $\mathcal{F}(X, E)$, nous le noterons $\mathcal{C}(X, E)$.
Voyons un nouvel outil pour établir la continuité d'une application $f : (X, d) \to (X', d')$. Étant donné un réel $k \geqslant 0$, considérons la propriété suivante :

$$\forall x \in X,\ \forall y \in X, \quad d'\big(f(x), f(y)\big) \leqslant k d(x, y). \tag{15}$$

Si cette propriété est vraie, f est continue. En effet, si $k := 0$, f est constante donc continue. Supposons $k > 0$ et soit $\varepsilon > 0$. Posons $\alpha := \varepsilon/k$. Si $a, x \in X$ et $d(x, a) \leqslant \alpha$, on a $d'\big(f(x), f(a)\big) \leqslant k d(x, a) \leqslant k\alpha = \varepsilon$, ce qui montre que f est continue en a.

Définition 16.

1. Soit $k \in \mathbb{R}_+$. On dit que f est k-lipschitzienne si la propriété (15) est vérifiée.
2. On dit que f est *lipschitzienne* s'il existe $k \in \mathbb{R}_+$ tel que f soit k-lipschitzienne.
3. On dit que f est *contractante* s'il existe $k \in [0, 1[$ tel que f soit k-lipschitzienne.

Toute fonction lipschitzienne est donc continue. Nous verrons que la réciproque est fausse.

Exemples. 1) Soient I un intervalle de $\mathbb{R}$ et $k \in \mathbb{R}_+$. Une fonction dérivable $f : I \to \mathbb{R}$ est k-lipschitzienne si, et seulement si, $|f'| \leqslant k$. La nécessité est évidente : si $x \in I$ est fixé,

faire tendre y vers x dans l'inégalité $|f(x)-f(y)| \leqslant k\,|x-y|$. Si $|f'| \leqslant k$, l'inégalité des accroissements finis donne $|f(x)-f(y)| \leqslant k\,|x-y|$ pour tous $x, y \in I$, donc f est k-lipschitzienne. Si I est fermé borné et si f est de classe $\mathcal{C}^1$, elle est M-lipschitzienne, M étant la borne supérieure (finie) de $|f'|$ sur I.

2) La fonction $f : x \mapsto x^2$ est continue mais n'est pas lipschitzienne sur $\mathbb{R}$. Soit en effet $k > 0$. L'inégalité $\left|x^2 - y^2\right| \leqslant k\,|x-y|$ est fausse si $x := 2k$ et $y := 0$, par exemple. D'après la remarque 1), on peut aussi conclure en remarquant que f' n'est pas bornée.

3) Soient E un e.v. n. et $\lambda \in K$. L'homothétie $x \mapsto \lambda x$ est $|\lambda|$-lipschitzienne.

4) Soit A une partie non vide d'un espace métrique X. La fonction $x \mapsto d(x, A)$ est 1-lipschitzienne, donc continue. En effet, soient $x, y \in X$. On a $d(x, a) \leqslant d(x, y) + d(y, a)$ pour tout $a \in A$, d'où $d(x, A) \leqslant d(x, y) + d(y, A)$ en passant à la borne inférieure pour $a \in A$. De même $d(y, A) \leqslant d(y, x) + d(x, A)$, d'où $|d(x, A) - d(y, A)| \leqslant d(x, y)$.

5) Soient (X_1, d_1) et (X_2, d_2) deux espaces métriques et $(X_1 \times X_2, d)$ leur produit. La projection $p : (x_1, x_2) \mapsto x_1$ de $X_1 \times X_2$ sur X_1 est 1-lipschitzienne, donc continue. En effet, si $x_1, y_1 \in X_1$ et $x_2, y_2 \in X_2$, $d_1\big(p\big(x_1, x_2\big)\big), p\big(y_1, y_2\big)\big) = d_1(x_1, y_1)$ est majoré par $d\big((x_1, x_2), (y_1, y_2)\big) := \max\big(d_1(x_1, y_1), d_1(x_2, y_2)\big)$.

6) Soit S la sphère unité de l'espace euclidien usuel $\mathbb{R}^n$, $n \geqslant 1$. Munissons $\mathbf{P}^{n-1}(\mathbb{R})$ de la distance δ de l'exemple 4 de la page 397. Pour tous $x, y \in S$, on a par définition $\delta(\mathbb{R}x, \mathbb{R}x') \leqslant \|x - x'\| = d(x, x')$. L'application $x \mapsto \mathbb{R}x$ de S dans $\mathbf{P}^{n-1}(\mathbb{R})$ est donc 1-lipschitzienne, en particulier continue.

7) Une application $f : (X, d) \to (X', d')$ vérifiant $d'\big(f(x), f(y)\big) = d(x, y)$ pour tous $x, y \in X$ est appelée une *isométrie*. Une isométrie est bien sûr injective, et elle est 1-lipschitzienne, donc continue. Dans un e.v. n., toute translation $x \mapsto x + a$ est une isométrie. Dans un produit $X \times Y$, si $y \in Y$ est fixé, $x \mapsto (x, y)$ est une isométrie de X dans $X \times Y$.

Proposition 15. Si (X, d) est un espace métrique, la fonction distance $d : X \times X \to \mathbb{R}$ est 2-lipschitzienne, donc continue. De même, si $(E, \|\cdot\|)$ est un e.v. n., la fonction norme $\|\cdot\| : E \to \mathbb{R}$ est 1-lipschitzienne, donc continue.

Démonstration. Pour E, c'est évident : pour tous $x, y \in E$, on a $\big|\|x\| - \|y\|\big| \leqslant \|x - y\|$, *cf.* l'inégalité (1). Pour X, notons D la distance sur $X \times X$ (*cf.* la proposition1 de la page 398). Considérons deux points $z := (x, y)$ et $z' := (x', y')$ de $X \times X$. Par définition, $D(z, z') = \max\big(d(x, x'), d(y, y')\big)$. Alors $d(x, y) \leqslant d(x, x') + d(x', y') + d(y', y)$, d'où $d(x, y) - d(x', y') \leqslant d(x, x') + d(y, y') \leqslant 2D(z, z')$, et aussi $d(x', y') - d(x, y) \leqslant 2D(z, z')$ par symétrie. Ainsi $\big|d(x, y) - d(x', y')\big| \leqslant 2D(z, z')$, d'où la conclusion. ∎

Voici une autre caractérisation utile de la continuité :

Théorème 16. Une application $f : (X, d) \to (Y, \delta)$ est continue si, et seulement si, pour tout fermé (resp. ouvert) F de Y, $f^{-1}(F)$ est fermé (resp. ouvert) dans X.

Démonstration. Rappelons que $f^{-1}(F) := \big\{x \in X \mid f(x) \in F\big\}$. Supposons que f soit continue et soit F un fermé de Y. Soit (u_n) une suite de $f^{-1}(F)$ convergeant vers $a \in X$.

Alors $\big(f(u_n)\big)$ converge vers $f(a)$; comme $f(u_n) \in F$ pour tout n, $f(a) \in \overline{F} = F$, ce qui montre que $f^{-1}(F)$ est fermé dans X. Si V est un ouvert de Y, la formule évidente $X \setminus f^{-1}(V) = f^{-1}(Y \setminus V)$ montre alors que $f^{-1}(V)$ est ouvert dans X.
Supposons inversement que l'image réciproque de tout fermé de Y soit fermée dans X. Par dualité, cela revient à dire que l'image réciproque de tout ouvert de Y est ouverte dans X. Soient $a \in X$ et $\varepsilon > 0$. L'image réciproque par f de la boule $B\big(f(a), \varepsilon\big)$ étant ouverte et contenant a, elle contient une boule $B(a, \alpha)$ de centre a. Alors $\big(x \in B(a, \alpha)\big)$ implique $\big(f(x) \in B\big(f(a), \varepsilon\big)\big)$, ce qui montre que f est continue en a. ∎

Ce théorème fournit un procédé commode pour montrer qu'une partie est fermée (ou ouverte). Ainsi, soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction continue. Puisque $\{0\}$ est fermé dans $\mathbb{R}$, l'hypersurface d'équation $f(x) = 0$ est fermée dans $\mathbb{R}^n$. Soient $f : X \to \mathbb{R}$ une fonction continue et $t \in \mathbb{R}$. Puisque $]t, +\infty[$ est ouvert alors que $[t, +\infty[$ et $\{t\}$ sont fermés dans $\mathbb{R}$, $\{x \in X \mid f(x) > t\}$ est un ouvert de X, alors que $\{x \in X \mid f(x) \geqslant t\}$ et $\{x \in X \mid f(x) = t\}$ sont fermés. Ainsi, par exemple, le demi-espace de $\mathbb{R}^3$ défini par $x - 2y + z > 0$ est ouvert. Dans le module I.6, nous verrons que, dans $M_n(\mathbb{R})$, $GL_n(\mathbb{R})$ est ouvert et $SL_n(\mathbb{R})$ est fermé, à cause de la continuité de l'application $\det : M_n(\mathbb{R}) \to \mathbb{R}$.

Exercice 6.

Soient $X, Y_1, \ldots, Y_n$ et $Y := Y_1 \times Y_2 \times \cdots \times Y_n$. Montrer qu'une application $f : X \to Y$ est continue si, et seulement si, chacune de ses composantes $f_i : X \to Y_i$ l'est.

Solution. Pour tout i, la $i^{ème}$ projection $(y_1, \ldots, y_n) \mapsto y_i$ est une application continue p_i de Y dans Y_i et, par définition, $f_i = p_i \circ f$. Si f est continue, chaque f_i est donc continue, comme composée d'applications continues. Supposons inversement que chaque f_i soit continue. Pour établir la continuité de f, il suffit de montrer que, si $(x_k)_k$ est une suite de X convergeant vers x, la suite $\big(f(x_k)\big)_k$ converge vers $f(x)$ (théorème 11 de la page 409). Or, pour $i = 1, \ldots, n$, la suite $\big(f_i(x_k)\big)_k$ converge vers $f_i(x)$. La conclusion résulte donc de l'exemple 1 de la page 401, puisque $f(x_k) = \big(f_1(x_k), \ldots, f_n(x_k)\big)$ pour tout $k \in \mathbb{N}$.

Exercice 7.

Soient $n, p \in \mathbb{N}^*$ et $k \in \mathbb{N}$.
Montrer que $\big\{A \in M_{n,p}(\mathbb{R}) \mid \operatorname{rang} A \geqslant k\big\}$ est ouvert dans $M_{n,p}(\mathbb{R})$.
Que dire de $\big\{A \in M_{n,p}(\mathbb{R}) \mid \operatorname{rang} A \leqslant k\big\}$ et $\big\{A \in M_{n,p}(\mathbb{R}) \mid \operatorname{rang} A = k\big\}$?

Solution. Posons $U_k := \big\{A \in M_{n,p}(\mathbb{R}) \mid \operatorname{rang} A \geqslant k\big\}$. Puisque $U_0 = M_{n,p}(\mathbb{R})$ et U_k est vide si $k > \min(n, p)$, on peut supposer que $1 \leqslant k \leqslant \min(n, p)$. Pour donner un sens à l'énoncé, on identifie $M_{n,p}(\mathbb{R})$ à $\mathbb{R}^{np}$, muni de sa norme usuelle (en réalité, le choix de la norme importe peu, *cf.* la section 4.1). Soit $A \in M_{n,p}(\mathbb{R})$. On sait que le rang de A est supérieur ou égal à k si, et seulement si, A possède un mineur d'ordre k non nul (*cf.* le module II.7 du cours de [L1]). Si L (resp. C) est une partie à k éléments de $[\![1, n]\!]$ $\big($resp. $[\![1, p]\!]\big)$, notons D_L^C le mineur de A obtenu en prenant dans A les lignes d'indices appartenant à L et les colonnes d'indices appartenant à C. Si L et C sont fixées,

l'application $A \mapsto D_L^C$ de $M_{n,p}(\mathbb{R})$ dans $\mathbb{R}$ est polynomiale, donc continue. Il en résulte que $U(L,C) := \left\{A \in M_{n,p}(\mathbb{R}) \mid D_L^C \neq 0\right\}$ est un ouvert de $M_{n,p}(\mathbb{R})$. Or U_k est la réunion des $U(L,C)$ lorsque L (resp. C) décrit l'ensemble des parties à k éléments de $[\![1,n]\!]$ (resp. $[\![1,p]\!]$), donc U_k est ouvert, comme réunion d'ouverts.
L'ensemble $\left\{A \in M_{n,p}(\mathbb{R}) \mid \operatorname{rang} A \leqslant k\right\}$ est le complémentaire de U_{k+1}, il est donc fermé. Par contre, si $0 < k < \min(n,p)$, l'ensemble $E_k := \left\{A \in M_{n,p}(\mathbb{R}) \mid \operatorname{rang} A = k\right\}$ n'est ni ouvert ni fermé. Prenons par exemple $n = p = 2$ et $k = 1$ (le cas général est analogue). Lorsque l'entier $m \geqslant 2$ tend vers $+\infty$, la matrice $\begin{pmatrix} 1/m & 0 \\ 0 & 0 \end{pmatrix} \in E_1$ tend vers $0 \notin E_1$, donc E_1 n'est pas fermé.
De même, $\begin{pmatrix} 1/m & 0 \\ 0 & 1 \end{pmatrix} \notin E_1$ tend vers $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \in E_1$, donc le complémentaire de E_1 n'est pas fermé, *i.e.* E_1 n'est pas ouvert.

Définition 17. On dit qu'une application $f : X \to Y$ est un *homéomorphisme* de X sur Y si elle est bijective, continue, et si sa réciproque f^{-1} est continue. Deux espaces métriques sont dits *homéomorphes* s'il existe un homéomorphisme de l'un sur l'autre.

Par exemple, le logarithme et l'exponentielle sont deux homéomorphismes réciproques entre $]0,+\infty[$ et $\mathbb{R}$. La fonction $\tanh$ est un homéomorphisme de $\mathbb{R}$ sur $]-1,1[$, car la bijection réciproque $\operatorname{argth} :]-1,1[\to \mathbb{R}$ est continue. Plus généralement, soient I un intervalle de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction continue strictement monotone. Alors f est un homéomorphisme de I sur l'intervalle image $f(I)$ (*cf.* le module IV.2 du cours de [L1]). Dans $\mathbb{R}^2$ un carré (plein) et un disque fermé sont homéomorphes (*cf.* l'exercice II.1.30). Une isométrie surjective f est un homéomorphisme : elle est bijective, et f^{-1} est aussi une isométrie. Dans un e.v. n., une homothétie $f : x \mapsto \lambda x$ ($\lambda \in K^*$ fixé) est un homéomorphisme, car f et sa réciproque $x \mapsto \lambda^{-1}x$ sont lipschitziennes ; f est une isométrie si, et seulement si, $|\lambda| = 1$. Enfin, la composée de deux homéomorphismes est un homéomorphisme, et la réciproque d'un homéomorphisme est aussi un homéomorphisme.
Il n'est pas toujours facile de montrer que deux espaces *ne sont pas* homéomorphes. Ainsi $[0,1]$ et $\mathbb{R}$ ne sont pas homéomorphes, car toute fonction continue de $[0,1]$ dans $\mathbb{R}$ est bornée. De même, $\mathbb{R}$ n'est pas homéomorphe à un cercle du plan, mais il faut un argument de compacité (section 3), ou de connexité comme pour $\mathbb{R}$ et $\mathbb{R}^2$ (section 5.2).
Soient $n,p \in \mathbb{N}^*$, $n \neq p$. Les espaces métriques $\mathbb{R}^p$ et $\mathbb{R}^n$ ne sont pas homéomorphes, mais la preuve dépasse nettement le cadre de cet ouvrage (c'est du domaine de la « topologie algébrique »). À titre de comparaison, le fait que $\mathbb{R}^p$ et $\mathbb{R}^n$ ne soient pas isomorphes comme espaces vectoriels est élémentaire : ils n'ont pas même dimension. C'est un exercice facile de montrer que $\mathbb{R}^p$ et $\mathbb{R}^n$ ne sont pas non plus *difféomorphes* : il n'existe pas de $\mathcal{C}^1$-difféomorphisme de $\mathbb{R}^p$ sur $\mathbb{R}^n$. Le lecteur pourra se reporter au module II.5.

Disons qu'une propriété est *topologique* si elle est invariante par homéomorphisme. Ainsi, pour une suite, la convergence est une propriété topologique : si (u_n) est une suite conver-

gente de X et si f est un homéomorphisme de X sur Y, la suite $\big(f(u_n)\big)$ converge car f est continue. La propriété « A est fermé dans X » est topologique : si f est un homéomorphisme de X sur Y, la continuité de f^{-1} montre, en vertu du théorème précédent, que $f(A)$ est fermé dans Y. Nous rencontrerons plus loin d'autres propriétés topologiques.
Certaines applications, sans être lipschitziennes, sont « un peu mieux » que continues :

Définition 18. Une application $f : (X, d) \to (X', d')$ est dite *uniformément continue* si elle vérifie la propriété suivante :

$$\forall \varepsilon > 0,\ \exists \alpha > 0,\ \forall x \in X,\ \forall y \in X,\ \big(d(x,y) \leqslant \alpha\big) \Longrightarrow \big(d'\big(f(x), f(y)\big) \leqslant \varepsilon\big). \quad (16)$$

Comparons avec le fait que f soit continue *en tout point* $x \in X$, ce qui se traduit ainsi :

$$\forall x \in X,\ \forall \varepsilon > 0,\ \exists \alpha > 0,\ \forall y \in X,\ \big(d(x,y) \leqslant \alpha\big) \Longrightarrow \big(d'\big(f(x), f(y)\big) \leqslant \varepsilon\big)\ (*).$$

Ainsi, « $\forall x \in X$ » a joué à saute-mouton ! Dans la propriété $(*)$, le nombre α dont l'existence est assurée dépend à la fois de x et de α (introduits avant lui) ; dans la propriété (16), au contraire, α ne dépend que du seul ε, *i.e.* on impose que α ne dépende plus de x. Cela montre que l'uniforme continuité implique la continuité. Par ailleurs, une application lipschitzienne est uniformément continue (examiner la preuve de sa continuité). Ainsi :

$$\textit{(lipschitzienne)} \Longrightarrow \textit{(uniformément continue)} \Longrightarrow \textit{(continue)}.$$

La fonction $x \mapsto x^2$ est continue mais pas uniformément continue. Sinon, en prenant un point, simplement $\varepsilon := 1$, il existerait $\alpha > 0$ tel que $|x - y| \leqslant \alpha$ implique $\big|x^2 - y^2\big| \leqslant 1$. Choisissant $x := y + \alpha$, on aurait $\big|x^2 - y^2\big| = \alpha\,|2y + \alpha| \leqslant 1$ pour tout $y \in \mathbb{R}$, ce qui est absurde.
La fonction $x \mapsto \sqrt{x}$ est uniformément continue sur $[0, 1]$ (*cf.* le théorème de Heine de la page 439) ; elle n'est pas lipschitzienne, car sa dérivée n'est pas bornée sur $]0, 1]$.

Un théorème de prolongement

Soient X un espace métrique, A une partie non vide de X et $f : A \to \mathbb{R}$ une fonction continue. Existe-t-il une fonction continue $F : X \to \mathbb{R}$ prolongeant f, *i.e.* telle que $F(a) = f(a)$ pour tout $a \in A$? La réponse est non en général : prendre $X = \mathbb{R}$, $A = \mathbb{R}^*$ et $f : x \mapsto 1/x$. Voici un cas particulier important où la réponse est positive :

Théorème 17 (Théorème de prolongement de Tietze-Urysohn).
Supposons que A soit *fermée* dans X, et soit $f : A \to [u, v]$ une fonction continue $(u \leqslant v)$. Il existe une fonction continue $F : X \to [u, v]$ prolongeant f.

Démonstration. À l'aide d'une transformation affine convenable $x \mapsto ax + b$ (où $a \neq 0$), on se ramène aisément au cas où $[u, v] = [-1, 1]$. Procédons alors en plusieurs étapes.

1. Soient Y, Z deux parties disjointes et fermées de X. Montrons qu'il existe une fonction continue $\varphi : X \to [0, 1]$ valant 0 sur Y et 1 sur Z.
   Si Y est vide, $\varphi := 1$ convient, de même $\varphi := 0$ convient si Z est vide. Supposons Y, Z non vides. La fonction $\varphi : x \mapsto d(x, Y)/\big[d(x, Y) + d(x, Z)\big]$ répond alors à la question. D'abord, la définition de φ est licite, *i.e.* le dénominateur ne s'annule jamais. En effet, si x appartient à Y, il n'appartient pas à Z, donc $d(x, Z) > 0$ puisque Z est fermée. De même

$(x \in Z) \Longrightarrow \big(d(x,Y) > 0\big)$. Ensuite φ est continue, car $x \mapsto d(x,Y)$ et $x \mapsto d(x,Z)$ le sont. Il est clair enfin que $\varphi(X) \subset [0,1]$ et que φ est nulle sur Y et vaut 1 sur Z.

2. Considérons un réel $t > 0$ et une fonction continue $g : A \to [-t,t]$. Il existe une fonction continue $h : X \to [-t/3, t/3]$ telle que $|g(a) - h(a)| \leqslant 2t/3$ pour tout $a \in A$.

   Posons $Y := g^{-1}\big([-t,-t/3]\big)$ et $Z := g^{-1}\big([t/3,t]\big)$. Ce sont deux parties disjointes et fermées de A, donc de X. D'après le point 1), il existe $\varphi : X \to [0,1]$ valant 0 sur Y et 1 sur Z. La fonction continue $h := t(2\varphi - 1)/3$ convient. D'abord $|2\varphi - 1| \leqslant 1$, donc $|h| \leqslant t/3$. Soit ensuite $a \in A$. Si $a \in Y$, $g(a) \in [-t,-t/3]$ et $h(a) = -t/3$, donc $|g(a) - h(a)| \leqslant 2t/3$. Si $a \in Z$, $g(a) \in [t/3,t]$ et $h(a) = t/3$, donc $|g(a) - h(a)| \leqslant 2t/3$. Enfin, si $a \notin Y \cup Z$, $|g(a)| \leqslant t/3$ et $|h(a)| \leqslant t/3$, d'où la même conclusion.

3. Appliquons le point 2) en prenant $t := 1$ et $g := f$. Il existe une fonction continue $h_1 : X \to \mathbb{R}$ telle que $|h_1| \leqslant 1/3$ sur X et $|h_1 - f| \leqslant 2/3$ sur A. Prenons ensuite $t := 2/3$ et $g := f - h_1$. Il existe une fonction continue $h_2 : X \to \mathbb{R}$ telle que $|h_2| \leqslant 2/3^2$ sur X et $|h_1 + h_2 - f| \leqslant 2^2/3^2$ sur A. En itérant ce procédé, on construit une suite $(h_n)_{n\geqslant 1}$ de fonctions continues de X dans $\mathbb{R}$ telle que, pour tout $n \geqslant 1$, on ait $|h_n| \leqslant 2^{n-1}/3^n$ sur X et $|h_1 + h_2 + \cdots + h_n - f| \leqslant 2^n/3^n$ sur A. Pour passer de h_n à h_{n+1}, on applique le point 2) en prenant $g := f - (h_1 + h_2 + \cdots + h_n)$ et $t := 2^n/3^n$.

4. Anticipons sur le module II.2, en considérant sur X la série de fonctions $\sum_{n\geqslant 1} h_n$.

   Cette série converge normalement (et donc uniformément) sur X, à cause des inégalités $|h_n| \leqslant 2^{n-1}/3^n$. La somme de cette série est donc une fonction continue $F : X \to \mathbb{R}$ (théorème 8 de la page 480). Soit $a \in A$. Pour tout $n \geqslant 1$, on a $|h_1(a) + h_2(a) + \cdots + h_n(a) - f(a)| \leqslant 2^n/3^n$. En faisant tendre n vers $+\infty$, on en déduit que $F(a) = f(a)$, ce qui montre que F prolonge f. ∎

1.6 Applications (multi)linéaires continues

Si une application entre deux e.v. n. est *linéaire*, il est assez simple de savoir si elle est continue, comme le montre le théorème essentiel suivant :

Théorème 18. Une application *linéaire* L d'un e.v. n. E dans un e.v. n. F est continue si, et seulement s'il existe un réel $C \geqslant 0$ tel que $\|L(x)\|_F \leqslant C\|x\|_E$ pour tout $x \in E$.

Démonstration. Noter que l'on peut remplacer $C \geqslant 0$ par $C > 0$. Si L et C sont comme dans l'énoncé, L est C-lipschitzienne, donc continue. En effet, la linéarité de L entraîne :

$$\forall x, y \in E, \quad d\big(L(x),L(y)\big) = \|L(x)-L(y)\|_F = \|L(x-y)\|_F \leqslant C\|x-y\|_E = Cd(x,y).$$

Supposons inversement que L soit continue. En particulier L est continue en 0. Prenons $\varepsilon := 1$ dans la définition de la limite de L en 0. Il existe $\alpha > 0$ tel que, pour $y \in E$, $\big(\|y\|_E \leqslant \alpha\big) \Longrightarrow \big(\|L(y)\|_F \leqslant 1\big)$. Posons $C := 1/\alpha$, et soit $x \in E$. Vérifions que $\|L(x)\|_F \leqslant C\|x\|_E$. C'est évident si $x := 0$, car $L(0) = 0$. Si $x \neq 0$, soit $y := \alpha x/\|x\|_E$, de sorte que $\|y\|_E = \alpha$. Alors $\|L(y)\|_F \leqslant 1$, d'où :

$$\|L(x)\|_F = \alpha^{-1}\|x\|_E\,\|L(y)\|_F = C\|x\|_E\,\|L(y)\|_F \leqslant C\|x\|_E.$$

∎

Noter que cette preuve montre qu'une application *linéaire* continue en 0 est continue. Donnons dès maintenant un corollaire qui sera généralisé dans la section 4 :

Corollaire 19. Toute application linéaire f de K^n dans un e.v. n. F est *continue*.

Démonstration. Notons $(e_1, \dots, e_n)$ la base canonique de K^n. Considérons un vecteur $\mathbf{u} := (x_1, \dots, x_n)$ de K^n, de sorte que : $|x_1|^2 + \cdots + |x_n|^2 = \|\mathbf{u}\|^2$, et donc $|x_i| \leqslant \|\mathbf{u}\|$ pour tout i. En posant $C := \sum_{i=1}^{n} \|f(e_i)\|$, la continuité de f résulte des calculs suivants :

$$\|f(\mathbf{u})\| = \left\| \sum_{i=1}^{n} x_i f(e_i) \right\| \leqslant \sum_{i=1}^{n} \|x_i f(e_i)\| = \sum_{i=1}^{n} |x_i| \, \|f(e_i)\| \leqslant \|\mathbf{u}\| \sum_{i=1}^{n} \|f(e_i)\| = C\|\mathbf{u}\|.$$

Dans la suite, $E, F, \dots$ sont des e.v. n., et l'on omet les indices dans les normes, comme dans $\|x\|_E$. Soit $L \in \mathcal{L}(E, F)$. Puisque $L(0) = 0$, L est continue si, et seulement si, quand x décrit $E \setminus \{0\}$, le quotient $\|L(x)\|/\|x\|$ est borné. Cela justifie la définition suivante :

Définition 19. Soit L une application linéaire continue d'un e.v. n. E dans un e.v. n. F. On appelle *norme fonctionnelle* de L le réel $|||L|||$ défini ainsi :

$$|||L||| := \sup_{x \neq 0} \frac{\|L(x)\|}{\|x\|}. \tag{17}$$

La borne supérieure porte sur l'ensemble des vecteurs non nuls de E. Voici une conséquence évidente, mais importante, de cette définition :

$$\|L(x)\| \leqslant |||L||| \, \|x\| \quad \text{pour tout} \quad x \in E. \tag{18}$$

En fait, la borne supérieure en (17) a toujours un sens, mais elle est finie si, et seulement si, L est continue. Plusieurs formules voisines permettent dans ce cas de calculer $|||L|||$:

PROPRIÉTÉ. *Soit L une application linéaire d'un e.v. n. E dans un e.v. n. F. Pour que L soit continue, il faut, et il suffit qu'elle soit bornée sur la sphère unité (resp. la boule unité fermée) de E. Si tel est le cas, la norme $|||L|||$ peut être calculée de plusieurs manières :*

$$|||L||| := \sup_{x \neq 0} \frac{\|L(x)\|}{\|x\|} = \sup_{\|x\|=1} \|L(x)\| = \sup_{\|x\| \leqslant 1} \|L(x)\|. \tag{19}$$

En effet, notons dans l'ordre a, b, c les bornes supérieures (finies ou non) de l'énoncé ; $|||L||| = a$ par définition. La borne supérieure d'une partie de $\mathbb{R}$ augmente si l'on agrandit cette partie, donc $b \leqslant c \leqslant a$. Soit $x \in E \setminus \{0\}$, $x := \alpha u$, où $\alpha := \|x\|$ et $u := x/\|x\|$. Alors $\|u\| = 1$, donc $\|L(u)\| \leqslant b$, d'où $\|L(x)\| = \|\alpha L(u)\| = \|x\|\|L(u)\| \leqslant b\|x\|$, c'est-à-dire $\|L(x)\|/\|x\| \leqslant b$. Passons à la borne supérieure : $a \leqslant b$, d'où $a = b = c$.
Calculer $|||L|||$ revient à trouver la meilleure borne possible pour la fonction $x \mapsto \|L(x)\|$ sur la sphère unité de E. Ce calcul dépend des normes choisies sur E et F, même si la notation « simplifiée » $|||L|||$ ne fait apparaître ni E ni F ni les normes choisies sur E et F.

Proposition 20. Une forme linéaire f sur un e.v. n. E est continue si, et seulement si, son noyau est fermé.

Démonstration. Soit $H := \operatorname{Ker} f$. Si f est continue, H est fermé, car $\{0\}$ est un fermé de K et $H = f^{-1}\big(\{0\}\big)$. Supposons inversement que H soit fermé. Si f est nulle, elle est continue. Sinon, soit $a \in E$ tel que $f(a) \neq 0$, de sorte que $E = H \oplus Ka$. Puisque $a \notin H$, il existe $r > 0$ tel que $B(a,r) \cap H = \varnothing$. Soit $x \in E$ un vecteur de norme 1. Il s'écrit $x = h + \lambda a$, où $h \in H$ et $\lambda \in K$. Si $\lambda \neq 0$, on a $a - \lambda^{-1}x = -\lambda^{-1}h \in H$, donc $a - \lambda^{-1}x \notin B(a,r)$, c'est-à-dire $\big|\lambda^{-1}\big| \geqslant r$, soit $|\lambda| \leqslant r^{-1}$, ce qui reste vrai si $\lambda = 0$. D'où $|f(x)| = |f(h) + \lambda f(a)| = |\lambda f(a)| \leqslant r^{-1}\,|f(a)|$. Ainsi, f est bornée sur la sphère unité, elle est donc continue, en vertu de la propriété précédente. ∎

Exemples.

1. Dans un e.v. n. E, l'addition $E \times E \to E$ est continue, à cause des inégalités : $\|x+y\| \leqslant \|x\| + \|y\| \leqslant 2\max\big(\|x\|, \|y\|\big) = 2\|(x,y)\|$. Une homothétie $x \mapsto \lambda x$ est continue et de norme $|\lambda|$.

2. Soient E un espace préhilbertien et $a \in E$. L'application $\varphi_a : x \mapsto (\,a \mid x\,)$ est une forme linéaire sur E. Pour tout x, $|(\,a \mid x\,)| \leqslant \|a\|\|x\|$, donc $\varphi_a : E \to K$ est continue et $\|\!|\varphi_a|\!\| \leqslant \|a\|$. En fait, $\|\!|\varphi_a|\!\| = \|a\|$, car $|(\,a \mid a\,)| = \|a\|^2$. Le noyau de φ_a étant l'orthogonal $a^\perp$ de a, la proposition précédente montre que $a^\perp$ est fermé.

3. Soit $L \in \mathcal{L}(\mathbb{R}^2)$ l'application linéaire $(x,y) \mapsto (x+2y, 3x+4y)$. Ici $\mathbb{R}^2$ est muni de sa norme usuelle. Posons $X := (x,y)$ et $Y := L(X) = (x+2y, 3x+4y)$. Alors $\|Y\|^2 := Q(X)$, où $Q : (x,y) \mapsto (x+2y)^2 + (3x+4y)^2 = 10x^2 + 28xy + 20y^2$ est une forme quadratique sur $\mathbb{R}^2$. La matrice de Q dans la base canonique de $\mathbb{R}^2$ est $M := \begin{pmatrix} 10 & 14 \\ 14 & 20 \end{pmatrix}$. Les valeurs propres de M sont $\lambda := 15 + \sqrt{221}$ et $\mu := 15 - \sqrt{221}$. Dans le module I.4 nous avons montré que, sur le cercle unité C de $\mathbb{R}^2$, la borne supérieure (resp. inférieure) de Q est λ (resp. μ). La borne supérieure de $\|L\|$ sur C est donc $\sqrt{\lambda}$. Cette borne étant finie, L est continue, et l'on a :
$$\|\!|L|\!\| = \sup_{X \in C} \|L(X)\| = \sup_{X \in C} \sqrt{Q(X)} = \sqrt{15 + \sqrt{221}}.$$

4. Considérons l'espace $E := \mathcal{C}\big([a,b], \mathbb{R}\big)$ muni de la norme $\|\cdot\|_\infty$ (exemple 6 de la page 395). L'application $L : f \mapsto \displaystyle\int_a^b f$ de E dans $\mathbb{R}$ est linéaire. Pour toute fonction $f \in E$, on a $\left|\displaystyle\int_a^b f(x)\,\mathrm{d}x\right| \leqslant \displaystyle\int_a^b |f(x)|\,\mathrm{d}x \leqslant \int_a^b \|f\|_\infty = (b-a)\|f\|_\infty$.
Ainsi d'une part L est continue, et d'autre part $\|\!|L|\!\| \leqslant b-a$. En fait $\|\!|L|\!\| = b-a$ (prendre $f := 1$).

Proposition 21. On munit $\mathbb{C}^n$ de sa norme hermitienne usuelle $\|\cdot\|$ et $M_n(\mathbb{C})$ de la norme correspondante $|||\cdot|||$. Si $M \in M_n(\mathbb{C})$, on note $\rho(M)$ le plus grand des modules des valeurs propres de M. Pour toute matrice $A \in M_n(\mathbb{C})$, on a : $|||A||| = \rho(A^*A)^{1/2}$, où A^* désigne, rappelons-le, la matrice adjointe de A.

Démonstration. Cela résulte directement de la dernière section du module I.4. La matrice $M := A^*A$ étant hermitienne positive, ses valeurs propres sont réelles positives. De plus :

$$\rho(M) = \sup\big(\mathrm{sp}(M)\big) = \sup_{\|X\|=1} X^*MX = \sup_{\|X\|=1} X^*A^*AX$$
$$= \sup_{\|X\|=1} (\,AX \mid AX\,) = \sup_{\|X\|=1} \|AX\|^2 = |||A|||^2. \qquad \square$$

Revenons à l'exemple 3 ci-dessus. La matrice de L dans la base canonique de $\mathbb{R}^2$ est $A := \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, et $A^*A = {}^tAA = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 10 & 14 \\ 14 & 20 \end{pmatrix} = M$. La matrice A étant réelle, on a l'égalité suivante (vérification facile laissée au lecteur) :

$$\sup_{X\in\mathbb{R}^2,\,\|X\|=1} \|AX\| = \sup_{X\in\mathbb{C}^2,\,\|X\|=1} \|AX\|.$$

La proposition précédente donne donc à nouveau $|||A||| = \rho(M) = \sqrt{\lambda}$.

Proposition et définition 22. Soient E, F deux espaces vectoriels normés. L'ensemble des applications linéaires continues de E dans F, noté $\mathcal{L}_c(E,F)$, est un sous–espace vectoriel de $\mathcal{L}(E,F)$. De plus, $L \mapsto |||L|||$ est une norme sur $\mathcal{L}_c(E,F)$. En particulier, l'espace vectoriel normé $E_c^* := \mathcal{L}_c(E,K)$ est appelé *dual topologique* de E.

Démonstration. Soient $f, g \in \mathcal{L}_c(E,F)$ et $\alpha, \beta \in K$. Pour tout $x \in E$ de norme 1, on a $\|f(x)\| \leqslant |||f|||$ et $\|g(x)\| \leqslant |||g|||$, d'où $\|(\alpha f + \beta g)(x)\| \leqslant |\alpha|\,|||f||| + |\beta|\,|||g|||$. Ainsi $\alpha f + \beta g$ est continue, d'où la première assertion. En prenant ci-dessus $\alpha := 1$ et $\beta := 1$, on obtient : $|||f+g||| \leqslant |||f||| + |||g|||$. Les deux autres propriétés des normes sont évidentes. $\square$

Proposition 23. Soient E, F, G des e.v. n. et $f \in \mathcal{L}_c(E,F)$, $g \in \mathcal{L}_c(F,G)$. Alors $g \circ f \in \mathcal{L}_c(E,G)$ et $|||g \circ f||| \leqslant |||g|||\,|||f|||$, mais il n'y a pas toujours égalité.

Démonstration. Soit $x \in E$.

Appliquons deux fois l'inégalité (18). Puisque f est continue, $\|f(x)\| \leqslant |||f|||\|x\|$. Puisque g est continue, $\big\|g\big(f(x)\big)\big\| \leqslant |||g|||\|f(x)\| \leqslant |||g|||\,|||f|||\|x\|$. D'où la continuité de $g \circ f$ et l'inégalité annoncée. La dernière assertion est claire : déjà en dimension finie, on sait qu'il est possible d'avoir $f \neq 0$, $g \neq 0$ et $g \circ f = 0$. $\square$

Ce que nous venons de dire sur la continuité des applications linéaires se généralise aisément aux applications *multilinéaires*. Soient $E_1, \ldots, E_n$ et F des espaces vectoriels normés, et $E := E_1 \times E_2 \times \cdots \times E_n$ l'e.v. n. produit (exemple 5 de la page 395).

Théorème 24. Une application *multilinéaire* $f : E := E_1 \times E_2 \times \cdots \times E_n \to F$ est continue si, et seulement s'il existe un réel $C \geqslant 0$ tel que, pour tout $(x_1, \ldots, x_n) \in E$, l'inégalité suivante soit vérifiée : $\|f(x_1, x_2 \ldots, x_n)\| \leqslant C\|x_1\|\|x_2\| \cdots \|x_n\|$.

Démonstration. Rappelons que, si $x := (x_1, \ldots, x_n) \in E$, $\|x\|$ est le maximum des $\|x_i\|$. Supposons que f soit continue, en particulier continue en 0. Il existe $\alpha > 0$ tel que, pour $y \in E$, $\big(\|y\| \leqslant \alpha\big) \Longrightarrow \big(\|f(y)\| \leqslant 1\big)$. Posons $C := \alpha^{-n}$. Vérifions que l'inégalité de l'énoncé est vraie pour tout $x := (x_1, \ldots, x_n) \in E$. C'est évident si l'un des x_i est nul, car alors $f(x) = 0$. Remarquons sinon que, si l'on multiplie l'un des x_i par $\lambda \in K$, chaque membre de l'inégalité est multiplié par $|\lambda|$. Cela permet de supposer que $\|x_i\| = \alpha$ pour tout i. Mais alors le second membre de l'inégalité vaut 1, et $\|f(x)\| \leqslant 1$ puisque $\|x\| = \alpha$.
Supposons inversement que $C \geqslant 0$ vérifie la propriété de l'énoncé. Traitons le cas $n := 2$, laissant le cas général au lecteur. Soient $a := (a_1, a_2) \in E$ et $h := (h_1, h_2) \in E$. La bilinéarité de f donne :

$$f(a_1 + h_1, a_2 + h_2) = f(a_1, a_2) + f(a_1, h_2) + f(h_1, a_2) + f(h_1, h_2). \qquad (*)$$

D'où $f(a+h) - f(a) = f(a_1, h_2) + f(h_1, a_2) + f(h_1, h_2)$. On en déduit la majoration suivante : $\|f(a+h) - f(a)\| \leqslant C\big[\|a_1\|\|h_2\| + \|h_1\|\|a_2\| + \|h_1\|\|h_2\|\big] \leqslant C\big[2\|a\|\|h\| + \|h\|^2\big]$. Il en résulte aussitôt que f est continue en a. ∎

Comme dans le cas linéaire, il résulte de ce théorème que f est continue si, et seulement si, sa restriction à $S_1 \times S_2 \times \cdots \times S_n$ est bornée, chaque S_i étant la sphère unité de E_i.

Corollaire 25. La loi externe $(\lambda, x) \mapsto \lambda x$ d'un espace vectoriel normé E est une application continue de $K \times E$ dans E.

Démonstration. L'application $f : (\lambda, x) \mapsto \lambda x$ de $K \times E$ dans E est bilinéaire, par définition d'un espace vectoriel. Puisque $\|\lambda x\| = |\lambda|\,\|x\|$ pour tout $(\lambda, x) \in K \times E$, la propriété du théorème est vraie avec $C := 1$. ∎

Corollaire 26. Soient E, F, G trois espaces vectoriels normés. La composition : $(f, g) \mapsto g \circ f$ est une application continue de $\mathcal{L}_c(E, F) \times \mathcal{L}_c(F, G)$ dans $\mathcal{L}_c(E, G)$.

Démonstration. Cette application est bilinéaire. La conclusion vient donc du théorème précédent et de la proposition 23 de la page ci-contre (prendre $C := 1$). ∎

Différentiabilité d'une application multilinéaire

Soient $f : E := E_1 \times E_2 \times \cdots \times E_n \to F$ une application multilinéaire continue et $C \geqslant 0$ comme dans le théorème. Montrons que f est « différentiable », et que sa « différentielle » Df_a en un point $a := (a_1, \ldots, a_n) \in E$ est donnée par la formule :

$$\forall h := (h_1, \ldots, h_n) \in E, \quad Df_a(h) \;=\; \sum_{k=1}^{n} f(a_1, a_2, \ldots, h_k, \ldots, a_n). \qquad (20)$$

Le $k^{ème}$ terme de la somme est obtenu en remplaçant a_k par h_k dans $f(a_1, \ldots, a_n)$. Notons $L(h)$ le second membre de cette égalité. Alors $L : E \mapsto F$ est linéaire. Elle est

continue, à cause de la majoration $\|L(h)\| \leqslant nC\|a\|^{n-1}\|h\|$, obtenue facilement en se rappelant que $\|a_j\| \leqslant \|a\|$ pour tout j, et de même pour h.
Posons $\Delta(h) := f(a+h) - f(a) - L(h)$, et montrons que, si h tend vers 0, $\|\Delta(h)\|$ est négligeable devant $\|h\|$. Si $n = 1$, $\Delta(h) = 0$, supposons donc $n \geqslant 2$. Puisque f est multilinéaire, $f(a+h)$ est somme de 2^n termes t du type suivant : J étant une partie arbitraire de $[\![1, n]\!]$, on pose $t = f(x_1, \dots, x_n)$ où, pour $i = 1, \dots, n$, $x_i = h_i$ si $i \in J$ et $x_i = a_i$ si $i \notin J$. Pour un terme tel t, on a $\|t\| \leqslant C\|h\|^k\|a\|^{n-k}$, où $k = \operatorname{card}(J)$. Il en résulte que, si J est fixée et $k \geqslant 2$, $\|t\|$ est négligeable devant $\|h\|$ lorsque h tend vers 0. Or, $\Delta(h)$ est la somme des termes t pour toutes les parties J de $[\![1, n]\!]$ telles que $\operatorname{card}(J) \geqslant 2$, donc $\|\Delta(h)\|$ est négligeable devant $\|h\|$ lorsque h tend vers 0.
Nous avons prouvé que f était « différentiable » en a, et que sa « différentielle » était L.
Explication des guillemets : dans cet ouvrage (*cf.* le module II.5), comme dans le cours de [L1], il n'est question de fonctions différentiables qu'en dimension finie.

Dérivation d'un déterminant

Soient $n \in \mathbb{N}^*$, I un intervalle ouvert de $\mathbb{R}$ et, pour tout $(i,j) \in [\![1, n]\!]^2$, $a_{i,j} : I \to \mathbb{R}$ une fonction dérivable. Si $t \in I$, notons $A(t)$ la matrice de terme général $a_{i,j}(t)$, $C_j(t)$ les colonnes de $A(t)$ ($1 \leqslant j \leqslant n$), de sorte que $A(t) = [C_1(t); \dots; C_n(t)]$. Soit $D(t)$ le déterminant de $A(t)$, i.e. le déterminant de $\big(C_1(t), \dots, C_n(t)\big)$ dans la base canonique de $\mathbb{R}^n$ (qui sera omise). Montrons que la fonction $t \mapsto D(t)$ est dérivable sur I, sa dérivée étant donnée par la formule :

$$\begin{aligned} D'(t) = &\det\big(C_1'(t), C_2(t), \dots, C_n(t)\big) + \det\big(C_1(t), C_2'(t), \dots, C_n(t)\big) \\ &+ \dots + \det\big(C_1(t), \dots, C_{n-1}'(t), C_n(t)\big) + \det\big(C_1(t), C_2(t), \dots, C_n'(t)\big)\,. \end{aligned}$$

Utilisons librement le module IV.7 du cours de [L1]. La fonction D apparaît comme composée : $D = g \circ f$, où $f : I \to (\mathbb{R}^n)^n$ est donnée par $t \mapsto \big(C_1(t), \dots, C_n(t)\big)$ et $g : (\mathbb{R}^n)^n \to \mathbb{R}$ est la fonction déterminant $[C_1; \dots; C_n] \mapsto \det(C_1, \dots, C_n)$. La fonction vectorielle f est dérivable, sa dérivée étant $t \mapsto \big(C_1'(t), \dots, C_n'(t)\big)$. Quant à g, c'est une application n-linéaire, continue parce que l'on est en dimension finie (*cf.* la proposition 61 de la page 441). Ce qui précède montre que g est différentiable, sa différentielle en un point $C := (C_1, \dots, C_n) \in (\mathbb{R}^n)^n$ étant donnée par la formule suivante, où $H_1, \dots, H_n \in \mathbb{R}^n$ sont quelconques et $H := (H_1, \dots, H_n)$:

$$Dg_C(H) \;=\; \sum_{k=1}^{n} g(C_1, C_2, \dots, H_k, \dots, C_n) \;=\; \sum_{k=1}^{n} \det(C_1, C_2, \dots, H_k, \dots, C_n).$$

Par composition, $D = g \circ f$ est différentiable, *i.e.* dérivable, et sa dérivée vaut :

$$D'(t) \;=\; Dg_{C(t)}\big(f'(t)\big) \;=\; Dg_{C(t)}\big(C_1'(t), \dots, C_n'(t)\big)\,,$$

d'où la conclusion.

1.7 Normes équivalentes

Lorsqu'on considère deux normes $\|\cdot\|$ et $\|\cdot\|'$ sur un même espace vectoriel E, les propriétés topologiques des espaces vectoriels normés $(E,\|\cdot\|)$ et $(E,\|\cdot\|')$ sont en général différentes. Nous allons étudier un cas important où ces propriétés sont les mêmes.

Définition 20. Deux normes $\|\cdot\|$, $\|\cdot\|'$ sur un espace vectoriel E sont dites *équivalentes* s'il existe $C, C' > 0$ tels que $\|x\|' \leqslant C\|x\|$ et $\|x\| \leqslant C'\|x\|'$ pour tout $x \in E$.

C'est évidemment une relation d'équivalence entre normes sur E. Comme exemple, considérons sur K^n la norme usuelle $\|\cdot\|$ et les normes $\|\cdot\|_1$, $\|\cdot\|_\infty$ de l'exemple 4 de la page 395. Soit $x := (x_1,\dots,x_n) \in K^n$. On a donc :

$$\|x\| = \Big(|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2\Big)^{1/2},\ \|x\|_1 = \sum_{i=1}^{n} |x_i|,\ \|x\|_\infty = \max\big(|x_1|,\dots,|x_n|\big).$$

D'abord, $|x_i| \leqslant \|x\|$ et $|x_i| \leqslant \|x\|_\infty$ pour tout i, d'où $\|x\|_\infty \leqslant \|x\|$ et $\|x\| \leqslant \sqrt{n}\,\|x\|_\infty$. Les normes $\|\cdot\|$ et $\|\cdot\|_\infty$ sont donc équivalentes. Ensuite, il existe un i tel que $|x_i| = \|x\|_\infty$, donc $\|x\|_\infty \leqslant \|x\|_1$. Vu l'inégalité évidente $\|x\|_1 \leqslant n\|x\|_\infty$, les normes $\|\cdot\|_1$ et $\|\cdot\|_\infty$ sont équivalentes. En conclusion, les trois normes considérées sont équivalentes. Dans la section 4.1, nous montrerons d'ailleurs que toutes les normes sur un espace vectoriel de dimension finie sont équivalentes.

Soit $E := \mathcal{C}\big([a,b],K\big)$ l'espace des fonctions continues de $[a,b]$ dans K, $a < b$. Montrons que les deux normes $\|\cdot\|_\infty$ et $\|\cdot\|_2$ sur E ne sont pas équivalentes. Soit $f \in E$.

Pour tout $x \in [a,b]$, on a $|f(x)| \leqslant \|f\|_\infty$, d'où par intégration :

$$\|f\|_2^2 := \int_a^b \big|f(x)\big|^2\,\mathrm{d}x \leqslant \int_a^b \|f\|_\infty^2\,\mathrm{d}x = (b-a)\|f\|_\infty^2, \quad \text{d'où} \quad \|f\|_2 \leqslant \sqrt{b-a}\,\|f\|_\infty.$$

Par contre, il n'existe pas de réel $C \geqslant 0$ tel que $\|f\|_\infty \leqslant C\|f\|_2$ pour toute $f \in E$. Supposons au contraire qu'un tel C existe. Prenons pour f la fonction $x \mapsto (x-a)^n$, où $n \in \mathbb{N}$. Alors $\|f\|_\infty = (b-a)^n$ et $\|f\|_2^2 = \displaystyle\int_a^b (x-a)^{2n}\,\mathrm{d}x = (b-a)^{2n+1}/(2n+1)$.

Ainsi $(b-a)^{2n} \leqslant C^2(b-a)^{2n+1}/(2n+1)$, soit $2n+1 \leqslant C^2(b-a)$, ceci pour tout $n \in \mathbb{N}$. C'est bien sûr absurde. Les deux normes données ne sont donc pas équivalentes.

Voici une traduction utile de l'équivalence de deux normes :

Proposition 27. Deux normes $\|\cdot\|$, $\|\cdot\|'$ sur un espace vectoriel E sont équivalentes si, et seulement si, l'identité Id_E de E est un homéomorphisme de $\big(E,\|\cdot\|\big)$ sur $\big(E,\|\cdot\|'\big)$.

Démonstration. Notons f l'application $\mathrm{Id}_E : \big(E,\|\cdot\|\big) \to \big(E,\|\cdot\|'\big)$. Sa réciproque f^{-1} est $\mathrm{Id}_E : \big(E,\|\cdot\|'\big) \to \big(E,\|\cdot\|\big)$. Bien sûr, $f(x) = x = f^{-1}(x)$ pour tout $x \in E$. D'après le théorème 18 de la page 415, f est continue si, et seulement s'il existe $C > 0$ tel que $\|x\|' \leqslant C\|x\|$ pour tout x. De même f^{-1} est continue si, et seulement s'il existe $C' > 0$ tel que $\|x\| \leqslant C'\|x\|'$ pour tout x. D'où la conclusion. ∎

Conséquences

Soient $\|\cdot\|$, $\|\cdot\|'$ deux normes équivalentes sur un espace vectoriel E. Alors les espaces vectoriels normés $\big(E, \|\cdot\|\big)$ et $\big(E, \|\cdot\|'\big)$ ont les mêmes propriétés topologiques. C'est un bon exercice de vérifier par exemple les assertions suivantes :

- une suite (u_n) converge dans $\big(E, \|\cdot\|\big)$ si, et seulement si, elle converge dans $\big(E, \|\cdot\|'\big)$;
- une partie A de E est fermée (resp. ouverte, resp. dense) dans $\big(E, \|\cdot\|\big)$ si, et seulement si, elle possède la même propriété dans $\big(E, \|\cdot\|\big)'$;
- soit X un espace métrique. Une application $f : E \to X$ (ou $f : X \to E$) est continue lorsqu'on munit E de la norme $\|\cdot\|$ si, et seulement si, c'est vrai pour la norme $\|\cdot\|'$.

Une partie A de E est bornée pour $\|\cdot\|$ si, et seulement si, elle est bornée pour $\|\cdot\|'$. Supposons par exemple qu'il existe $M \geqslant 0$ tel que $\|x\| \leqslant M$ pour tout $x \in A$. Il existe $C > 0$ tel que $\|x\|' \leqslant C\|x\|$ pour tout $x \in E$, et alors $\|x\|' \leqslant CM$ pour tout $x \in A$. Cependant, « être borné » n'est pas une propriété topologique : ainsi, la fonction $\tanh$ est un homéomorphisme de $\mathbb{R}$, qui n'est pas borné, sur $]-1, 1[$, qui est borné.

2 Espaces métriques complets

Nous abordons l'étude des espaces complets, qui jouent un rôle central en analyse.

2.1 Suites de Cauchy — Espaces complets

Comment savoir si une suite réelle (u_n) converge ? Si l'on connaît sa future « limite » ℓ, il suffit d'appliquer la définition de la convergence de (u_n) vers ℓ. Si (u_n) est croissante, il suffit de voir si elle est majorée. Dans le cas général, le cours de [L1] n'offre pas d'outil systématique. La notion (fondamentale) de suite de Cauchy permet de combler cette lacune.

Définition 21. Dans un espace métrique X, une suite (u_n) est appelée *suite de Cauchy* (on dit aussi que la suite (u_n) *est de Cauchy*) si elle vérifie la condition suivante :

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \geqslant N, \forall p \geqslant N, \quad d(u_n, u_p) \leqslant \varepsilon. \tag{21}$$

Autrement dit, pour tout $\varepsilon > 0$, il doit exister un rang à partir duquel la distance de deux termes quelconques de la suite est au plus égale à ε. Donnons deux résultats très simples portant sur les suites de Cauchy dans un espace métrique X.

Proposition 28. Toute suite convergente est une suite de Cauchy.

Démonstration. Supposons que (u_n) converge vers $\ell \in X$, et soit $\varepsilon > 0$. Il existe $N \in \mathbb{N}$ tel que $n \geqslant N \implies d(u_n, \ell) \leqslant \varepsilon$. Si $n \geqslant N$ et $p \geqslant N$, on a donc $d(u_n, \ell) \leqslant \varepsilon$ et $d(\ell, u_p) \leqslant \varepsilon$. L'inégalité triangulaire donne : $d(u_n, u_p) \leqslant d(u_n, \ell) + d(\ell, u_p) \leqslant \varepsilon + \varepsilon = 2\varepsilon$. La suite (u_n) est donc une suite de Cauchy (remplacer ε par $\varepsilon/2$). ∎

Proposition 29. Toute suite de Cauchy (u_n) est bornée.

Démonstration. Il existe $N \in \mathbb{N}^*$ tel que les conditions $n \geqslant N$ et $p \geqslant N$ entraînent $d(u_n, u_p) \leqslant 1$. En particulier $n \geqslant N \Longrightarrow d(u_n, u_N) \leqslant 1$. Alors (u_n) est bornée car, pour tout $n \in \mathbb{N}$, on a $d(u_n, u_N) \leqslant \max\big(d(u_0, u_N), \dots, d(u_{N-1}, u_N), 1\big)$.

La réciproque est fausse : dans $\mathbb{R}$, la suite $\big((-1)^n\big)$ est bornée. Ce n'est pas une suite de Cauchy, car la distance entre deux termes consécutifs vaut toujours 2. Si l'on choisit $\varepsilon := 1$ (pourquoi en a-t-on le droit ?), la condition (21) n'est donc pas satisfaite.
La réciproque de la proposition 28 de la page ci-contre est également fausse. Prenons par exemple $X := \mathbb{Q}$, muni de la distance usuelle $d : (x, y) \mapsto |x - y|$. Soit a un nombre réel irrationnel (par exemple $a := \sqrt{2}$). Nous savons qu'il existe une suite (u_n) de $\mathbb{Q}$ convergeant vers a. D'après la proposition 28 de la page précédente, (u_n) est une suite de Cauchy de $\mathbb{R}$, donc aussi de l'espace métrique $\mathbb{Q}$. Si (u_n) avait dans $\mathbb{Q}$ une limite ℓ, ℓ serait aussi limite de (u_n) dans $\mathbb{R}$.
D'où $\ell = a$ (unicité de la limite dans $\mathbb{R}$), et donc $a \in \mathbb{Q}$, ce qui est faux.

Proposition 30.
Si une suite de Cauchy a une valeur d'adhérence, elle est convergente.

Démonstration. Soit ℓ une valeur d'adhérence de la suite de Cauchy (u_n). Soit $\varepsilon > 0$. Il existe $N \in \mathbb{N}$ tel que $d(u_n, u_p) \leqslant \varepsilon$ dès que $n \geqslant N$ et $p \geqslant N$. Puisque ℓ est valeur d'adhérence de (u_n), il existe $m \geqslant N$ tel que $d(u_m, \ell) \leqslant \varepsilon$. Pour tout $n \geqslant N$, on a : $d(u_n, \ell) \leqslant d(u_n, u_m) + d(u_m, \ell) \leqslant \varepsilon + \varepsilon = 2\varepsilon$, et donc (u_n) converge vers ℓ.

Dans certains espaces métriques (« complets » au sens de la définition ci-après), toute suite de Cauchy converge. Ce n'est donc pas le cas de $\mathbb{Q}$, qui n'est pas assez « gros » ; *cf.* à ce sujet le théorème 34 de la page 425 et l'exercice II.1.21.
Nous montrerons en revanche que $\mathbb{R}$ est « complet ».

Définition 22 (Espaces complets, espaces de Banach). Un espace métrique dans lequel toute suite de Cauchy converge est dit *complet*. Si un e.v. n. est complet pour la distance associée à sa norme, on dit que c'est un e.v. n. complet ou un *espace de Banach*.

En rapprochant cette définition de la proposition 28 de la page précédente, nous obtenons ceci :

Corollaire 31 (critère de Cauchy). Dans un espace métrique *complet* (en abrégé *espace complet*), une suite est convergente si, et seulement si, c'est une suite de Cauchy.

Ce corollaire et la proposition 28 de la page ci-contre fournissent les règles pratiques suivantes :

PROPRIÉTÉ. *Dans un espace métrique X, considérons une suite (u_n).*

1. *Pour démontrer que (u_n)* diverge*, il* suffit *de vérifier qu'elle n'est pas de Cauchy.*
2. *Supposons X* complet*. Pour que (u_n)* converge*, il* suffit *qu'elle soit de Cauchy.*

Voici deux exemples d'e.v. n. complets (le premier sera généralisé dans la section 4).

Théorème 32.
Pour tout $p \in \mathbb{N}^*$, $\mathbb{R}^p$ est complet. En particulier, $\mathbb{R}$ et $\mathbb{C}$ sont complets.

Démonstration. Soit $(\mathbf{u}_n)$ une suite de Cauchy de $\mathbb{R}^p$. D'après la proposition 29 de la page précédente, $(\mathbf{u}_n)$ est bornée. D'après le théorème de Bolzano-Weierstraß (*cf.* le cours de [L1]), on peut en extraire une sous-suite, convergeant vers un certain vecteur $\mathbf{u}$. Alors $\mathbf{u}$ est valeur d'adhérence de $(\mathbf{u}_n)$. La proposition 30 de la page précédente montre donc que $(\mathbf{u}_n)$ converge vers $\mathbf{u}$. Ainsi $\mathbb{R}^p$ est complet. En particulier, $\mathbb{R}$ est complet, et $\mathbb{C}$ aussi, car il s'identifie à $\mathbb{R}^2$ en tant qu'espace euclidien. ∎

Théorème 33. L'espace préhilbertien $\ell^2(K)$ est complet.

Démonstration. Soit (u_n) une suite de Cauchy de $\ell^2(K)$. Pour tout $n \in \mathbb{N}$, u_n est une suite de K : $u_n := (u_{k,n})_{k\geqslant 0}$, et la série $\sum_{k=0}^{+\infty} |u_{k,n}|^2$ converge. Soit $\varepsilon > 0$. Il existe un entier N tel que, pour tous $m, n \geqslant N$, on ait $\|u_m - u_n\|_2 \leqslant \varepsilon$ dès que $m, n \geqslant N$, soit $\sum_{k=0}^{+\infty} |u_{k,m} - u_{k,n}|^2 \leqslant \varepsilon^2$.
Fixons un entier k. Pour tous $m, n \geqslant N$, l'inégalité précédente donne : $|u_{k,m} - u_{k,n}| \leqslant \varepsilon$. Il en résulte que $(u_{k,n})_{n\geqslant 0}$ est une suite de Cauchy de K. Puisque K est complet, cette suite converge, notons v_k sa limite. Nous obtenons ainsi une suite $v := (v_k)_{k\geqslant 0}$ de K. Nous allons montrer que $v \in \ell^2(K)$ et que (u_n) converge vers v, ce qui prouvera que $\ell^2(K)$ est complet.
Soit $T \in \mathbb{N}$. Si $m, n \geqslant N$, l'inégalité ci-dessus entraîne $\sum_{k=0}^{T} |u_{k,m} - u_{k,n}|^2 \leqslant \varepsilon^2$. Dans cette inégalité, fixons l'entier $m \geqslant N$ et faisons tendre n vers $+\infty$. Il vient : $\sum_{k=0}^{T} |u_{k,m} - v_k|^2 \leqslant \varepsilon^2$. C'est vrai pour tout T, de sorte que la série de terme général $|u_{k,m} - v_k|^2$ converge. Autrement dit, $u_m - v \in \ell^2(K)$, d'où $v \in \ell^2(K)$. En outre $\sum_{k=0}^{+\infty} |u_{k,m} - v_k|^2 \leqslant \varepsilon^2$, c'est-à-dire $\|u_m - v\|_2 \leqslant \varepsilon$. Ainsi $\|u_m - v\|_2 \leqslant \varepsilon$ dès que $m \geqslant N$, ce qui prouve que (u_n) converge vers v. ∎

Comme application de ce théorème, montrons que $\ell^2(K)$ ne possède pas de base orthonormée, ainsi que nous l'avons signalé dans le module I.4.
Supposons qu'une telle base $(e_i)_{i\in I}$ existe. L'ensemble I étant infini, il existe une bijection s de $\mathbb{N}$ sur une partie J de I. Soit V le sous-espace vectoriel de E engendré par les $e_{s(n)}$, $n \in \mathbb{N}$. Pour tout $i \in I \setminus J$, $e_i^\perp$ est fermé (exemple 2 de la page 417), donc V est fermé, comme intersection des $e_i^\perp$, $i \in I \setminus J$. La série $\sum_{n\geqslant 0} 2^{-n} e_{s(n)}$ est absolument convergente, donc convergente car $\ell^2(K)$ est complet. Sa somme x appartient à V puisque V est fermé. Il existe donc $p \in \mathbb{N}$ et $\lambda_0, \ldots, \lambda_p \in K$ tels que $x = \sum_{n=0}^{p} \lambda_n e_{s(n)}$. On en déduit que $(\, e_{s(p+1)} \mid x\,) = 0$. C'est absurde, car la forme linéaire $y \mapsto (\, e_{s(p+1)} \mid y\,)$ est continue, donc $(\, e_{s(p+1)} \mid x\,) = \sum_{n=0}^{+\infty} \big(\, e_{s(p+1)} \;\big|\; 2^{-n} e_{s(n)}\,\big) = 2^{-(p+1)}$, (*cf.* la remarque de la page 429).

Sur un espace vectoriel E, considérons deux normes $\|\cdot\|$ et $\|\cdot\|'$. Il se peut que E soit complet pour $\|\cdot\|$ mais pas pour $\|\cdot\|'$. Ainsi, nous verrons dans le module suivant que, si $a < b$, l'espace $\mathcal{C}\big([a,b],K\big)$ est complet pour la norme $\|\cdot\|_\infty$ mais pas pour la norme $\|\cdot\|_2$.
Supposons par contre que les deux normes soient équivalentes. Alors E est complet pour l'une si, et seulement s'il est complet pour l'autre. En effet, il existe deux réels $A, B > 0$ tels que $A\|x\| \leqslant \|x\|' \leqslant B\|x\|$ pour tout $x \in E$. Soit (u_n) une suite de E. Supposons par exemple que (u_n) soit une suite de Cauchy pour $\|\cdot\|$. Soit $\varepsilon > 0$. Il existe $N \in \mathbb{N}$ tel que $\|u_n - u_p\| \leqslant \varepsilon$ dès que $n, p \geqslant N$. Alors $\|u_n - u_p\|' \leqslant \varepsilon B$ dès que $n, p \geqslant N$, donc (u_n) est de Cauchy pour $\|\cdot\|'$. Le même argument montre que si (u_n) converge dans $\big(E, \|\cdot\|\big)$ et si ℓ est sa limite, (u_n) converge aussi vers ℓ dans $\big(E, \|\cdot\|'\big)$.
Quelles sont les parties de $\mathbb{R}$ qui sont complètes (en tant qu'espaces métriques) ? Le théorème suivant répond à cette question, de façon plus générale :

Théorème 34. Soient X un espace métrique et A un sous-espace (partie) de X.

1. Si A est un espace complet, c'est un fermé de X.
2. Réciproquement, *si X est complet* et si A est fermé, A est un espace complet.

Démonstration. Supposons A complet, et soit $x \in \overline{A}$. Il existe une suite de A convergeant vers x. Cette suite est de Cauchy, dans X donc dans A. Puisque A est complet, elle converge *dans* A. Sa limite dans A est égale à x, par unicité de la limite dans X, donc $x \in A$.

Supposons que X soit complet et que A soit fermé dans X. Soit (a_n) une suite de Cauchy de A. C'est aussi une suite de Cauchy de X. Puisque X est complet, (a_n) converge dans X, notons ℓ sa limite. Alors ℓ est limite d'une suite de A, donc $\ell \in A$ puisque A est fermé. Finalement, (a_n) converge vers ℓ *dans* A. L'espace A est donc complet. ∎

Ainsi, $[0,1]$ et $[0,+\infty[$ sont complets, mais $]0,1]$, $]0,1[$ et $]0,+\infty[$ ne le sont pas. Comme l'exponentielle est un homéomorphisme de $\mathbb{R}$, qui est complet, sur $]0,+\infty[$, qui ne l'est pas, on voit que *la complétude n'est pas une propriété topologique*.

Proposition 35. Tout produit fini d'espaces complets est complet.

Démonstration. Par récurrence sur le nombre d'espaces, on se ramène au cas d'un produit $X \times Y$ de deux espaces complets. Soit $\big((x_n, y_n)\big)$ une suite de Cauchy de $X \times Y$. Soit $\varepsilon > 0$. Il existe $N \in \mathbb{N}$ tel que les conditions $n \geqslant N$ et $p \geqslant N$ entraînent $\max\big(d(x_n,x_p), d(y_n,y_p)\big) = d\big((x_n,y_n),(x_p,y_p)\big) \leqslant \varepsilon$, c'est-à-dire $d(x_n,x_p) \leqslant \varepsilon$ *et* $d(y_n,y_p) \leqslant \varepsilon$. Les suites (x_n) et (y_n) sont donc de Cauchy. Puisque X et Y sont complets, ces suites convergent, notons x et y leurs limites respectives. Alors la suite $\big((x_n,y_n)\big)$ converge vers (x,y) (*cf.* l'exemple 1 de la page 401).

L'espace $X \times Y$ est donc complet. ∎

Voici un autre résultat important sur les espaces complets :

Théorème 36. Soient X un espace métrique complet et (F_n) une suite de parties fermées non vides de X, décroissante pour l'inclusion et telle que $\lim\limits_{n\to+\infty} \mathrm{Diam}(F_n) = 0$. Alors l'intersection F de tous les F_n est un singleton.

Démonstration. Comme intersection de fermés, F est fermée. Soient ensuite $x, y \in F$. Pour tout n, x et y appartiennent à F_n, donc $d(x,y) \leqslant \mathrm{Diam}(F_n)$, par définition du diamètre. D'où $d(x,y) \leqslant \lim\limits_{n\to+\infty} \mathrm{Diam}(F_n) = 0$, c'est-à-dire $x = y$.

Il reste à montrer que F n'est pas vide, *i.e.* nous devons *construire* un point de F. Les suites de Cauchy sont l'outil voulu. Pour tout n, soit a_n un point de F_n (par hypothèse, $F_n \neq \varnothing$). Soit $\varepsilon > 0$. À partir d'un certain rang N, on a $\mathrm{Diam}(F_n) \leqslant \varepsilon$. Si $N \leqslant n < p$, on a $a_n \in F_n$ et $a_p \in F_p \subset F_n$, donc $d(a_n, a_p) \leqslant \mathrm{Diam}(F_n) \leqslant \varepsilon$. La suite (a_n) est donc de Cauchy. Puisque X est complet, cette suite a une limite a. Si $n \in \mathbb{N}$, a est la limite de $(a_k)_{k\geqslant n}$, qui est une suite de F_n, donc $a \in F_n$ car F_n est fermé. D'où $a \in F$. ∎

Dans $\mathbb{R}$ (qui est complet), le diamètre d'un intervalle borné est sa longueur. Ainsi :

Corollaire 37 (segments emboîtés). Soit (I_n) une suite de segments emboîtés ($I_n \supset I_{n+1}$ pour tout n) non vides de $\mathbb{R}$. Si la longueur de I_n tend vers 0 quand n tend vers $+\infty$, l'intersection des I_n est réduite à un point.

En exercice, le lecteur pourra aussi démontrer ce corollaire à l'aide de suites adjacentes.

Il y a bien sûr des e.v. n. qui ne sont pas complets. Soit par exemple V le sous-espace de $\ell^2(\mathbb{R})$ formé des suites à support fini. Pour tout $n \in \mathbb{N}$, notons $e_n \in \ell^2(\mathbb{R})$ la suite $(0, 0, \ldots, 1, 0, \ldots)$, le 1 étant à la position n. Anticipons ici sur la section 2.3. La série $\sum\limits_{n\geqslant 0} (n+1)^{-1} e_n$ converge et sa somme x est la suite de terme général $(n+1)^{-1}e_n$. Cette série vérifie donc le critère de Cauchy, dans $\ell^2(\mathbb{R})$ comme dans V. Mais $x \notin V$, donc V n'est pas complet. De même, dans le module suivant, nous montrerons que, si $a < b$, l'espace $\mathcal{C}\big([a,b],\mathbb{R}\big)$ muni de la norme $\|\cdot\|_\infty$ est complet. Nous montrerons aussi que le sous-espace vectoriel $\mathcal{P}$ de $\mathcal{C}\big([a,b],\mathbb{R}\big)$ formé des fonctions polynomiales est *dense* dans $\mathcal{C}\big([a,b],\mathbb{R}\big)$. Comme $\mathcal{P} \neq \mathcal{C}\big([a,b],\mathbb{R}\big)$, $\mathcal{P}$ n'est pas fermé dans $\mathcal{C}\big([a,b],\mathbb{R}\big)$. Le théorème 34 de la page précédente montre alors que $\mathcal{P}$ n'est pas complet. L'espace $\mathcal{C}\big([-1,1],\mathbb{R}\big)$, muni de la norme $\|\cdot\|_2$ associée au produit scalaire $(\,f \mid g\,) = \displaystyle\int_{-1}^{1} f(t)g(t)\,\mathrm{d}t$ n'est pas complet (*cf.* l'exercice II.1.42).

Lorsque l'espace d'arrivée d'une fonction est complet, l'existence d'une limite de cette fonction en un point (*cf.* la définition 14 de la page 408) peut être prouvée sans connaître la limite *a priori*, comme pour les suites (*cf.* le critère de Cauchy 31 de la page 423). Plus précisément, on dispose du résultat ci-dessous :

Proposition 38. Soient (X, d) un espace métrique, A une partie de X et a un point de X *adhérent à* A ($a \in \overline{A}$). Par ailleurs, soit (Y, δ) un espace métrique complet. Pour qu'une application $f : A \to Y$ possède une limite en a, il faut et il suffit que la condition suivante soit vérifiée. Pour tout $\varepsilon > 0$, il existe un voisinage U de a dans X tel que $\delta\big(f(x), f(x')\big) \leqslant \varepsilon$ pour tous x, $x' \in A \cap U$.

Démonstration. La nécessité est simple. Supposons en effet que f possède en a une limite ℓ. Soit $\varepsilon > 0$; d'après la définition précitée, il existe $\alpha > 0$ tel que, pour tout $x \in A$, l'inégalité $d(x, a) \leqslant \alpha$ implique $\delta\big(f(x), \ell\big) \leqslant \varepsilon/2$. Soit U la boule ouverte $B(a, \alpha)$ dans X. C'est un voisinage de a et, pour tous x, $x' \in A \cap U$, on a $\delta\big(f(x), \ell\big) \leqslant \varepsilon/2$ et $\delta\big(f(x'), \ell\big) \leqslant \varepsilon/2$, donc $\delta\big(f(x), f(x')\big) \leqslant \varepsilon$ par inégalité triangulaire.

Supposons inversement que la condition de l'énoncé soit satisfaite. Soit (a_n) une suite de A convergeant vers a. L'hypothèse montre évidemment que la suite $\big(f(a_n)\big)$ est de Cauchy, elle converge donc puisque Y est complet. Pour conclure, en vertu du théorème 10 de la page 409, il suffit de montrer que la limite ℓ de $\big(f(a_n)\big)$ ne dépend pas de la suite (a_n). Soient donc (a_n) une autre suite de A convergeant vers a et ℓ' la limite de la suite $\big(f(a'_n)\big)$. Soient $\varepsilon > 0$ et U un voisinage de a comme dans l'énoncé. Pour tout entier n assez grand, on a d'une part $\delta\big(f(a_n), \ell\big) \leqslant \varepsilon$ et $\delta\big(f(a'_n), \ell'\big) \leqslant \varepsilon$, d'autre part a_n, $a'_n \in U$, d'où $\delta\big(f(a_n), f(a'_n)\big) \leqslant \varepsilon$. Par inégalité triangulaire, $\delta(\ell, \ell') \leqslant 3\varepsilon$. C'est vrai pour tout $\varepsilon > 0$, donc $\ell = \ell'$. ∎

2.2 Théorème du point fixe

Le théorème du point fixe est un résultat essentiel en analyse. Dans ce volume, nous en verrons des applications en analyse numérique (résolution d'équations), en calcul différentiel (théorème des fonctions implicites), et dans la théorie des équations différentielles.

Rappelons que, si f est une application d'un ensemble X dans lui-même, on appelle *point fixe* de f tout point $x \in X$ tel que $f(x) = x$. Ainsi, trouver d'éventuels points fixes de f, c'est résoudre l'équation $f(x) = x$.

Théorème 39 (Théorème du point fixe). Soit X un espace métrique complet (non vide). Toute application *contractante* f de X dans lui-même possède un unique point fixe a. De plus, pour tout point $b \in X$, on peut obtenir a comme limite de la suite (u_n) définie par son premier terme $u_0 := b$ et la relation de récurrence $u_{n+1} := f(u_n)$.

Démonstration. Elle est assez simple, si l'on se rappelle que toute suite de Cauchy de X converge, et si l'on sait calculer la somme des termes d'une progression géométrique. Soit donc $f : X \to X$ une application contractante. Par définition, il existe un nombre $k \in [0, 1[$, tel que l'inégalité suivante soit vraie pour tous $x, y \in X$:

$$d\big(f(x), f(y)\big) \leqslant k d(x, y). \tag{22}$$

Prouvons d'abord l'existence d'un point fixe. Partons d'un point $b \in X$, et définissons la suite (u_n) comme dans l'énoncé. Puisque $u_2 := f(u_1)$ et $u_1 := f(u_0)$, l'inégalité (22)

entraîne $d(u_1,u_2) \leqslant kd(u_0,u_1)$. De même $d(u_2,u_3) \leqslant kd(u_1,u_2) \leqslant k^2 d(u_0,u_1)$. Par récurrence, on en déduit $d(u_n,u_{n+1}) \leqslant k^n d(u_0,u_1)$ pour tout $n \in \mathbb{N}$. Considérons maintenant deux entiers n,p tels que $n<p$. L'inégalité triangulaire donne :

$$\begin{aligned} d(u_n,u_p) &\leqslant d(u_n,u_{n+1}) + d(u_{n+1},u_{n+2}) + \cdots + d(u_{p-1},u_p) \\ &\leqslant d(u_0,u_1)\big[k^n + k^{n+1} + \cdots + k^{p-1}\big] \\ &= \frac{k^n(1-k^{p-n})d(u_0,u_1)}{1-k} \leqslant \frac{k^n d(u_0,u_1)}{1-k}\cdot \end{aligned}$$

En posant $C := d(u_0,u_1)/(1-k)$, nous obtenons l'inégalité suivante pour $n \leqslant p$:

$$d(u_n,u_p) \leqslant Ck^n. \tag{23}$$

Il en résulte que (u_n) est une suite de Cauchy. Soit en effet $\varepsilon > 0$. Puisque $k<1$, il existe un $N \in \mathbb{N}$ tel que $Ck^N \leqslant \varepsilon$. Considérons alors deux entiers $n,p \geqslant N$. Si par exemple $n \leqslant p$, l'inégalité (23) donne : $d(u_n,u_p) \leqslant Ck^n \leqslant Ck^N \leqslant \varepsilon$. Puisque X est complet, la suite (u_n) converge, soit a sa limite.

L'application f est continue, car k-lipschitzienne. Puisque (u_n) converge vers a, la suite $\big(f(u_n)\big) = (u_{n+1})$ converge vers $f(a)$, d'où $f(a) = a$ (unicité de la limite). Ainsi a est un point fixe de f.

L'assertion d'unicité est évidente. Soient en effet a, a' deux points fixes de f. Puisque $f(a) = a$ et $f(a') = a'$, l'inégalité $d\big(f(a),f(a')\big) \leqslant kd(a,a')$ s'écrit $d(a,a') \leqslant kd(a,a')$, soit $(1-k)d(a,a') \leqslant 0$. Comme $k<1$, on a $d(a,a') \leqslant 0$, d'où $a' = a$. ∎

Il convient de noter que cette preuve nous renseigne sur la vitesse de convergence de la suite (u_n). En effet, dans l'inégalité (23), fixons l'indice n, et faisons tendre p vers $+\infty$. Puisque u_p tend vers a, on obtient ceci (par passage à la limite dans des inégalités larges) :

$$d(u_n,a) \leqslant Ck^n, \quad \text{où} \quad C = \frac{d(u_0,u_1)}{1-k}\cdot \tag{24}$$

Voici une illustration très simple.

Exercice 8.

Calculer $\sqrt{5}$ à 10^{-6} près, à l'aide de la fonction $f : x \mapsto \dfrac{1}{2}\left(x + \dfrac{5}{x}\right)$.

Solution. Le domaine de définition de f est $\mathbb{R}^*$.

L'équation $f(x) = x$ s'écrit $x + 5/x = 2x$, ce qui équivaut à $x = 5/x$, soit $x^2 = 5$. Il y a deux points fixes : $\sqrt{5}$ et $-\sqrt{5}$.

Pour appliquer le théorème, il faut d'abord choisir un intervalle X complet, c'est-à-dire fermé. Comme $2^2 < 5 < 2,5^2 = 6,25$, prenons $X := [2,5/2]$. Étudions f'. Sur l'intervalle X, $f'(x) = \big[1 - 5/x^2\big]/2$ croît de $-1/8$ à $1/10$ et ne s'annule qu'en $\sqrt{5}$. Ainsi f décroît de $9/4$ à $\sqrt{5}$, puis croît de $\sqrt{5}$ à $9/4$. Cela montre que X est stable par f : $f(X) \subset X$, c'était nécessaire pour appliquer le théorème. Notons encore f l'application $x \mapsto f(x)$ de X dans X, son seul point fixe est $\sqrt{5}$.

Montrons que f est contractante. Sur X, $|f'|$ est majorée par $k := 1/8$, donc f est k-lipschitzienne (exemple 1 de la page 410). Choisissons $b := 2$. Alors $u_0 := 2$, $u_1 := 2,25$, d'où $C := 2/7$. Pour obtenir une précision de 10^{-6}, il suffit de prendre $n := 7$, car $Ck^7 < 2.10^{-7}$. L'inégalité (24) donne $\left|u_7 - \sqrt{5}\right| < 2.10^{-7}$, donc $u_7 = 2,236067\ldots$ est une valeur approchée de $\sqrt{5}$ à 10^{-6} près. En réalité, on constate que $u_3 = 2,236067977\ldots$ est déjà une valeur approchée de $\sqrt{5}$ à 10^{-9} près, nous laissons le lecteur le *démontrer* à titre d'exercice.

Dans le théorème du point fixe, les hypothèses sont bien nécessaires. Ainsi l'application $x \mapsto x/2$ de $X := \left]0, 1\right]$ dans lui-même est contractante (de rapport de contraction $1/2$), mais elle n'a pas de point fixe. Ici, l'espace X n'est pas complet, car il n'est pas fermé dans $\mathbb{R}$ (en fait, le seul point fixe de $x \mapsto x/2$ dans $\mathbb{R}$ est $0 \notin X$). Autre exemple : $\mathbb{R}$ est complet, mais $f : x \mapsto \sqrt{x^2+1}$ n'a pas de point fixe. On a $|f'(x)| < 1$ pour tout $x \in \mathbb{R}$, donc $|f(x) - f(y)| < |x - y|$ dès que $x \neq y$. Cette application n'est cependant pas contractante, bien qu'elle soit 1-lipschitzienne, donc le théorème ne s'applique pas.

2.3 Séries dans un espace vectoriel normé

La convergence des séries numériques a été étudiée dans le cours de [L1]. Lorsqu'on remplace les nombres par des fonctions, on obtient des « séries de fonctions », dont l'étude fait l'objet du module suivant et sera utilisée dans le module II.4. Dans cette section, nous considérons plus généralement des séries à valeurs dans un e.v. n. et donnons les définitions et résultats indispensables pour la suite.

Définition 23. Soit (u_n) une suite de l'e.v. n. E. Pour tout $n \in \mathbb{N}$, posons $S_n := \sum_{k=0}^{n} u_k$. On dit que la série $\sum_{n\geqslant 0} u_n$ *converge* ou *est convergente* si la suite (S_n) de ses sommes partielles est convergente (sinon la série est *divergente*). Dans ce cas, la limite S de la suite (S_n) est appelée *somme* de la série $\sum_{n\geqslant 0} u_n$, et l'on écrit : $S = \sum_{n=0}^{+\infty} u_n$.

Si une série $\sum_{n\geqslant 0} u_n$ converge, son terme général u_n tend vers 0. En effet, S_n et S_{n-1} tendent vers S quand n tend vers $+\infty$, donc $u_n = S_n - S_{n-1}$ tend vers $S - S = 0$. La réciproque est fausse, nous l'avons vu dans $\mathbb{R}$.

Remarque. Supposons que $\sum_{n\geqslant 0} u_n$ converge et soit S sa somme. Si les u_n appartiennent tous à un sous-espace vectoriel *fermé* V de E, alors $S \in V$ puisque chaque S_n appartient à V (*cf.* le théorème 6 de la page 404).

Soient F un e.v. n. et $f \in \mathcal{L}_c(E, F)$. Pour tout n, on a $\sum_{k=0}^{n} f(u_k) = f(S_n)$. On en déduit que la série $\sum_{n\geqslant 0} f(u_n)$ converge et a pour somme $f(S)$.

Ainsi $f\left(\sum_{n=0}^{+\infty} u_n\right) = \sum_{n=0}^{+\infty} f(u_n)$.

Dans $\mathbb{R}^p$, l'étude d'une série vectorielle $\sum_{n\geqslant 0} \mathbf{u}_n$ se ramène immédiatement à l'étude de p séries numériques. En effet, si $\mathbf{u}_n := (u_{1,n}, \dots, u_{p,n})$ pour tout n, il est clair que la série $\sum_{n\geqslant 0} \mathbf{u}_n$ converge si, et seulement si, chacune des séries numériques $\sum_{n\geqslant 0} u_{k,n}$ converge $(k = 1, \dots, p)$. Dans ce cas, on a de plus :

$$\sum_{n=0}^{+\infty} \mathbf{u}_n \;=\; \left(\sum_{n=0}^{+\infty} u_{1,n}, \sum_{n=0}^{+\infty} u_{2,n}, \dots, \sum_{n=0}^{+\infty} u_{p,n} \right).$$

Considérons en général une série $\sum_{n\geqslant 0} u_n$ dans un e.v. n. E. Pour des entiers n, p tels que $1 \leqslant n < p$, on a : $S_p - S_{n-1} = \sum_{k=n}^{p} u_k$.

En appliquant le corollaire 31 de la page 423 (critère de Cauchy) à la suite (S_n), on obtient pour les séries un critère de convergence essentiel :

Théorème 40 (critère de Cauchy). Dans un e.v. n. *complet* E, une série $\sum_{n\geqslant 0} u_n$ converge si, et seulement si, pour tout $\varepsilon > 0$, il existe un $N \in \mathbb{N}$ possédant la propriété suivante : pour tous entiers n, p tels que $N \leqslant n < p$, on a $\left\| \sum_{k=n}^{p} u_k \right\| \leqslant \varepsilon$.

Déjà dans $\mathbb{R}$ (qui est complet), le critère de Cauchy est un outil puissant. Reprenons deux exemples traités dans le cours de [L1]. Pour tout entier $n \geqslant 1$, on a $\sum_{k=n+1}^{2n} 1/k \geqslant 1/2$, car chacun des n termes de la somme est minoré par $1/(2n)$. La série harmonique $\sum_{n\geqslant 1} 1/n$ est donc divergente, car elle ne vérifie pas le critère de Cauchy (prendre $\varepsilon := 1/3$).
En sens inverse, la série $\sum_{n\geqslant 1} 1/n^2$ converge. En effet, si $2 \leqslant n < p$, on a :

$$\sum_{k=n+1}^{p} \frac{1}{k^2} \leqslant \sum_{k=n+1}^{p} \frac{1}{k(k-1)} = \sum_{k=n+1}^{p} \left[\frac{1}{k-1} - \frac{1}{k} \right] = \frac{1}{n} - \frac{1}{p} \leqslant \frac{1}{n}.$$

Si $\varepsilon > 0$ est donné, soit $N \in \mathbb{N}^*$ tel que $N > 1/\varepsilon$. Alors, $\sum_{k=n+1}^{p} 1/k^2 \leqslant 1/n \leqslant 1/N \leqslant \varepsilon$ dès que $p > n \geqslant N$. Le critère de Cauchy est donc vérifié, d'où la conclusion.

Exercice 9.

Soient E et F deux e.v. n.. On munit $\mathcal{L}_c(E, F)$ de la norme $|||\cdot|||$. Montrer que si F est complet, $\mathcal{L}_c(E, F)$ l'est aussi.

Solution.

Soit (f_n) une suite de Cauchy de $\mathcal{L}_c(E, F)$. Il s'agit de prouver que cette suite converge. Soit $\varepsilon > 0$. Il existe $N \in \mathbb{N}$ tel que $|||f_n - f_p||| \leqslant \varepsilon$ dès que $n \geqslant N$ et $p \geqslant N$. Si $x \in E$ est fixé, $\big(f_k(x)\big)$ est une suite de Cauchy de F. En effet, dès que $n \geqslant N$ et $p \geqslant N$, on a $\|f_n(x) - f_p(x)\| \leqslant |||f_n - f_p|||\,\|x\| \leqslant \varepsilon\|x\|$. Puisque F est complet, cette suite converge. La

formule $f(x) := \lim_{n\to+\infty} f_n(x)$ définit donc une application $f : E \to F$, linéaire en vertu de la proposition 2 de la page 401. Dans les inégalités ci-dessus, fixons x et un entier $n \geqslant N$. En faisant tendre p vers $+\infty$, on obtient $\|f_n(x) - f(x)\| \leqslant \varepsilon\|x\|$. Cela vaut pour tout x, donc $f_n - f$ est continue, et par suite f est continue. Enfin, la dernière inégalité étant vraie pour tout x, $|||f_n - f||| \leqslant \varepsilon$ dès que $n \geqslant N$, ce qui montre que f est limite de la suite (f_n) de $\mathcal{L}_c(E, F)$. D'où la conclusion.

Dans le cours de [L1], nous avons vu que si une série à terme complexes est absolument convergente, elle converge. Le théorème ci-après est une généralisation très importante de ce résultat. Sa preuve, très simple, fait encore apparaître la puissance du critère de Cauchy.

Définition 24. Dans un e.v. n. E, une série $\sum_{n\geqslant 0} u_n$ est dite *absolument convergente* si la série $\sum_{n\geqslant 0} \|u_n\|$ est convergente.

Dans $\mathbb{C}$, nous retrouvons la notion convergence absolue définie en [L1]. La somme de deux séries absolument convergentes est absolument convergente (le vérifier).

Théorème 41. Dans un e.v. n. complet, toute série absolument convergente est convergente. La réciproque est fausse.

Démonstration. Soit $\sum_{n\geqslant 0} u_n$ une série absolument convergente dans un e.v. n. complet E. Puisque la série numérique $\sum_{n\geqslant 0} \|u_n\|$ converge, elle vérifie le critère de Cauchy. Soit alors $\varepsilon > 0$. Il existe un $N \in \mathbb{N}$ tel que $\sum_{k=n}^{p} \|u_k\| \leqslant \varepsilon$ dès que les entiers n, p vérifient $N \leqslant n < p$. Pour de tels entiers, l'inégalité triangulaire (généralisée) donne :

$$\left\| \sum_{k=n}^{p} u_k \right\| \leqslant \sum_{k=n}^{p} \|u_k\| \leqslant \varepsilon.$$

La série $\sum_{n\geqslant 0} u_n$ vérifiant ainsi le critère de Cauchy, elle converge, car E est complet.

Nous savons que la réciproque est déjà fausse dans $\mathbb{R}$. Ainsi, la série harmonique alternée $\sum_{n\geqslant 1} (-1)^n/n$ est convergente, mais pas absolument convergente. ∎

À propos de convergence absolue, voici un résultat utile, qui sera généralisé dans [MPA]. Pour la preuve, nous appliquerons le théorème de Fubini (théorème 15 de la page 485).

Théorème 42. Soit $\sum_{n\geqslant 0} u_n$ une série *absolument convergente* à termes complexes.

Si s est une permutation de $\mathbb{N}$, la série $\sum_{n\geqslant 0} u_{s(n)}$ converge absolument, et les deux séries ont la même somme.

Démonstration. Si $(n, p) \in \mathbb{N}^2$, posons $a_{n,p} := u_n$ si $n = s(p)$ et $a_{n,p} := 0$ si $n \neq s(p)$. Si $n \in \mathbb{N}$ est fixé, la série $\sum_{p\geqslant 0} a_{n,p}$ converge absolument, car son seul terme éventuellement

non nul est $a_{n,s^{-1}(n)} = u_n$. Ainsi $s_n := \sum_{p=0}^{+\infty} |a_{n,p}| = |u_n|$ et $\sum_{p=0}^{+\infty} a_{n,p} = u_n$. Puisque $\sum_{n\geqslant 0} s_n = \sum_{n\geqslant 0} |u_n|$ converge, par hypothèse, le théorème de Fubini peut être appliqué.

Si $p \in \mathbb{N}$ est fixé, $a_{s(p),p} = u_{s(p)}$ est le seul terme éventuellement non nul de la série $\sum_{n\geqslant 0} a_{n,p}$. Ainsi $t_p := \sum_{n=0}^{+\infty} |a_{n,p}| = \left|u_{s(p)}\right|$ et $\sum_{n=0}^{+\infty} a_{n,p} = u_{s(p)}$. Le théorème de Fubini nous apprend d'abord que la série $\sum_{p\geqslant 0} t_p = \sum_{p\geqslant 0} \left|u_{s(p)}\right|$ converge. Ce théorème donne de plus l'égalité $\sum_{n=0}^{+\infty} \left(\sum_{p=0}^{+\infty} a_{n,p} \right) = \sum_{p=0}^{+\infty} \left(\sum_{n=0}^{+\infty} a_{n,p} \right)$, *i.e.* $\sum_{n=0}^{+\infty} u_n = \sum_{p=0}^{+\infty} u_{s(p)}$, comme désiré. ∎

Si les u_n sont réels positifs, la somme (éventuellement infinie) de la série $\sum_{n\geqslant 0} u_{s(n)}$ est indépendante de la permutation s : il suffit d'appliquer le théorème à chacune des séries $\sum_{n\geqslant 0} u_n$ et $\sum_{n\geqslant 0} u_{s(n)}$ (nous laissons les détails au lecteur).

3 Espaces métriques compacts

La notion de compacité est essentielle dans la plupart des branches des mathématiques et de leurs applications. Nous en poursuivons l'étude, commencée dans le cours de [L1].

3.1 Définition et premières propriétés

Dans le module IV.7 du cours de [L1], nous avons vu que, pour une partie K de $\mathbb{R}^p$, les deux propriétés suivantes sont équivalentes :

(i) la partie K est compacte, c'est-à-dire (par définition) fermée et bornée,

(ii) de toute suite de K on peut extraire une sous-suite convergeant *vers un point de* K.

Ainsi, un intervalle I de $\mathbb{R}$ est compact si, et seulement si, c'est un segment : $I := [a, b]$, où $a, b \in \mathbb{R}$. Pour une partie d'un e.v. n. E, les propriétés (i) et (ii) ne sont plus toujours équivalentes (sauf en dimension finie, nous le verrons). Pour généraliser la notion de compacité, il s'avère qu'il est préférable de partir de la propriété (ii) :

Définition 25. Un espace métrique X est dit *compact* si de toute suite de X on peut extraire une suite convergeant dans X (propriété dite de *Bolzano-Weierstraß*). Cela revient à dire que toute suite de X possède au moins une valeur d'adhérence.

Contrairement à la complétude (*cf.* page 425), *la compacité est une propriété topologique*. En effet, soit $f : X \to Y$ un homéomorphisme et supposons X compact. Soit (y_n) une suite de Y. La suite $\left(f^{-1}(y_n)\right)$ a une sous-suite convergente $\left(f^{-1}(y_{n_k})\right)$, dont l'image (y_{n_k}) par f converge puisque f est continue. Cela prouve que Y est compact lui aussi.

Avec cette définition, les parties compactes de $\mathbb{R}^n$ sont donc les mêmes que dans le cours de [L1] : ce sont les parties fermées et bornées. Nous verrons que c'est encore vrai pour les parties d'un e.v. n. de dimension finie (théorème 58 de la page 440). Voici un exemple montrant que cela peut être faux dans un e.v. n. de dimension infinie.

Exemple. Considérons l'e.v. n. $\ell^\infty(\mathbb{R})$ (exemple 6 de la page 395). Pour tout $k \in \mathbb{N}$, notons δ_k la suite dont tous les termes sont nuls sauf celui d'indice k, qui vaut 1. On a clairement $\|\delta_k\|_\infty = 1$ et $\|\delta_h - \delta_k\|_\infty = 1$ dès que $h, k \in \mathbb{N}$ sont distincts. On en déduit qu'aucune sous-suite de la suite (δ_k) ne peut converger, car une telle suite n'est pas de Cauchy. La sphère unité de $\ell^\infty(\mathbb{R})$ est fermée et bornée, mais elle ne vérifie pas la propriété (ii).

Donnons quelques propriétés simples des espaces compacts.

Proposition 43. Tout espace métrique compact est borné.

Démonstration. Montrons qu'un espace métrique X non borné n'est pas compact. Soient $a, b \in X$, $a \neq b$. Par hypothèse, aucune boule de centre a n'est égale à X. Nous pouvons donc définir par récurrence une suite (a_n) de X en posant $a_0 := b$ et en imposant à chaque pas la condition $d(a, a_{n+1}) > 2d(a, a_n)$. La suite $\big(d(a, a_n)\big)$ a pour limite $+\infty$. Si (a_{n_k}) est une sous-suite de (a_n), la suite (a_{n_k}) n'est pas bornée, car la suite $\big(d(a, a_{n_k})\big)$ a pour limite $+\infty$. Il n'existe donc aucune sous-suite convergente de (a_n), ce qui montre que X n'est pas compact. ∎

Ainsi $\mathbb{R}$ n'est pas compact, donc n'est pas homéomorphe à un cercle de $\mathbb{R}^2$. Un e.v. n. non nul E n'est jamais compact, car il n'est pas borné. Soit en effet $a \in E$, $a \neq 0$. Si $n \in \mathbb{N}$ tend vers $+\infty$, $\|na\| = n\|a\|$ tend vers $+\infty$, d'où notre assertion.

Théorème 44. Tout espace métrique compact est complet (la réciproque est fausse).

Démonstration. Il s'agit de montrer que toute suite de Cauchy (u_n) d'un espace métrique compact X est convergente. C'est immédiat : la compacité de X montre que (u_n) possède une valeur d'adhérence ℓ. D'après la proposition 30 de la page 423, (u_n) converge donc vers ℓ. La réciproque est fausse : par exemple $\mathbb{R}$ est complet, mais pas compact. ∎

Proposition 45. Dans un espace compact X , une suite (u_n) converge si, et seulement si, elle possède une unique valeur d'adhérence.

Démonstration. Si (u_n) converge, la remarque de la page 402 montre que sa limite est sa seule valeur d'adhérence. Supposons inversement que (u_n) possède une valeur d'adhérence unique ℓ. Soit $\varepsilon > 0$. Posons $E := \big\{n \in \mathbb{N} \mid d(u_n, \ell) > \varepsilon\big\}$. Si E est infini, il existe une suite strictement croissante $(n_k)_{k \geqslant 0}$ formée d'éléments de E. Puisque X est compact, la suite $(u_{n_k})_{k \geqslant 0}$ possède une valeur d'adhérence ℓ', qui est aussi valeur d'adhérence de (u_n). Vu l'hypothèse, $\ell' = \ell$. Pour tout $k \in \mathbb{N}$, $d(u_{n_k}, \ell) > \varepsilon$. Par passage à la limite lorsque k tend vers $+\infty$, on en déduit $0 = d(\ell', \ell) \geqslant \varepsilon$, ce qui est absurde. Ainsi, E est fini. Soit $N \in \mathbb{N}$ un majorant de E. Pour tout $n > N$, n n'appartient pas à E, donc $d(u_n, \ell) \leqslant \varepsilon$, ce qui montre que (u_n) converge vers ℓ. ∎

Noter l'analogie entre le théorème suivant et le théorème 34 de la page 425 :

Théorème 46. Soient X un espace métrique et A une partie de X.

1. Si A est un espace compact, c'est un fermé de X.
2. Réciproquement, *si* X *est compact* et si A est fermée, A est un espace compact.

Démonstration. L'assertion 1 résulte du théorème précédent et du théorème 34 de la page 425. Inversement, supposons que X soit compact et que A soit fermée dans X. Soit (a_n) une suite de A. Elle possède une valeur d'adhérence b dans X ; b est limite d'une sous-suite de (a_n), donc $b \in \overline{A} = A$. Toute suite de A a donc une valeur d'adhérence dans A, *i.e.* A est compacte. ∎

L'intersection F d'une famille $(F_i)_{i\in I}$ de parties compactes d'un espace X est compacte, dès que l'ensemble I est non vide. En effet, soit $i_0 \in I$. Pour tout i, $F_i \cap F_{i_0}$ est une partie fermée de F_{i_0}, elle est compacte puisque F_{i_0} l'est. Mais F est aussi l'intersection des $F_i \cap F_{i_0}$ pour $i \in I$, donc F est compacte car fermée dans F_{i_0}. De même :

Proposition 47. Dans un espace métrique, toute réunion *finie* de parties compactes est compacte. Par ailleurs, tout produit fini d'espaces compacts est compact.

Démonstration. Traitons le cas de deux parties compactes A et B d'un espace X. Soit (u_n) une suite de $A \cup B$. Si $T := \{n \in \mathbb{N} \mid u_n \in A\}$ est infini, on peut extraire de (u_n) une sous-suite (u_{n_k}) formée d'éléments de A. Puisque A est compacte, on peut extraire de (u_{n_k}) une sous-suite convergeant vers un point $a \in A$. Cette sous-suite est aussi une sous-suite de (u_n), convergeant vers $a \in A \cup B$. Si T est fini, l'ensemble $\{n \in \mathbb{N} \mid u_n \in B\}$ est infini, et l'on conclut de même. D'où la compacité de $A \cup B$.
Soient de même X, Y deux espaces compacts et $\big((x_n, y_n)\big)$ une suite de $X \times Y$. Puisque X est compact, il existe une sous-suite $(x_{n_k})_{k\geqslant 0}$ de (x_n) convergeant vers un point $x \in X$. Puisque Y est compact, il existe une sous-suite $(y_{n_{k_p}})_{p\geqslant 0}$ de $(y_{n_k})_{k\geqslant 0}$ convergeant vers un point y de Y. Alors la suite $\big((x_{n_{k_p}}, y_{n_{k_p}})\big)_{p\geqslant 0}$ est extraite de $\big((x_n, y_n)\big)$, et elle converge vers (x, y). D'où la compacité de $X \times Y$. Noter les deux extractions successives. ∎

Dans le cas d'un produit dénombrables d'espaces compacts, signalons le résultat suivant, qui sera démontré dans [MPA] :

Proposition 48. Soit $\big((X_n), d_n\big)_{n\in\mathbb{N}}$ une famille dénombrable d'espaces métriques compacts. Munissons $X := \prod_{n\in\mathbb{N}} X_n$ de la distance d définie par la formule (10) de la page 398. L'espace métrique (X, d) est alors compact.

Ainsi le cube de Hilbert $[0, 1]^{\mathbb{N}}$ et l'ensemble de Cantor $\{0, 1\}^{\mathbb{N}}$ (*cf.* page 399) sont compacts. Pour l'ensemble de Cantor, nous retrouverons ce fait dans la section suivante.
Avec des hypothèses de nature différente, voici une proposition donnant une conclusion un peu plus faible que le théorème 36 de la page 426.

Proposition 49. Soient X un espace métrique et (F_n) une suite de parties compactes non vides de X, décroissante pour l'inclusion. Alors $F := \bigcap_{n\geqslant 0} F_n$ est non vide.

Démonstration. Pour tout n, soit $u_n \in F_n$ (F_n est non vide). Puisque $F_n \subset F_0$ pour tout n, (u_n) est une suite de l'espace compact F_0, elle possède une valeur d'adhérence $a \in F_0$. Pour tout $p \in \mathbb{N}$, a est aussi valeur d'adhérence de $(u_n)_{n\geqslant p}$. Mais F_p est fermée et $(n \geqslant p) \Longrightarrow (u_n \in F_n \subset F_p)$, donc $a \in F_p$. C'est vrai pour tout p, d'où $a \in F$. ∎

Le théorème suivant donne une caractérisation importante des espaces métriques compacts. La preuve en sera faite dans [MPA]. Rappelons qu'un *recouvrement* d'un ensemble X est une famille $(A_i)_{i\in I}$ de parties de X telle que $\bigcup_{i\in I} A_i = X$.

Théorème 50.
Pour un espace métrique X, les conditions suivantes sont équivalentes :

1. L'espace X est compact, *i.e.* vérifie la propriété de Bolzano-Weierstraß.
2. De tout recouvrement de X par des ouverts, on peut extraire un recouvrement fini (propriété dite de *Borel-Lebesgue*).
3. Soit $(F_i)_{i\in I}$ une famille de fermés de X d'intersection vide. Il existe une partie *finie* J de I telle que $\bigcap_{i\in J} A_i = X$ soit déjà vide.

À titre d'exercice facile, le lecteur peut établir l'équivalence des conditions 2 et 3 (par passage aux complémentaires), et montrer que la proposition précédente résulte du théorème.

3.2 Fonctions continues sur un compact

Théorème 51. Soient X, Y deux espaces métriques et $f : X \to Y$ une application continue. Si X est compact, son image par f est une partie compacte de Y.

Retenons donc ceci : *toute image continue d'un compact est compacte*. Plus précisément, si $f : X \to Y$ est continue et si A est une partie compacte de X, son image $f(A)$ est compacte (appliquer le théorème à la restriction $f_{|A}$).

Démonstration. Soit (y_n) une suite de $\operatorname{Im} f$. Pour tout n, il existe $x_n \in X$ tel que $f(x_n) = y_n$. On peut extraire de (x_n) une sous-suite (x_{n_k}) convergeant vers un point $a \in X$, car X est compact. Comme f est continue en a, (y_{n_k}) converge vers $f(a) \in \operatorname{Im} f$. Ainsi (y_{n_k}) est une sous-suite de (y_n) convergeant dans $\operatorname{Im} f$, donc $\operatorname{Im} f$ est compact. ∎

Corollaire 52. Soient X, Y deux espaces métriques et $f : X \to Y$ une application continue *bijective*. Si X est compact, f est un *homéomorphisme* de X sur Y.

Démonstration. Nous devons montrer que la réciproque $f^{-1} : Y \to X$ est continue. Cela revient à dire que, si F est un fermé de X, son image réciproque par f^{-1}, à savoir $f(F)$, est fermée dans Y. Puisque X est compact, F l'est aussi. Il en résulte que $f(F)$ est compact (appliquer le théorème à $f_{|_F}$), donc fermé dans Y. ∎

Comme illustration, revenons sur l'ensemble de Cantor $\big(\{0,1\}^{\mathbb{N}}, d\big)$ (voir page 399). Soit $f : \{0,1\}^{\mathbb{N}} \to [0,1]$ l'application définie par la formule :

$$f(u) \;=\; \frac{2}{3}\sum_{n=0}^{+\infty} \frac{u_n}{3^n} \quad \text{lorsque} \quad u = (u_n)_n \in \{0,1\}^{\mathbb{N}}.$$

L'image de f est une partie K_3 de $[0,1]$ appelée *ensemble triadique de Cantor*.

Proposition 53. L'application f est un homéomorphisme de $\{0,1\}^{\mathbb{N}}$ (muni de la distance d) sur K_3. De plus, K_3 est un compact de $\mathbb{R}$.

Démonstration. Soient $u = (u_n)_n$ et $v = (v_n)$ deux éléments de $\{0,1\}^{\mathbb{N}}$ distincts. Notons p le plus petit indice tel que $u_p \neq v_p$. À l'ordre près, on peut supposer que $u_p = 1$ et $v_p = 0$. Pour tout n, posons $a_n = 1$ si $u_n \neq v_n$ et $a_n = 0$ si $u_n = v_n$. Par définition de d, on a $d(u,v) = \sum\limits_{n=p}^{+\infty} 2^{-n} a_n = 2^{-p} + \sum\limits_{n=p+1}^{+\infty} 2^{-n} a_n$, d'où $2^{-p} \leqslant d(u,v) \leqslant 2^{1-p}$. De la même façon, $|f(u) - f(v)| = (2/3)\sum\limits_{n=p}^{+\infty} 3^{-n} a_n = (2/3)\big[3^{-p} + \sum\limits_{n=p+1}^{+\infty} 3^{-n}\,|u_n - v_n|\big]$, d'où l'on déduit aisément l'encadrement $3^{-(p+1)} \leqslant |f(u) - f(v)| \leqslant 3 \cdot 3^{-(p+1)}$. Cela prouve l'injectivité de f.

Montrons que $f^{-1} : K_3 \to \{0,1\}^{\mathbb{N}}$ est *uniformément* continue. Soit $\varepsilon > 0$. Il existe $m \in \mathbb{N}$ tel que $2^{1-m} \leqslant \varepsilon$. Posons $\alpha = 3^{-(m+1)}$. Soient u, v, p comme ci-dessus, supposons que $|f(u) - f(v)| \leqslant \alpha$. Alors $3^{-(p+1)} \leqslant 3^{-(m+1)}$, d'où $p \geqslant m$ puis $d(u,v) \leqslant 2^{1-p} \leqslant 2^{1-m} \leqslant \varepsilon$, ce qui établit l'uniforme continuité de f^{-1}. Un argument analogue montrerait que f est uniformément continue, mais c'est inutile : si K_3 est compact, le corollaire précédent, appliqué à f^{-1}, montrera que f^{-1} est un homéomorphisme, et donc f est un homéomorphisme.

Il reste à établir que K_3 est un fermé de $\mathbb{R}$, c'est l'objet de la construction qui suit. Coupons le segment $F_0 := [0,1]$ en trois segments égaux et enlevons le tiers central (ouvert). Il reste un fermé $F_1 := [0,1/3] \cup [2/3,1]$. Ôtons de chacun des segments composant F_1 son tiers central. Il reste le fermé $F_2 := [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1]$:

Itérons le procédé. Après n itérations, il reste un fermé F_n réunion de 2^n segments disjoints, chacun étant de longueur 3^{-n}. En ôtant de chacun de ces segments son tiers central ouvert, on

obtient un fermé F_{n+1} réunion de 2^{n+1} segments disjoints, chacun étant de longueur $3^{-(n+1)}$. L'intersection F des F_n, $n \in \mathbb{N}$, est fermée bornée, donc compacte. Il suffit ainsi de montrer que $K_3 = F$.

Nous laissons le lecteur vérifier par récurrence sur n que les 2^n segments constituant F_n sont les $[3^{-n}t, 3^{-n}(t+1)]$, t décrivant l'ensemble T_n des entiers de $[\![0, 3^n - 1]\!]$ dont l'écriture en base 3 ne comporte pas le chiffre 1, *i.e.* l'ensemble des $\sum_{k=0}^{n-1} 3^k a_k$, chaque a_k valant 0 ou 2.

Soit $u = (u_n)_n \in \{0,1\}^{\mathbb{N}}$. Montrons que $f(u) \in F$. Si $n \in \mathbb{N}^*$, posons $t_n := 2\sum_{k=0}^{n-1} 3^{n-k-1}u_k$. Alors $t_n \in T_n$, et l'on a :

$$\frac{t_n}{3^n} \;=\; \frac{2}{3}\sum_{k=0}^{n-1}\frac{u_k}{3^k} \;\leqslant\; f(u) \;\leqslant\; \frac{t_n}{3^n} + (2/3)\sum_{k=n}^{+\infty}\frac{1}{3^k} \;=\; \frac{t_n+1}{3^n}\cdot$$

Ainsi $f(u) \in F_n$ pour tout n, d'où $K_3 \subset F$. Soit inversement $x \in F$. Pour tout $n \in \mathbb{N}$, il existe un unique $t_n \in T_n$ tel que $x \in [3^{-n}t_n, 3^{-n}(t_n+1)]$. Vu le procédé définissant F_{n+1} à partir de F_n, chaque t_{n+1} est égal à $3t_n$ ou $3t_n + 2$, donc s'écrit $t_{n+1} := 3t_n + 2u_n$, où $u_n \in \{0,1\}$. Pour tout $n \in \mathbb{N}$, on montre (par récurrence sur n) que $t_n = 2\sum_{k=0}^{n-1} 3^{n-k-1}u_k$. Il en résulte aussitôt que $x = f(u) \in K_3$, où $u := (u_n)_n$. Ainsi $F \subset K_3$, d'où la proposition. ∎

L'espace K_3 est un prototype d'ensemble « fractal ». Il est à la fois « assez gros » (il est en bijection avec $\mathbb{R}$) et « de longueur nulle », *cf.* l'exercice II.1.43. L'exercice II.1.27 en donne une définition « dynamique ».

On déduit du théorème 51 de la page 435 l'un des résultats les plus importants en analyse :

Théorème 54. Sur un espace métrique *compact* X, toute fonction réelle continue f est bornée et atteint ses bornes $\sup f$ et $\inf f$.
Autrement dit, il existe $a, b \in X$ tels que $f(a) \leqslant f(x) \leqslant f(b)$ pour tout $x \in X$.

Démonstration. Le théorème précédent montre que $\operatorname{Im} f$ est une partie compacte de $\mathbb{R}$. Elle est donc bornée, *i.e.* les bornes $\inf(\operatorname{Im} f)$ et $\sup(\operatorname{Im} f)$ sont finies. Mais $\operatorname{Im} f$ est aussi fermée, donc elle contient ces bornes, d'après l'exemple de la page 405. ∎

Le théorème précédent généralise le théorème des bornes vu dans le cours de [L1], pour un compact de $\mathbb{R}^n$. C'est un pur résultat *d'existence*. Il dit ainsi que $m := \inf f$ est finie, et que l'équation $f(x) = m$ a *au moins* une solution x (*i.e.* f atteint son minimum absolu en x). Le théorème ne dit pas comment trouver les solutions. Pour résoudre ce « problème de minimum », il faut faire appel à d'autres méthodes, par exemple au calcul différentiel.

Exercice 10.

Dans le plan, considérons trois cercles $\mathcal{C}_i$ $(i = 1,2,3)$ extérieurs les uns aux autres. Montrer qu'il existe un triangle $M_1M_2M_3$, où $M_i \in \mathcal{C}_i$ pour tout i, dont le périmètre $p := M_1M_2 + M_2M_3 + M_3M_1$ est minimum.

Solution. Pour chaque i, $\mathcal{C}_i$ est une partie compacte de $\mathbb{R}^2$, donc $X := \mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{C}_3$ est une partie compacte de $E := \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2$. La fonction distance $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$

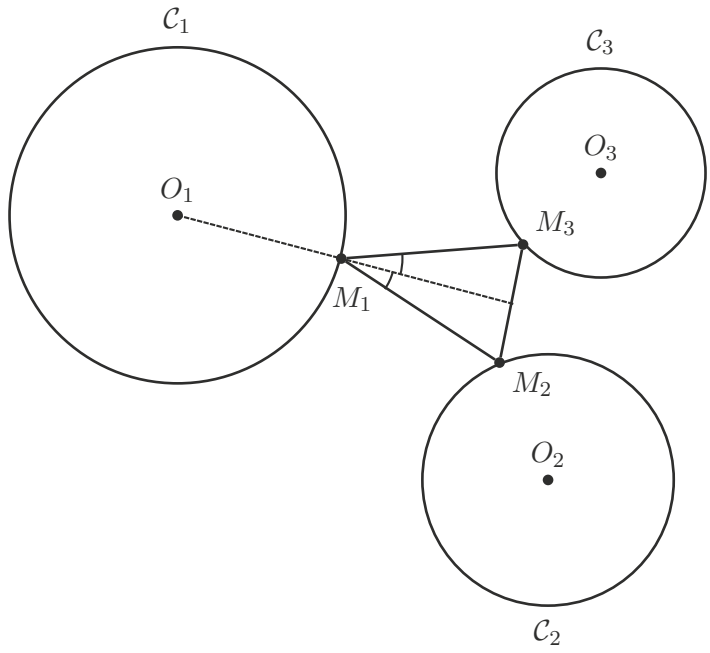


FIGURE II.1.2 – Triangle de périmètre minimum.

étant continue, la fonction $(M_1, M_2, M_3) \mapsto p := M_1M_2 + M_2M_3 + M_3M_1$ est continue sur $(\mathbb{R}^2)^3$. La restriction f de cette fonction à X est encore continue. Puisque X est compact, le théorème précédent montre que f atteint son minimum absolu en un point (M_1, M_2, M_3) de X. On dit que $M_1M_2M_3$ est un « triangle de périmètre minimum », puisque l'on a $M_1M_2 + M_2M_3 + M_3M_1 \leqslant N_1N_2 + N_2N_3 + N_3N_1$ pour tout $(N_1, N_2, N_3) \in \mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{C}_3$.
La détermination d'un tel triangle se ramène à la résolution d'un système d'équations algébriques, mais cela dépasse le cadre de cet exercice. En revanche, l'exercice II.1.28 indique une propriété géométrique simple de ce triangle.

Corollaire 55. Dans un espace métrique X, la distance entre deux parties compactes non vides A et B est toujours atteinte.

Si donc B est compacte non vide et $x \in X$, $d(x, B)$ est atteinte, car $\{x\}$ est compact.

Démonstration. La fonction $(a, b) \mapsto d(a, b)$ est continue sur $A \times B$.

Comme $A \times B$ est compact (proposition 47 de la page 434), cette fonction atteint sa borne inférieure $d(A, B)$. ∎

Comme application de ce corollaire, l'exercice suivant propose un exemple d'espace métrique ne provenant pas d'un e.v. n. (il est prolongé par les exercices II.1.26 et II.1.27).

Exercice 11.
Soit $\mathcal{K}$ l'ensemble des compacts non vides de $\mathbb{R}^n$. On note d la distance usuelle sur $\mathbb{R}^n$. Pour tout couple $(A,B)\in\mathcal{K}^2$, on pose :

$$\delta(A,B) \;:=\; \max\left(\sup_{a\in A} d(a,B),\ \sup_{b\in B} d(A,b)\right). \tag{25}$$

Montrer que δ est une distance sur $\mathcal{K}$ (elle est appelée *distance de Haussdorf*).

Solution. Noter que $\delta(A,B)$ est un maximum de deux bornes supérieures, et chaque $d(a,B)$, par exemple, est une borne inférieure ! Soient $A,B\in\mathcal{K}$ et $b_0\in B$. Pour tout $a\in A$, $d(a,B) := \inf\big\{d(a,b)\mid b\in B\big\}\leqslant d(a,b_0)$, donc la première borne supérieure de l'énoncé est finie. La formule (25) définit donc une application (symétrique) $\delta : \mathcal{K}\times\mathcal{K}\to\mathbb{R}_+$.

Soit $A\in\mathcal{K}$. Pour tout $a\in A$, $d(a,A)=0$, d'où $\delta(A,A)=0$. Soient inversement $A,B\in\mathcal{K}$ tels que $\delta(A,A)=0$. Chacune des deux bornes supérieures figurant en (25) est nulle. Ainsi $d(a,B)=0$ pour tout $a\in A$; puisque B est compact, le corollaire précédent montre que $a\in B$ pour tout $a\in A$, donc $A\subset B$. Symétriquement, $B\subset A$, d'où $A=B$.

Il reste l'inégalité triangulaire. Soient $A,B,C\in\mathcal{K}$. Partons d'un point $a\in A$. Puisque B est compact, la distance $d(a,B)$ est atteinte en un point $b\in B$: $d(a,b)=d(a,B)$. De même il existe un point $c\in C$ tel que $d(b,c)=d(b,C)$. Alors :

$$d(a,b)=d(a,B)\leqslant\sup_{x\in A} d(x,B)\leqslant\delta(A,B),\ d(b,c)=d(b,C)\leqslant\sup_{y\in B} d(y,C)\leqslant\delta(B,C),$$

d'où $d(a,c)\leqslant d(a,b)+d(b,c)\leqslant\delta(A,B)+\delta(B,C)$. Vu la définition de $d(a,C)$ comme borne inférieure, $d(a,C)\leqslant d(a,c)$. Ainsi $d(a,C)\leqslant\delta(A,B)+\delta(B,C)$ pour tout $a\in A$. Passons à la borne supérieure : $\sup\limits_{a\in A} d(a,C)\leqslant\delta(A,B)+\delta(B,C)$. De manière symétrique, $\sup\limits_{c\in C} d(A,c)\leqslant\delta(A,B)+\delta(B,C)$. D'où $\delta(A,C)\leqslant\delta(A,B)+\delta(B,C)$, par définition de $\delta(A,C)$.

Voici un dernier résultat sur la compacité, utile notamment pour l'intégrale de Riemann :

Théorème 56 (Heine). Soient X,Y deux espaces métriques et $f : X\to Y$ une application continue. Si X est compact, f est *uniformément continue*.

Démonstration. Raisonnons par l'absurde, en prenant la négation de « f est uniformément continue ». Il existe donc un réel $\varepsilon>0$ ayant la propriété suivante : pour tout $\alpha>0$, il existe $(a,b)\in X^2$ tel que $d(a,b)\leqslant\alpha$ mais $d\big(f(a),f(b)\big)>\varepsilon$. En particulier, pour tout entier $n\geqslant 1$, il existe $(a_n,b_n)\in X^2$ tel que $d(a_n,b_n)\leqslant 1/n$ mais $d\big(f(a_n),f(b_n)\big)>\varepsilon$. Puisque X est compact, on peut extraire de (a_n) une sous-suite $(a_{n_k})_{k\geqslant 0}$ convergeant vers un point $a\in X$. Puisque $d(a_{n_k},b_{n_k})\leqslant 1/n_k$ pour tout k, la suite $(b_{n_k})_{k\geqslant 0}$ converge aussi vers a. La continuité de f entraîne alors que les deux suites $\big(f(a_{n_k})\big)_{k\geqslant 0}$ et $\big(f(b_{n_k})\big)_{k\geqslant 0}$ convergent vers $f(a)$. C'est absurde, car $d\big(f(a_{n_k}),f(b_{n_k})\big)>\varepsilon$ pour tout $k\geqslant 0$. ∎

4 Espaces vectoriels normés de dimension finie

4.1 Théorèmes fondamentaux

Nous allons démontrer simultanément les deux théorèmes fondamentaux suivants concernant les e.v. n. de dimension finie (dans le second, *chacune* des conclusions est essentielle) :

Théorème 57. Sur un espace vectoriel réel ou complexe de dimension finie, toutes les normes (et il en existe) sont équivalentes.

Théorème 58. Soit E un espace vectoriel normé de dimension finie $n \geqslant 1$. Alors :

1. L'espace E est complet.
2. Une partie de E est compacte si, et seulement si, elle est fermée et bornée.
3. Toute application *linéaire* de E dans un e.v. n. quelconque est continue.

Démonstration. L'existence d'une norme sur un K-espace vectoriel de dimension finie (K étant égal à $\mathbb{R}$ ou $\mathbb{C}$) a été prouvée dans l'exemple 7 de la page 396. Par ailleurs, une norme sur un $\mathbb{C}$-espace vectoriel E est aussi une norme sur le $\mathbb{R}$-espace vectoriel E, et toute application $\mathbb{C}$-linéaire est *a fortiori* $\mathbb{R}$-linéaire. Nous pouvons donc supposer que E est un $\mathbb{R}$-espace vectoriel de dimension finie $n \geqslant 1$, le cas où E est nul étant trivial. L'idée générale est que savons déjà que le deuxième théorème est vrai lorsque $E := \mathbb{R}^n$ (*cf.* le théorème 32 de la page 424, le théorème de Bolzano-Weierstraß et le corollaire 19 de la page 416), et nous allons nous ramener à ce cas.

Soit $\|\cdot\|$ une norme quelconque sur E. Choisissons une base $(e_1, \dots, e_n)$ de E, et notons g l'isomorphisme linéaire $(x_1, \dots, x_n) \mapsto x_1e_1 + x_2e_2 + \cdots + x_ne_n$ de $\mathbb{R}^n$ sur E. Munissons $\mathbb{R}^n$ de sa norme usuelle, également notée $\|\cdot\|$. D'après le corollaire 19 de la page 416, g est continue. Sur la sphère unité S de $\mathbb{R}^n$, la fonction $x \mapsto \|g(x)\|$ est continue et ne s'annule pas. Il existe donc deux réels m, M, $0 < m \leqslant M$, tels que $m \leqslant \|g(x)\| \leqslant M$ pour tout $x \in S$. La propriété d'homogénéité des normes et la linéarité de g entraînent alors :

$$\forall x \in \mathbb{R}^n, \quad m\|x\| \leqslant \|g(x)\| \leqslant M\|x\|. \qquad (*)$$

Pour tout $y \in E$ on en déduit, en posant $x := g^{-1}(y)$, l'inégalité $\|g^{-1}(y)\| \leqslant m^{-1}\|y\|$, ce qui prouve que g^{-1} est continue. Ainsi g est un *homéomorphisme* de $\mathbb{R}^n$ sur E.

Soient F un e.v. n. et $f : E \to F$ une application linéaire. D'après le corollaire 19 de la page 416, $f \circ g : \mathbb{R}^n \to F$ est continue. Comme g^{-1} est continue, $f = (f \circ g) \circ g^{-1}$ l'est aussi, ce qui prouve l'assertion 3 du théorème 58.

Soient A une partie de E et $B := g^{-1}(A)$. Puisque g est un homéomorphisme, A est compacte (resp. fermée) si, et seulement si, B l'est. La propriété $(*)$ montre que A est bornée si, et seulement si, B l'est. L'assertion 2 du théorème 58 est alors vraie, puisque les compacts de $\mathbb{R}^n$ sont les fermés bornés.

Soit (u_n) une suite de Cauchy de E. La propriété $(*)$ fait que $\big(g^{-1}(u_n)\big)$ est une suite de Cauchy de $\mathbb{R}^n$, elle converge donc, car $\mathbb{R}^n$ est complet. Puisque g est continue, il en résulte que (u_n) converge. L'espace E est donc complet. Le théorème 58 de la page précédente est ainsi démontré.

Soit $\|\cdot\|'$ une autre norme sur E. Notons g' l'application $x \mapsto g(x)$ de $\big(\mathbb{R}^n, \|\cdot\|\big)$ dans $\big(E, \|\cdot\|'\big)$ et $\varphi : \big(E, \|\cdot\|\big) \to \big(E, \|\cdot\|'\big)$ l'application identité. Évidemment, $g' = \varphi \circ g$. Nous avons montré que g est un homéomorphisme, et de même pour g' $\big($remplacer $\|\cdot\|$ par $\|\cdot\|'\big)$. Il en résulte que φ est un homéomorphisme, *i.e.* que $\|\cdot\|$ et $\|\cdot\|'$ sont équivalentes (*cf.* la proposition 27 de la page 421). Le théorème 57 de la page précédente est donc démontré. ∎

Comme application, il sera prouvé dans le module I.6 que le groupe orthogonal $O_n(\mathbb{R})$ est compact, parce que c'est un fermé borné de $M_n(\mathbb{R})$.

Corollaire 59. Tout sous-espace vectoriel V de dimension finie d'un e.v. n. quelconque E est fermé dans E.

Démonstration. Le théorème précédent montre que V est complet, il est donc fermé, en vertu du théorème 34 de la page 425. ∎

Corollaire 60. Dans un e.v. n. de dimension finie, les boules fermées et les sphères sont compactes.

Démonstration. Cela résulte immédiatement de l'assertion 2 du théorème 58 de la page précédente. ∎

En dimension infinie, l'exemple de la page 433 montre que ce résultat n'est pas toujours vrai. On peut montrer qu'il est en fait toujours faux (c'est un théorème de Riesz).

Proposition 61. Soient $E_1, \dots, E_n$ des e.v. n. de dimension finie. Toute application n-linéaire de $E_1 \times E_2 \times \cdots \times E_n$ dans un e.v. n. quelconque est continue.

Démonstration. C'est vrai si $n := 1$. Soit $n \geqslant 2$; supposons la conclusion vraie au rang $n-1$. Soit $f : E := E_1 \times E_2 \times \cdots \times E_n \to F$ une application n-linéaire. Choisissons une base $(e_1, \dots, e_p)$ de E_n. Les normes sur E_n étant équivalentes, on peut supposer que, si $x_n := \lambda_1 e_1 + \cdots + \lambda_p e_p \in E_n$ (les λ_j sont dans K), on a $\|x_n\| = \max(|\lambda_j|)$. Pour chaque j, $(x_1, \dots, x_{n-1}) \mapsto f(x_1, \dots, x_{n-1}, e_j)$ est une application $(n-1)$-linéaire de $E_1 \times \cdots \times E_{n-1}$ dans F. Vu l'hypothèse de récurrence, il existe $C \geqslant 0$ tel que, pour tout $(x_1, \dots, x_{n-1}) \in E_1 \times \cdots \times E_{n-1}$ et tout $j \in [\![1, p]\!]$, on ait : $\|f(x_1, \dots, x_{n-1}, e_j)\| \leqslant C\|x_1\|\|x_2\| \cdots \|x_{n-1}\|$. Nous laissons le lecteur en déduire que $\|f(x_1, \dots, x_n)\| \leqslant nC\|x_1\|\|x_2\| \cdots \|x_n\|$ pour tout $(x_1, \dots, x_n) \in E$. ∎

Exercice 12.

Soit V un sous-espace vectoriel de dimension finie d'un e.v. n. E.
Montrer que, pour tout $x \in E$, la distance $d(x, V)$ est atteinte.

Solution. Si $x \in V$, $0 = d(x, V) = d(x, x)$. Supposons $x \notin V$. En particulier $x \neq 0$, donc $r := \|x\| > 0$. La boule $\overline{B}(x, r)$ est fermée et bornée, mais pas nécessairement compacte, car nous ne supposons pas que E est de dimension finie. Par contre, posons $F := V \cap \overline{B}(x, r)$. C'est une partie fermée et bornée de V, elle est donc compacte puisque V, lui, est de dimension finie. La fonction $y \mapsto d(x, y) = \|x - y\|$ est continue sur le compact F, elle atteint donc son minimum absolu en un point $a \in F$. En particulier, puisque $0 \in F$, $d(x, a) \leqslant d(x, 0) = r$. Soit alors $y \in V$. Si $y \in F$, on a $d(x, y) \geqslant d(x, a)$. C'est vrai aussi si $y \notin F$, c'est-à-dire si $d(x, y) > r$, car dans ce cas $d(x, y) > r \geqslant d(x, a)$. Ainsi $d(x, a) \leqslant d(x, y)$ pour tout $y \in V$, d'où $d(x, V) = d(x, a)$, puisque $a \in V$.

Exercice 13.

Soit f une application linéaire d'un e.v. n. E *de dimension finie* dans un e.v. n. F. Il existe $x \in E$ tel que $\|x\| = 1$ et $\|f(x)\| = |||f|||$.

Solution. Bien sûr, f est continue. L'application $x \mapsto \|f(x)\|$ de la sphère unité S de E dans $\mathbb{R}$ est continue, elle atteint son maximum absolu, à savoir $|||f|||$, car S est compacte.

Remarque. Soit E un espace vectoriel réel ou complexe de dimension finie. Puisque toutes les normes sur E sont équivalentes, les *Conséquences* de la page 422 s'appliquent. Soit par exemple (u_n) une suite de E. De deux choses l'une : ou bien (u_n) converge dans $\big(\|\cdot\|\big)$ quel que soit le choix de la norme $\|\cdot\|$ sur E, ou bien elle diverge quel que soit ce choix. On peut dire que la convergence est « intrinsèque », *i.e.* indépendante du choix d'une norme. De même, lorsqu'on dit qu'une partie de E est ouverte (ou fermée, bornée, compacte, dense), il s'agit d'une propriété intrinsèque.

Soit f une application linéaire d'un e.v. n. E dans un e.v. n. F. Si $\dim E$ est finie, f est continue, nous l'avons vu. Si c'est F qui est de dimension finie, ce n'est plus toujours vrai. Ainsi, munissons $E := \mathcal{C}\big([-1, 1], \mathbb{R}\big)$ de la norme $\|\cdot\|_2$ associée au produit scalaire $(f \mid g) = \displaystyle\int_{-1}^{1} f(t)g(t)\,\mathrm{d}t$ (*cf.* le module I.4). La forme linéaire $L : f \mapsto f(1)$ sur E n'est pas continue car, si l'on pose $f_n(t) := t^n$ pour tous $n \in \mathbb{N}^*$ et $t \in [-1, 1]$, on a pour tout n : $\|f_n\|_2 = \big[2/(2n+1)\big]^{1/2}$ et $L(f_n) = 1$. Les exercices II.1.6 et II.1.7 fournissent d'autres contre-exemples. Si E est de dimension infinie, le lemme de Zorn (voir le module I.1 du cours de [L1]) permet de montrer qu'il existe toujours des formes linéaires $f : E \to K$ non continues.

4.2 Normes matricielles

Soit $n \in \mathbb{N}^*$. Identifions toute matrice $A \in M_n(\mathbb{C})$ avec l'endomorphisme $X \mapsto AX$ de $\mathbb{C}^n$ (X est un vecteur unicolonne), endomorphisme *continu*, car on est en dimension finie. *Choisissons* une norme $\|\cdot\|$ sur $\mathbb{C}^n$. La norme fonctionnelle $|||\cdot|||$ correspondante est alors définie par la formule :

$$|||A||| \;:=\; \sup_{X\neq 0} \frac{\|AX\|}{\|X\|}. \tag{26}$$

La norme $|||\cdot|||$ *dépend* du choix de $\|\cdot\|$. La multiplication matricielle correspondant à la composition des endomorphismes, la proposition 23 de la page 418 donne la propriété :

$$\forall A, B \in M_n(\mathbb{C}), \quad |||AB||| \;\leqslant\; |||A|||\,|||B|||. \tag{27}$$

En général, une norme sur $M_n(\mathbb{C})$ possédant cette propriété est appelée *norme matricielle*. Dans le cas présent, la norme matricielle $|||\cdot|||$ est dite *subordonnée* à la norme $\|\cdot\|$ sur $\mathbb{C}^n$.

Exemple. Explicitons par exemple la norme $|||\cdot|||_\infty$ subordonnée à la norme $\|\cdot\|_\infty$ sur $\mathbb{C}^n$, donnée par $\|X\|_\infty := \max(|x_1|, \dots, |x_n|)$ si $X := (x_1, \dots, x_n) \in \mathbb{C}^n$ (voir d'autres exemples dans les exercices). Soit $A := (a_{i,j}) \in M_n(\mathbb{C})$, voyons comment calculer $|||A|||_\infty$.
Soit $X := (x_1, \dots, x_n) \in \mathbb{C}^n$, et posons $Y := AX := (y_1, \dots, y_n)$. Soit $i \in [\![1, n]\!]$. De $y_i = \sum_{j=1}^{n} a_{i,j}x_j$, on tire $|y_i| \leqslant \sum_{j=1}^{n} |a_{i,j}|\,|x_j| \leqslant \|X\|_\infty \sum_{j=1}^{n} |a_{i,j}|$, car $|x_j| \leqslant \|X\|_\infty$ pour tout j.
Si l'on note L_i la ligne d'indice i, la somme $\sum_{j=1}^{n} |a_{i,j}|$ est la norme $\|L_i\|_1$ de L_i (*cf.* l'exemple 4 de la page 395). Ainsi $|y_i| \leqslant \|L_i\|_1\|X\|_\infty$. En faisant varier i, on obtient :

$$\|AX\|_\infty = \|Y\|_\infty := \max(|y_1|, \dots, |y_n|) \leqslant \max(\|L_1\|_1, \dots \|L_n\|_1)\|X\|_\infty.$$

C'est vrai pour tout $X \in \mathbb{C}^n$, donc $\|A\|_\infty \leqslant \max(\|L_1\|_1, \dots \|L_n\|_1) := M$. En fait il y a égalité :

$$\|A\|_\infty \;=\; \max(\|L_1\|_1, \dots \|L_n\|_1). \tag{28}$$

En effet, on peut supposer $A \neq 0$. Soit i un indice tel que $\|L_i\|_1$ soit maximum. On définit $X := (x_1, \dots, x_n)$ en posant, pour tout j, $x_j := 0$ si $a_{i,j} = 0$ et $x_j = \bar{a}_{i,j}/\,|a_{i,j}|$ si $a_{i,j} \neq 0$. Alors $\|X\|_\infty = 1$, et $a_{i,j}x_j = |a_{i,j}|$ pour tout j, donc $y_i = \sum_j |a_{i,j}| = \|L_i\|_1 = M$. Il en résulte que $\|AX\|_\infty = \|Y\|_\infty \geqslant |y_i| = M$, d'où $|||A|||_\infty \geqslant M$, ce qui prouve l'égalité (28).

Lorsqu'on munit $M_n(\mathbb{C})$ d'une norme matricielle, on obtient ce que l'on appelle une *algèbre de Banach*. En général, appelons *algèbre normée* toute K-algèbre $\mathcal{A}$ munie d'une norme vérifiant la propriété suivante : $\|ab\| \leqslant \|a\|\|b\|$ pour tous $a, b \in \mathcal{A}$. Une algèbre normée complète est aussi appelée *algèbre de Banach*. Dans une algèbre normée $\mathcal{A}$, la multiplication : $\mathcal{A} \times \mathcal{A} \to \mathcal{A}$ est bilinéaire ; elle est continue, d'après le théorème 24 de la page 419. Si $x \in \mathcal{A}$, on a $\|x^n\| \leqslant \|x\|^n$ pour tout entier $n \geqslant 1$ (récurrence immédiate).
Si l'algèbre $M_n(\mathbb{C})$ est munie d'une norme matricielle, c'est une algèbre normée (d'après l'inégalité (27)), complète car de dimension finie. Si E est un e.v.n., la proposition 23 de la page 418 montre que $\big(\mathcal{L}_c(E), |||\cdot|||\big)$ est une algèbre normée. Si l'algèbre $\mathcal{C}\big([a, b], K\big)$ est munie de la norme $\|\cdot\|_\infty$, c'est une algèbre normée (la multiplication étant la multiplication ordinaire des fonctions), qui est complète, nous le verrons dans le module suivant.

Exercice 14.
Soit $\mathcal{A}$ une algèbre de Banach et x un élément de $\mathcal{A}$ tel que $\|x\| < 1$. Montrer que $1 - x$ est inversible, son inverse étant $\sum_{n=0}^{+\infty} x^n$.

Solution. Pour tout n, on a $\|x^n\| \leqslant \|x\|^n$. Puisque $\|x\| < 1$, la série $\sum_{n\geqslant 0} \|x^n\|$ converge, *i.e.* la série $\sum_{n\geqslant 0} x^n$ converge absolument. Puisque $\mathcal{A}$ est complète, cette série converge (théorème 41 de la page 431), notons S sa somme, limite de la suite (S_p), où $S_p := \sum_{n=0}^{p} x^n$.
L'application linéaire $t \mapsto t(1-x)$ de $\mathcal{A}$ dans elle-même est continue (pourquoi ?). Il en résulte que $S(1-x)$ est la limite de $S_p(1-x) = 1 - x^{p+1}$ quand p tend vers $+\infty$. D'où $S(1-x) = 1$, car la suite (x^n) converge vers 0.
On montre de même que $(1-x)S = 1$.

5 Connexité. Convexité

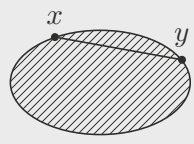
La notion de convexité (étude des parties et fonctions convexes) a de nombreuses applications : géométrie différentielle, analyse fonctionnelle, optimisation, programmation linéaire, etc. (*cf.* [MPA1] et [MPA]). Nous en présentons les aspects élémentaires. Cette section porte également sur la notion de connexité. On doit d'ailleurs prendre garde à ne pas confondre « connexe » et « convexe », malgré la ressemblance de ces termes.

5.1 Parties convexes

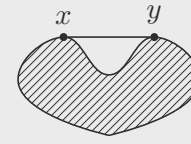
Dans cette section, $E, F, \ldots$ sont des espaces vectoriels réels, de dimension finie ou non.
L'espace vectoriel E peut être considéré comme espace affine (*cf.* le cours de [L1]). Si x, y sont deux « points » de E, le « vecteur » que l'on note en général $\overrightarrow{xy}$ dans un espace affine est simplement $y - x$. Soient $x_1, \ldots, x_n \in E$. Si $\lambda_1, \ldots, \lambda_n$ sont des réels *de somme* 1, le barycentre de $(x_1, \lambda_1), \ldots, (x_n, \lambda_n)$ est $\lambda_1 x_1 + \cdots + \lambda_n x_n$ (se reporter à la définition). Le lecteur doit s'en convaincre et voir, par exemple, que le barycentre de $(x, 1)$ et $(y, 1)$ est $(x + y)/2$.
Soient $x, y \in E$, $x \neq y$. Lorsque λ décrit $\mathbb{R}$, le point $(1-\lambda)x + \lambda y$ décrit la droite affine passant par x et y. Lorsque λ décrit $[0, 1]$, $(1-\lambda)x + \lambda y$ décrit le *segment* déterminé par x et y et noté $[x, y]$. On pose aussi $]x, y[:= [x, y] \setminus \{x, y\}$, $[x, y[:= [x, y] \setminus \{y\}$ et $]x, y] := [x, y] \setminus \{x\}$.

Définition 26. Une partie A de E est dite *convexe* si elle contient le segment joignant deux quelconques de ses points, ce qui se traduit ainsi :

$$\forall x \in A,\ \forall y \in A,\ \forall \lambda \in [0,1], \quad (1-\lambda)x + \lambda y \in A. \tag{29}$$

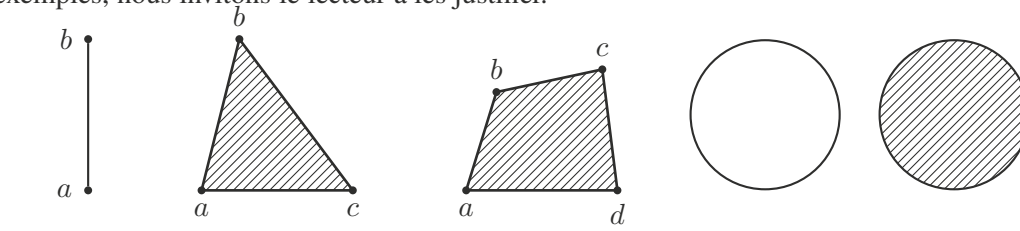


convexe



non convexe

Dans cette définition, $(1-\lambda)x+\lambda y$ est l'image de y par l'homothétie de centre x et de rapport λ. Disons que la partie A est *étoilée* par rapport à l'un de ses points x si A si toute homothétie h de centre x et de rapport $\lambda \in [0,1]$ laisse stable A, ce qui signifie que $h(A) \subset A$. La définition précédente se traduit ainsi : *une partie A est convexe si, et seulement si, elle est étoilée par rapport à chacun de ses points*.
La partie vide, un point (singleton) et E sont convexes. Tout sous-espace affine de E est convexe, c'est un bon exercice de le vérifier. Les parties convexes de $\mathbb{R}$ sont exactement les intervalles : la convexité d'un intervalle est évidente, et la réciproque est un exercice facile. Toute intersection de parties convexes de E est convexe. Si $S \subset E$, l'intersection $\mathrm{C}onv(S)$ des parties convexes de E contenant S est donc, pour l'inclusion, la plus petite partie convexe de E contenant S, on l'appelle *enveloppe convexe* de S. Voici quelques exemples, nous invitons le lecteur à les justifier.



$\mathrm{Conv}(\{a,b\})$ $\mathrm{Conv}(\{a,b,c\})$ $\mathrm{Conv}(\{a,b,c,d\})$ S $\mathrm{Conv}(S)$

FIGURE II.1.3 – Exemples d'enveloppes convexes.

On appelle *combinaison convexe* des points $x_1,\dots,x_n \in E$ tout $x \in E$ de la forme $x=\lambda_1 x_1+\cdots+\lambda_n x_n$, où les λ_i sont des réels *positifs* de somme 1.

Proposition 62.

1. Si $C \subset E$ est convexe, toute combinaison convexe de points de C est dans C.
2. Soit $f : E \to F$ une application affine. Si $C \subset E$ est convexe, $f(C)$ est convexe ; si $D \subset F$ est convexe, $f^{-1}(D)$ est convexe.
3. Si $S \subset E$, $\mathrm{C}onv(S)$ est formée des combinaisons convexes de points de S.

Démonstration. Si $x=\lambda_1 x_1+\cdots+\lambda_n x_n$ est une combinaison convexe de points de C, montrons par récurrence sur n que $x \in C$. C'est vrai si $n=1$. Supposons $n \geqslant 2$ et le résultat vrai pour $n-1$. Si $\lambda_n := 1$, $x = x_n \in C$. Sinon, $s := \lambda_1+\cdots+\lambda_{n-1} = 1-\lambda_n > 0$. On a $x=(1-\lambda_n)x'+\lambda_n x_n$, où $x' := \lambda_1 s^{-1}x_1+\cdots+\lambda_{n-1}s^{-1}x_{n-1}$. Vu l'hypothèse de récurrence, $x' \in C$, donc $x=(1-\lambda_n)x'+\lambda_n x_n \in C$.
Pour l'assertion 2, f conserve les barycentres (*cf.* le cours de [L1]). Si $x,y \in E$ et $\lambda \in [0,1]$, on a donc $f\big((1-\lambda)x+\lambda y\big) = (1-\lambda)f(x)+\lambda f(y)$, d'où aussitôt la conclusion.
Pour l'assertion 3, $\mathrm{C}onv(S)$ contient toute combinaison convexe de points de S, d'après l'assertion 1. Il suffit alors de vérifier que l'ensemble des combinaisons convexes de points de S est convexe, ce qui résulte facilement de la propriété d'associativité des barycentres. ∎

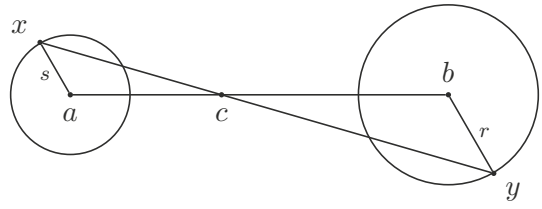
Comme exemple, soient $f : E \to \mathbb{R}$ une forme linéaire non nulle et $c \in \mathbb{R}$. L'hyperplan affine d'équation $f(x) = c$ est convexe ; le demi-espace défini par la condition $f(x) \geqslant c$ est convexe, car c'est $f^{-1}([c, +\infty[)$.
De même avec les conditions $f(x) \leqslant c$, $f(x) > c$ ou $f(x) < c$. Toute intersection de parties de ce type est donc convexe. On obtient d'autres exemples de convexes à l'aide de fonctions convexes, *cf.* la section 5.3.

Exercice 15.
Montrer que $S_n^{++}(\mathbb{R})$, ensemble des matrices réelles $n \times n$ symétriques définies positives est un ouvert convexe de l'espace $S_n(\mathbb{R})$ des matrices réelles $n \times n$ symétriques.
Solution. Soient $A, B \in S_n^{++}(\mathbb{R})$ et $\lambda \in [0, 1]$. Pour tout $X \in \mathbb{R}^n$ non nul, l'application $A \mapsto {}^tXAX$ de $S_n(\mathbb{R})$ dans $\mathbb{R}$ est linéaire, donc $\{A \in S_n(\mathbb{R}) \mid {}^tXAX > 0\}$ est un convexe de $S_n(\mathbb{R})$. Lorsque X varie, l'intersection de ces convexes est $S_n^{++}(\mathbb{R})$, qui est donc convexe.
Puisque $S_n(\mathbb{R})$ est de dimension finie, nous n'avons pas besoin de le munir d'une norme pour parler d'ouverts, etc (*cf.* la remarque de la page 442). D'après la proposition 49 de la page 134, une matrice $A \in S_n(\mathbb{R})$ est définie positive si, et seulement si, tous ses mineurs principaux A_k sont strictement positifs. Si $k \in [\![1, n]\!]$ est donné, l'application $A \mapsto A_k$ de $S_n(\mathbb{R})$ dans $\mathbb{R}$ est k-linéaire en les k premières colonnes de A, elle est donc continue (proposition 61 de la page 441). Il en résulte que $S_n^{++}(\mathbb{R})$ est ouvert dans $S_n(\mathbb{R})$, comme intersection (finie) des ouverts $\{A \in S_n(\mathbb{R}) \mid A_k > 0\}$ pour $k = 1, \ldots, n$.

La définition de la convexité est *géométrique*. Supposons désormais que E soit un e.v. n.. Les convexes de E ont alors des propriétés topologiques remarquables.

Lemme 63. Soient C une partie convexe de E et $a, b \in E$, $a \neq b$. Si a est adhérent à C et si b est intérieur à C, tout point c de $]a, b[$ est intérieur à C.

Démonstration. La preuve « se lit » sur la figure :



Soit f l'unique homothétie de centre c transformant b en a, elle est de rapport $t < 0$, en fait $t := -\alpha/(1-\alpha)$ si $c := (1-\alpha)a + \alpha b$. Comme f est un homéomorphisme de E sur E et $\mathrm{Int}(C)$ est un voisinage de b, $f\big(\mathrm{Int}(C)\big)$ est un voisinage de $f(b) = a$. Puisque $a \in \overline{C}$, ce voisinage rencontre C en un point $x := f(y)$, où $y \in \mathrm{Int}(C)$. De $x - c = t(y - c)$, on déduit $(t-1)(c-x) = t(y-x)$. L'homothétie g de centre x et de rapport $\lambda := t/(t-1)$ transforme ainsi y en c. Puisque C est un voisinage de y, $g(C)$ est un voisinage de $g(y) = c$. Mais $\lambda \in]0, 1[$ et C est convexe, donc $g(C) \subset C$, et ainsi C est un voisinage de c. Autrement dit, $c \in \mathrm{Int}(C)$. ∎

Proposition 64. Soit C une partie convexe de E.

1. L'intérieur et l'adhérence de C sont convexes.
2. Si $\mathrm{Int}(C)$ n'est pas vide, on a $\overline{\mathrm{Int}(C)} = \overline{C}$ et $\mathrm{Int}\big(\overline{C}\big) = \mathrm{Int}(C)$.

Démonstration. Si $u, v \in \mathrm{Int}(C)$, le lemme montre que $]u, v[\subset \mathrm{Int}(C)$, donc $\mathrm{Int}(C)$ est convexe. Soient $x, y \in \overline{C}$ et $\lambda \in [0,1]$. Il existe deux suites $(x_n), (y_n)$ de C convergeant vers x, y respectivement, et $\big((1-\lambda)x_n + \lambda y_n\big)$ converge vers $(1-\lambda)x + \lambda y$. Pour tout n, $(1-\lambda)x_n + \lambda y_n \in C$, donc $(1-\lambda)x + \lambda y \in \overline{C}$, ce qui montre que $\overline{C}$ est convexe.

De toutes façons, $\overline{\mathrm{Int}(C)} \subset \overline{C}$ puisque $\mathrm{Int}(C) \subset C$, et de même $\mathrm{Int}(C) \subset \mathrm{Int}(\overline{C})$ puisque $C \subset \overline{C}$. Supposons désormais que a soit un point intérieur à C. Soit $x \in \overline{C}$, montrons que $x \in \overline{\mathrm{Int}(C)}$. C'est vrai si $x := a$; sinon, x est adhérent à $]a, x[$, d'où la conclusion puisque $]a, x[\subset \mathrm{Int}(C)$ en vertu du lemme. Ainsi $\overline{\mathrm{Int}(C)} = \overline{C}$.

Soit enfin $y \in \mathrm{Int}\big(\overline{C}\big)$. Il existe une boule B de centre y contenue dans $\overline{C} = \overline{\mathrm{Int}(C)}$. Comme B est un voisinage du point y adhérent à $\mathrm{Int}(C)$, B rencontre $\mathrm{Int}(C)$ en un point x. Si $x := y$, on a $y \in \mathrm{Int}(C)$. Sinon, soit z le symétrique de x par rapport à y, de sorte que y est le milieu de $[x, z]$. Comme $x \in \mathrm{Int}(C)$ et $z \in B \subset \overline{C}$, le lemme montre que $]x, z[\subset \mathrm{Int}(C)$, d'où $y \in \mathrm{Int}(C)$. Ainsi $\mathrm{Int}\big(\overline{C}\big) \subset \mathrm{Int}(C)$, d'où $\mathrm{Int}\big(\overline{C}\big) = \mathrm{Int}(C)$. ∎

En dimension finie, la topologie des convexes *bornés* est assez simple. La proposition suivante, dont la preuve est esquissée, illustre ce fait.

Proposition 65. Soit C une partie convexe d'un e.v.n. E de dimension finie $n \geqslant 1$. L'intérieur de C est non vide si, et seulement si, l'espace *affine* E est engendré par C. Si c'est le cas, et si C est *bornée*, alors $\mathrm{Int}(C), \overline{C}$ et la frontière $\partial(C)$ de C sont homéomorphes respectivement à la boule unité ouverte, la boule unité fermée et la sphère unité de E.

Démonstration. Supposons $\mathrm{Int}(C)$ non vide. À une translation près, on peut supposer que C contient une boule de centre 0. Cette boule contenant une base de E, C engendre E comme espace vectoriel, donc comme espace affine. Inversement, si C engendre l'espace affine E, on peut supposer, à translation près, que C contient 0 et une base $(x_1, \dots, x_n)$ de E. Soit U l'ensemble (non vide) des $x := \sum_{i=1}^{n} \lambda_i x_i$, où $\lambda_i \in]0,1[$ pour tout i et $s := \sum_{i=1}^{n} \lambda_i < 1$. D'abord U est ouvert, comme intersection finie de demi-espaces ouverts. Pour tout x comme ci-dessus, $x := \sum_{i=0}^{n} \lambda_i x_i$, où $\lambda_0 := 1 - s \in]0,1[$ et $x_0 := 0$, donc $x \in C$ car C est convexe. Ainsi $C \supset U$, d'où $\mathrm{Int}(C) \neq \varnothing$.

On suppose désormais que C est bornée et d'intérieur non vide. En vertu de la proposition précédente, on peut supposer que C est aussi fermée, donc compacte. À une translation et une homothétie près, on peut supposer que C contient la boule unité $B := B(0,1)$. Posons $S := S(0,1)$. Comme $0 \notin \partial(C) = C \setminus \mathrm{Int}(C)$, on dispose de l'application continue $f : x \mapsto x/\|x\|$ de $\partial(C)$ dans S.

Soit $x \in S$. L'application $t \mapsto tx$ de $\mathbb{R}$ dans E est linéaire, donc $\{t \in \mathbb{R}_+ \mid tx \in C\}$ est une partie convexe, *i.e.* un intervalle I_x de $\mathbb{R}$. Puisque C contient B, on a $t \in I_x$ pour tout $t > 0$ assez petit. De plus I_x est fermé borné puisque C est fermée bornée. Ainsi $I_x = [0, k]$, avec $k > 0$. On voit aisément que $kx \in \partial(C) = C \setminus \mathrm{Int}(C)$ et que $tx \notin C$ si $t > k$. Le lemme 63 de la page ci-contre montre de plus que $tx \in \mathrm{Int}(C)$ pour tout $t \in [0, k[$. Il en résulte que $kx \in \partial(C)$ est l'unique

antécédent de x par f, donc f est bijective. Puisque $\partial(C)$ est compacte, f est un homéomorphisme de $\partial(C)$ sur S (théorème 52 de la page 436), donc $f^{-1} : S \mapsto \partial(C)$ est continue.

On prolonge f^{-1} « radialement », en définissant $g : \overline{B}(0,1) \to E$ par $g(x) := \|x\| f^{-1}(x/\|x\|)$ si $x \neq 0$ et $g(0) := 0$. La fonction g est continue, en 0 parce que C est bornée, et en dehors de 0 de façon évidente. Si $x \in \overline{B}(0,1)$ et $x \neq 0$, $f^{-1}(x/\|x\|) \in \partial(C) \subset C$, donc $g(x) \in C$ parce que C est convexe et $\|x\| \leqslant 1$. Le lecteur vérifiera aisément que g est une bijection de $\overline{B}(0,1)$ sur C. Cette bijection applique S sur $\partial(C)$, donc $B(0,1)$ sur $C \setminus \partial(C) = \text{Int}(C)$. Enfin g est un homéomorphisme parce que $\overline{B}(0,1)$ est compacte. ∎

5.2 Espaces connexes

Un espace métrique donné X est-il formé d'un ou de plusieurs morceaux ? Cette question est formulée de façon intuitive. En voici une traduction précise : nous savons que X et la partie vide sont des parties à la fois ouvertes et fermées de X. Y-a-t-il d'autres parties de X ouvertes et fermées ? La réponse est oui, en général. Cela conduit en fait à une définition :

Définition 27. Un espace métrique X est dit *connexe* si la partie vide et X sont les seules parties de X à la fois ouvertes et fermées.

Il ressort de cette définition que l'espace vide est connexe, de même que tout singleton. Par ailleurs, la connexité est une *propriété topologique* : si deux espaces sont homéomorphes et si l'un est connexe, l'autre l'est également (le vérifier). L'espace $\mathbb{R}^*$ n'est pas connexe, car $]0, +\infty[$ est ouvert dans $\mathbb{R}^*$, et son complémentaire $]-\infty, 0[$ dans $\mathbb{R}^*$ l'est aussi.
Notons plus généralement que, si (U_i) est une famille d'ouverts non vides et deux à deux disjoints recouvrant un espace X, chaque U_i est aussi fermé. En effet, pour tout i, U_i est le complémentaire de la réunion des U_j pour $j \neq i$, réunion qui est ouverte. Le lecteur pourra en déduire que chacune des propriétés ci-dessous caractérise la connexité d'un espace X :

1. Il n'existe pas de partition de X formée de deux ouverts non vides.
2. Il n'existe pas de partition de X formée d'au moins deux ouverts non vides.

Ainsi l'espace $\mathbb{R}\backslash\mathbb{Z}$ n'est pas connexe, car les intervalles ouverts $]n, n+1[$, n parcourant $\mathbb{Z}$, en forment une partition. En sens inverse, voici un exemple important d'espace connexe :

Théorème 66. Les parties connexes de $\mathbb{R}$ sont exactement les intervalles. En particulier $\mathbb{R}$ lui-même est connexe.

Démonstration. Soit X une partie connexe de $\mathbb{R}$ ayant au moins deux éléments, de sorte que les bornes (éventuellement infinies) $m := \inf X$ et $M := \sup X$ sont distinctes. Soit $c \in]m, M[$. Puisque c n'est pas un majorant de X, $]c, +\infty[\cap X \neq \varnothing$, de même $]-\infty, c[\cap X \neq \varnothing$. Si c n'appartenait pas à X, $]-\infty, c[\cap X$ et $]c, +\infty[\cap X$ seraient deux ouverts non vides et disjoints de X, ce qui contredirait la connexité de X. Ainsi X contient $]m, M[$, donc X est l'un des intervalles d'extrémités m, M.
Montrons inversement que tout intervalle I de $\mathbb{R}$ est connexe. Raisonnons par l'absurde, en supposant que $I = A \cup B$, où A, B sont deux ouverts de I, non vides et disjoints. Soit $f : I \to \mathbb{R}$ l'application valant 0 sur A et 1 sur B ; l'image de f est donc $\{0, 1\}$, de plus

$f^{-1}\big(\{0\}\big) = A$ et $f^{-1}\big(\{1\}\big) = B$. Il en résulte que f est continue car, si U est un ouvert de $\mathbb{R}$, $f^{-1}(U) = f^{-1}\big(\{0,1\} \cap U\big)$ vaut $\varnothing, A, B$ ou I, donc $f^{-1}(U)$ est un ouvert de I. La conclusion est alors aisée : l'image de l'intervalle I par la fonction continue f est un intervalle de $\mathbb{R}$ (*cf.* le cours de [L1]), c'est absurde car cette image est $\{0,1\}$. ∎

Le lecteur attentif pourra déduire de cette preuve le résultat suivant :

Proposition 67. Un espace métrique X est connexe si, et seulement si, toute fonction continue de X dans le sous-espace $\{0,1\}$ de $\mathbb{R}$ est constante.

Comme application, montrons que, si A est une partie connexe de l'espace X, son adhérence $\overline{A}$ est connexe. Soit $f : \overline{A} \to \{0,1\}$ une fonction continue. Puisque A est connexe, $f_{|A}$ est constante et égale à c, où $c \in \{0,1\}$. Alors $f^{-1}\big(\{c\}\big)$ est un fermé de $\overline{A}$ (donc de X) contenant A, donc $f^{-1}\big(\{c\}\big) = \overline{A}$. Ainsi f est constante et égale à c.

> ***Remarque.*** Soient A et B deux parties de l'espace X. On suppose que B est connexe et rencontre l'intérieur de A et l'intérieur de $X \setminus A$. Alors B rencontre la frontière ∂A de A. En effet, $V_1 := B \cap \mathrm{Int}(A)$ et $V_2 := B \cap \mathrm{Int}\big(X \setminus A\big)$ sont deux ouverts de B disjoints, non vides par hypothèse. Puisque B est connexe, $V_1 \cup V_2 \neq B$. Or, par définition de ∂A (définition 11 de la page 407), $\mathrm{Int}(A)$, $\mathrm{Int}\big(X \setminus A\big)$ et ∂A forment une partition de X, et il en résulte que B rencontre ∂A. On donne le nom de *passage des frontières* à ce résultat.

Proposition 68. L'image d'un espace connexe par une application continue est connexe.

Démonstration. Soient X, Y deux espaces métriques et $f : X \to Y$ une application continue. En supposant X connexe, il s'agit de prouver que $f(X)$ l'est aussi. L'application $x \mapsto f(x)$ de X dans $f(X)$ est surjective et continue, ce qui nous ramène au cas où f est surjective, ce que nous supposons. Si V est une partie ouverte et fermée de Y, $f^{-1}(V)$ est ouverte et fermée dans X. Puisque X est connexe, deux cas sont possibles. Ou bien $f^{-1}(V) = \varnothing$, auquel cas $V = \varnothing$, à cause de la surjectivité de f, ou bien $f^{-1}(V) = X$, auquel cas $V = Y$ pour la même raison. L'espace Y est donc connexe. ∎

Ainsi le cercle unité de $\mathbb{R}^2$ est connexe, comme image de $[0, 2\pi]$ (qui est connexe) par l'application continue $t \mapsto (\cos t, \sin t)$. De même, si I est un intervalle de $\mathbb{R}$ et $f : I \to \mathbb{R}$ est continue, le graphe de f est une partie connexe de $\mathbb{R}^2$, car c'est l'image de I par l'application $x \mapsto \big(x, f(x)\big)$, qui est continue. Par exemple, une sinusoïde ou une parabole sont des parties connexes du plan.
Voici une généralisation du théorème des valeurs intermédiaires :

Corollaire 69. Soient X un espace connexe et $f : X \to \mathbb{R}$ une fonction continue. Si f prend les valeurs a et b, $a < b$, elle prend toute valeur $c \in]a, b[$.

Démonstration. D'après la proposition précédente, $J := f(X)$ est une partie connexe

de $\mathbb{R}$, c'est-à-dire un intervalle (théorème 66 de la page 448). Mais $a, b \in J$, donc J contient $]a, b[$. ∎

Exercice 16.

Pourquoi l'hyperbole d'équation $y = 1/x$ n'est-elle pas connexe ?

Solution. Soit $H := \{(x, 1/x) \mid x \in \mathbb{R}^*\}$ cette hyperbole. Elle a deux branches, à savoir $H_+ := \big\{(x, 1/x) \mid x \in]0, +\infty[\big\}$ et $H_- := \big\{(x, 1/x) \mid x \in]-\infty, 0[\big\}$. Puisque H_+ est l'intersection de H avec le demi-plan ouvert défini par $x > 0$, H_+ est un ouvert de H, et de même pour H_-. Ainsi H n'est pas connexe. Par contre H_+ et H_- sont connexes (appliquer la proposition précédente). Autre solution : l'image de H par la fonction continue $(x, y) \mapsto x$ est $\mathbb{R}^*$, qui n'est pas connexe, donc H n'est pas connexe.

Exercice 17.

Soit $(Y_i)_{i\in I}$ une famille de parties connexes de l'espace X. Si $j \in I$ est tel que $Y_j \cap Y_i \neq \varnothing$ pour tout $i \in I$, montrer que $Y := \bigcup_{i\in I} Y_i$ est connexe.

Solution. Soit $f : Y \to \{0, 1\}$ une application continue. Il s'agit de prouver que f est constante. Pour tout $i \in I$, la restriction de f à Y_i est constante, car Y_i est connexe. De plus, puisque $Y_j \cap Y_i \neq \varnothing$, f prend même valeur sur Y_j et sur Y_i. Ainsi f est constante.

Composantes connexes

Soit X un espace métrique. Définissons sur X une relation $\mathcal{R}$ ainsi : $a \mathcal{R} b$ si a et b appartiennent à une même partie connexe de X. Cette relation est évidemment symétrique ; elle est réflexive car tout $x \in X$ appartient à $\{x\}$, qui est connexe. Si $a \mathcal{R} b$ et $b \mathcal{R} c$, il existe deux parties connexes Y, Z de X telles que $a, b \in Y$ et $b, c \in Z$. Puisque $b \in Y \cap Z$, l'exercice précédent montre que $Y \cup Z$ est connexe. D'où $a \mathcal{R} c$, car $a, c \in Y \cup Z$.

Ainsi $\mathcal{R}$ est une relation d'équivalence, dont les classes sont appelées *composantes connexes* de X. Ces classes forment une partition de X, ce sont les « morceaux » dont nous avons parlé au début de cette section. Soient $a \in X$ et C la composante connexe de X contenant a, *i.e.* la classe de a modulo $\mathcal{R}$. D'abord C est connexe. En effet, pour tout $b \in C$, il existe une partie connexe S_b de X contenant a et b, et $S_b \subset C$, par définition de $\mathcal{R}$. Ainsi C est la réunion des S_b ($b \in C$), qui sont des parties connexes de X contenant a, donc C est connexe, d'après l'exercice précédent. En fait il est clair que C est, au sens de l'inclusion, la plus grande partie connexe de X contenant a. Comme $\overline{C}$ est connexe, on a donc $\overline{C} = C$, *i.e.* les composantes connexes de X sont *fermées*. Noter que X est connexe si, et seulement si, sa seule composante connexe est X.

À titre d'exemple, $\mathbb{R}^*$ a deux composantes connexes, à savoir $]-\infty, 0[$ et $]0, +\infty[$. En particulier, $\mathbb{R}^*$ n'est pas homéomorphe à $\mathbb{R}$, qui est connexe. Les composantes connexes de $\mathbb{R} \setminus \mathbb{Z}$ sont les intervalles $]n, n + 1[$, $n \in \mathbb{Z}$.

Pour démontrer qu'un espace métrique est connexe, il est parfois commode de vérifier que cet espace a une propriété plus forte que la connexité, à savoir la *connexité par arcs*.

Définition 28. Un espace métrique X est dit *connexe par arcs* si, pour deux points arbitraires a, b de X, il existe un arc (ou chemin continu) joignant a à b, c'est-à-dire une application continue $\gamma : [0, 1] \to X$ telle que $\gamma(0) = a$ et $\gamma(1) = b$.

Proposition 70. Tout espace métrique connexe par arcs est connexe.

Démonstration. Soit X un espace connexe par arcs.

Si $a, b \in X$, soit $\gamma : [0,1] \to X$ un chemin continu joignant a à b. Puisque $[0,1]$ est connexe, la proposition 68 de la page 449 montre que $\gamma\big([0,1]\big)$ est connexe. Puisque a et b appartiennent à $\gamma\big([0,1]\big)$, ils ont même composante connexe.

Cela est vrai pour tous $a, b \in X$, donc X est connexe. ∎

Proposition 71. Soit X une partie d'un e.v. n. E étoilée par rapport à l'un de ses points. Alors X est connexe par arcs. Toute partie *convexe* de E est donc connexe par arcs.

Démonstration. Supposons que X soit étoilée par rapport au point $a \in X$. Si $b \in X$, $t \mapsto (1-t)a + tb$ est un chemin continu joignant a à b $\big($on paramètre le segment $[a,b]\big)$. Soient maintenant $b, c \in X$. Il existe donc dans X un chemin continu β' joignant a à b et un chemin continu γ joignant a à c. Le chemin continu $\beta : t \mapsto \beta'(1-t)$ de X joint b à a. Définissons $\delta : [0,1] \to X$ par $\delta(t) := \beta(2t)$ si $t \in [0,1/2]$ et $\delta(t) := \gamma(2t-1)$ si $t \in \,]1/2, 1]$. Puisque $\beta(1) = a = \gamma(0)$, la fonction δ est continue, car ses restrictions à $[0,1/2]$ et $[1/2,1]$ le sont. Ainsi δ est un chemin continu de X joignant b à c (appelé *composé* de β et γ), ce qui prouve que X est connexe par arcs. La deuxième assertion est évidente, puisqu'une partie convexe est étoilée par rapport à chacun de ses points. ∎

Ainsi, dans un e.v. n. E, tout sous-espace affine est connexe par arcs, et toute boule (ouverte ou fermée) est connexe par arcs. Pour les sphères, c'est plus subtil :

Proposition 72. Soit E un e.v. n. réel, de dimension (finie ou non) au moins égale à 2. Toute sphère de E est connexe par arcs et *a fortiori* connexe.

Démonstration. Il suffit (pourquoi ?) de montrer que la sphère unité S de E est connexe par arcs. Soient $a, b \in S$. Comment définir un chemin continu $\gamma : [0,1] \to S$ joignant a à b ? C'est évident si a et b ne sont pas colinéaires. En effet, dans ce cas, $(1-t)a + tb \neq 0$ pour tout $t \in [0,1]$, et il suffit de poser $\gamma(t) := \big\|(1-t)a + tb\big\|^{-1}\big((1-t)a + tb\big)$. Supposons que a, b soient colinéaires, *i.e.* appartiennent à une même droite vectorielle D. Puisque $\dim E \geqslant 2$, il existe $x \in E \setminus D$, et alors $c := x/\|x\| \in S$ n'appartient pas à D. Comme nous venons de le montrer, il existe alors un chemin continu γ de S joignant a à c, et un autre δ joignant c à b. Comme ci-dessus, on en déduit par « composition » un chemin continu de S joignant a à b. ∎

C'est faux en dimension 1 : la sphère unité de $\mathbb{R}$ est $\{-1, 1\}$, elle n'est pas connexe. Si E est un e.v. n. *complexe* non nul, les sphères de E sont connexes par arcs (pourquoi ?).

Proposition 73. L'image par une application continue d'un espace connexe par arcs est connexe par arcs.

Démonstration. Soient X, Y deux espaces métriques, $f : X \to Y$ une application continue et $c, d \in f(X)$. Il existe $a, b \in X$ tels que $c = f(a)$ et $d = f(b)$. Si X est connexe par arcs, il existe un chemin continu $\gamma : [0,1] \to X$ joignant a à b. Alors $f \circ \gamma : [0,1] \to Y$ est un chemin continu joignant c à d. La conclusion est maintenant immédiate. ∎

Composantes connexes par arcs

Soit X un espace métrique. Définissons sur X une relation $\mathcal{R}$ ainsi : $a \mathcal{R} b$ si, et seulement s'il existe un chemin continu de X joignant a à b. Cette relation est évidemment réflexive car tout $x \in X$ appartient à $\{x\}$, qui est connexe par arcs. Elle est symétrique car, si γ est un chemin continu de a à b, $t \mapsto \gamma(1-t)$ est un chemin continu de b à a. Si $a \mathcal{R} b$ et $b \mathcal{R} c$, il existe un chemin continu de a à b et un chemin continu de b à c. En composant ces chemins comme ci-dessus, on obtient un chemin continu de a à c. Ainsi $\mathcal{R}$ est une relation d'équivalence, dont les classes sont appelées *composantes connexes par arcs* de X. On vérifie immédiatement que, si $a \in X$, la composante connexe par arcs de a est, au sens de l'inclusion, la plus grande partie connexe par arcs de X contenant a.

Les exemples d'espaces connexes mais pas connexes par arcs sont « pathologiques », *cf.* l'exercice II.1.41. Dans cas particulier important, ces deux notions coïncident :

Proposition 74. Pour un *ouvert* U d'un e.v.n. E, les notions de « connexe » et de « connexe par arcs » coïncident.

Démonstration. Supposons U connexe, et montrons qu'il est connexe par arcs (cela suffit, d'après la proposition 70 de la page précédente). Soient $a \in U$ et C la composante connexe par arcs de a dans U. Puisque U est ouvert, il existe $r > 0$ tel que $B(a,r) \subset U$. Comme $B(a,r)$ est connexe par arcs, on a $B(a,r) \subset C$, donc C est un ouvert de U. Ainsi les composantes connexes par arcs de U sont des ouverts. Ces composantes formant une partition de U, chacune d'elles est fermée dans U, comme complémentaire de la réunion des autres. Chaque composante connexe par arcs C de U est donc ouverte et fermée dans U, donc $C = U$ puisque U est connexe et C n'est pas vide. Il y a donc une seule composante connexe par arcs de U, d'où la conclusion. ∎

5.3 Fonctions convexes

Dans cette section, E désigne un espace vectoriel réel*, de dimension finie ou non, C est une partie* convexe *non vide de E, et l'on se donne une fonction $f : C \to \mathbb{R}$.*

Définition 29. La fonction f est dite *convexe* si, pour tous $x, y \in C$ et $\lambda \in [0,1]$, on a :

$$f\big((1-\lambda)x + \lambda y\big) \leqslant (1-\lambda)f(x) + \lambda f(y). \tag{30}$$

Si cette inégalité est stricte dès que $x \neq y$ et $\lambda \neq 0, 1$, f est dite *strictement convexe*. Si ces inégalités sont dans l'autre sens, f est dite *concave* (resp. strictement concave).

Ainsi f est concave (ou strictement concave) si, et seulement si, $-f$ est convexe (ou strictement convexe), ce qui permet de se limiter aux fonctions convexes. Toute forme affine (application affine de E dans $\mathbb{R}$) est à la fois convexe et concave. La réciproque constitue un bon exercice pour le lecteur.

Autre exemple : *toute norme* $\|\cdot\| : E \to \mathbb{R}$ *est convexe*. En effet, si $x, y \in E$ et $\lambda \in [0,1]$, on a :

$$\|(1-\lambda)x + \lambda y\| \leqslant \|(1-\lambda)x\| + \|\lambda y\| = |1-\lambda|\,\|x\| + |\lambda|\,\|y\| = (1-\lambda)\|x\| + \lambda\|y\|.$$

Il y a des liens entre parties convexes et fonctions convexes. Par exemple, si f est convexe et si $k \in \mathbb{R}$, $\big\{x \in C \mid f(x) \leqslant k\big\}$ et $\big\{x \in C \mid f(x) < k\big\}$ sont des parties convexes de E (le vérifier). En appliquant cette remarque à la norme, on en déduit ceci : *toute boule d'un espace vectoriel normé est* convexe*, donc connexe par arcs.*

Appelons *épigraphe* de f l'ensemble $\Gamma := \big\{(x,t) \in C \times \mathbb{R} \mid t \geqslant f(x)\big\}$. Alors f *est convexe si, et seulement si,* Γ *est une partie convexe de* $E \times \mathbb{R}$. En effet, x, y et λ étant donnés, posons : $A := \big(x, f(x)\big)$, $B := \big(y, f(y)\big)$, $M := \Big((1-\lambda)x + \lambda y, f\big((1-\lambda)x + \lambda y\big)\Big)$ et $P := (1-\lambda)A + \lambda B$. Si Γ est convexe, $P \in \Gamma$ (car $A, B \in \Gamma$), ce qui équivaut à l'inégalité (30). Inversement, si f est convexe et si $A' := (x,t), B' := (y,u) \in \Gamma$ et $\lambda \in [0,1]$, $(1-\lambda)t + \lambda u \geqslant (1-\lambda)f(x) + \lambda f(y) \geqslant f\big((1-\lambda)x + \lambda y\big)$, ce qui montre que $(1-\lambda)A' + \lambda B' \in \Gamma$, d'où la convexité de Γ.

Lorsque $E := \mathbb{R}$ et $C := I$ est un intervalle, soit $\mathcal{C}$ le graphe de f. Graphiquement, la convexité de f signifie que le point P est toujours au-dessus (au sens large) du point M de $\mathcal{C}$ correspondant, ou encore que, entre les points A et B, la courbe $\mathcal{C}$ est au-dessous de la corde AB. On dit aussi que $\mathcal{C}$ tourne sa *concavité* vers le haut si f est convexe, et vers le bas si f est concave. Retenons donc ceci, dans le cas où $E := \mathbb{R}$:

PROPRIÉTÉ. *Une fonction* $f : I \to \mathbb{R}$ *est convexe si, et seulement si, sa courbe représentative est au-dessous de chacune de ses cordes.*

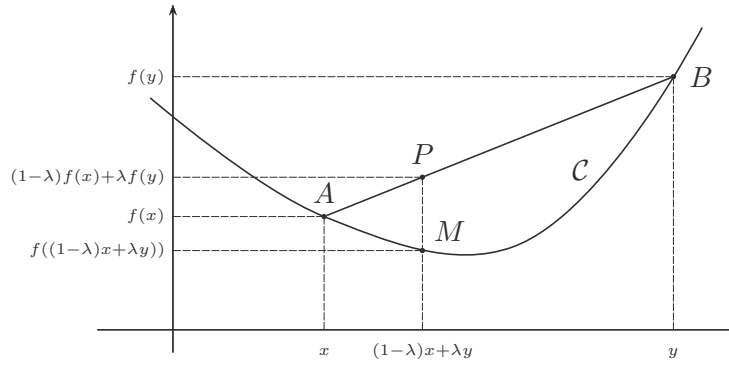


FIGURE II.1.4 – La courbe est au-dessous de ses cordes : f est convexe.

Supposons que $x, y \in I$, $x < y$ et $\lambda \in \,]0,1[$. Le point M est au-dessous de la droite (AB) si, et seulement si, $\mathrm{pente}(AM) \leqslant \mathrm{pente}(AB)$, soit $\mathrm{pente}(AB) \leqslant \mathrm{pente}(BM)$, c'est équivalent. Puisque $\lambda \mapsto z := (1-\lambda)x + \lambda y$ est une bijection de $]0,1[$ sur $]x,y[$, chacune des inégalités ci-dessous, si elle vaut pour tous $x, z, y \in I$ tels que $x < z < y$, *caractérise* la convexité de f :

$$\frac{f(z) - f(x)}{z - x} \;\leqslant\; \frac{f(y) - f(x)}{y - x} \;\leqslant\; \frac{f(y) - f(z)}{y - z}. \tag{31}$$

L'exercice suivant montre comment utiliser ces inégalités.

Exercice 18.
Montrer que si une fonction convexe $f : \mathbb{R} \to \mathbb{R}$ n'est pas constante, on a $\lim\limits_{+\infty} f = +\infty$ ou $\lim\limits_{-\infty} f = +\infty$. En particulier, f n'est pas majorée.

Solution. Elle se devine sur la figure précédente. La fonction $g : x \mapsto f(-x)$ est convexe, et son graphe se déduit du graphe $\mathcal{C}$ de f par symétrie par rapport à Oy. Puisque f n'est pas constante, il existe $A, M \in \mathcal{C}$ tels que $f(A) \neq f(M)$, soit $m := \text{pente}(AM) \neq 0$. Quitte à remplacer f par g, ce qui change le signe de m, on peut supposer $m > 0$. Notons x, z les abscisses respectives de A, M, avec par exemple $x < z$. Pour tout $y > z$, les inégalités (31) donnent : $f(y) \geqslant m(y - z) + f(z)$. D'où évidemment $\lim\limits_{y\to+\infty} f(y) = +\infty$.

Voici maintenant deux résultats généraux très simples faisant déjà apparaître la puissance de la notion de convexité. La section suivante montrera l'intérêt du premier de ces résultats.

Proposition 75. Soient $x_1, \dots, x_n \in C$ et $\lambda_1, \dots, \lambda_n \in [0, 1]$ des réels de somme 1. Si f est *convexe*, on a l'inégalité :

$$f\big(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n\big) \leqslant \lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_n f(x_n).$$

En particulier :

$$f\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right) \leqslant \frac{1}{n}\big[f(x_1) + f(x_2) + \cdots + f(x_n)\big]. \qquad (32)$$

Démonstration. Pour tout i, le point $M_i := \big(x_i, f(x_i)\big)$ appartient à l'épigraphe Γ de f. Puisque Γ est convexe, le barycentre de $(M_1, \lambda_1), \dots, (M_n, \lambda_n)$ appartient à Γ, d'où la conclusion, puisque ce barycentre est : $\big(\lambda_1 x_1 + \cdots + \lambda_n x_n, \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n)\big)$. La deuxième inégalité est obtenue en prenant $\lambda_i := 1/n$ pour tout i. ∎

Théorème 76. Supposons que E soit un e.v. n. et que $f : C \to \mathbb{R}$ soit convexe.

1. Si f présente en un point a un minimum *relatif*, c'est en fait un minimum *absolu*.
2. Si f est strictement convexe, elle présente un minimum en un point *au plus*, et il s'agit d'un minimum (absolu) *strict*.

Démonstration. Supposons que f présente en $a \in C$ un minimum relatif. Cela signifie qu'il existe un $r > 0$ tel que $f(x) \geqslant f(a)$ pour tout $x \in C \cap B(a, r)$. Soit $x \in C$, $x \neq a$. Choisissons $\lambda \in]0, 1[$ tel que $\lambda\|x-a\| \leqslant r$, et soit $y := (1-\lambda)a+\lambda x$, d'où $y - a = \lambda(x - a)$. Alors $d(a, y) = \lambda d(a, x) \leqslant r$, et $y \in C$ puisque C est convexe, d'où $f(a) \leqslant f(y)$. Or $f(y) \leqslant (1 - \lambda)f(a) + \lambda f(x)$ (convexité de f), donc $f(a) \leqslant (1 - \lambda)f(a) + \lambda f(x)$, soit $\lambda f(a) \leqslant \lambda f(x)$. Ainsi $f(a) \leqslant f(x)$, d'où l'assertion 1. Si f est strictement convexe, on a $f(y) < (1 - \lambda)f(a) + \lambda f(x)$, d'où l'on déduit $f(a) < f(x)$; f présente donc en a un minimum absolu strict, et l'unicité de a est claire. ∎

Limitons-nous désormais aux fonctions d'une variable réelle $f : I \to \mathbb{R}$, où I est un intervalle non trivial (*i.e.* ayant au moins deux points). La convexité de f est une propriété très forte, impliquant la continuité et (presque) la dérivabilité de f. Plus précisément :

Théorème 77. Supposons que I soit *ouvert*. Si la fonction $f : I \to \mathbb{R}$ est convexe, elle est continue et possède en chaque point x 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 :

$$f'_g(x) \leqslant f'_d(x) \leqslant \frac{f(y) - f(x)}{y - x} \leqslant f'_g(y) \leqslant f'_d(y). \tag{33}$$

Démonstration. Posons pour simplifier $p(x, y) = \big(f(y) - f(x)\big)/(y - x)$ dès que $x, y \in I$ sont distincts. Écrivons $I :=]u, v[$, et soit $a \in I$. Les inégalités (31) montrent que la fonction $\varphi : x \mapsto p(x, a)$ est croissante sur $]u, a[$ (prendre $y := a$) ainsi que sur $]a, v[$ (prendre $x := a$). De plus $(x < a < y) \Longrightarrow \big(p(a, x) \leqslant p(a, y)\big)$ (prendre $z := a$). Il en résulte que la fonction φ est croissante et majorée sur $]u, a[$, elle a donc en a^- une limite finie qui est, par définition, la dérivée à gauche de f en a. Cette limite est aussi la borne supérieure de φ sur $]u, a[$, ce qui établit la troisième égalité (33) (aux lettres près).
De même, la fonction φ étant croissante et minorée sur $]a, v[$, elle a en a^+ une limite finie qui est, par définition, la dérivée à droite de f en a. Cette limite est aussi la borne *inférieure* de φ sur $]a, v[$, et l'on en déduit la deuxième égalité (33). L'implication ci-dessus donne en outre l'inégalité $f'_g(a) \leqslant f'_d(a)$, d'où la première et la dernière inégalité (33). Enfin, puisque $f'_g(a)$ et $f'_d(a)$ existent, f est continue en a. ∎

Le fait que I soit ouvert est important. Soit par exemple $f : [0, 1] \to \mathbb{R}$ la fonction valant 1 en 0 et 0 ailleurs. Il est clair que f est convexe, mais elle n'est pas continue (à droite) en 0. De même, une fonction convexe sur un intervalle ouvert n'est pas forcément partout dérivable : $x \mapsto |x|$ est convexe sur $\mathbb{R}$ mais n'est pas dérivable en 0.

Corollaire 78. Si I est ouvert et si $f : I \to \mathbb{R}$ est convexe, sa courbe représentative $\mathcal{C}$ est entièrement située au-dessus de chacune de ses tangentes (à gauche ou à droite).

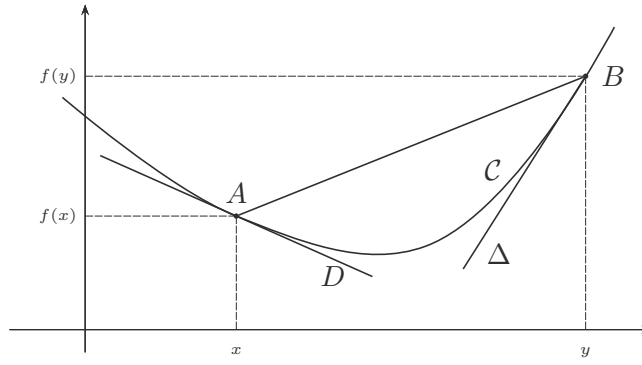


FIGURE II.1.5 – La courbe est au-dessus de chacune de ses tangentes.

Démonstration. Soient $x, y \in I$, $x < y$. Notons $A := \big(x, f(x)\big)$ et $B := \big(y, f(y)\big)$ les points de $\mathcal{C}$ correspondants. Soit D la tangente à droite à $\mathcal{C}$ en A, *i.e.* la droite

de pente $f'_d(x)$ passant par A. D'après le théorème précédent, $f'_d(x) \leqslant \text{pente}(AB)$, donc B est au-dessus de D. En faisant varier y, on voit que la partie de $\mathcal{C}$ à droite de A est entièrement au-dessus de D. Soit de même Δ la tangente à droite à $\mathcal{C}$ en B. Ici, $\text{pente}(AB) \leqslant f'_d(y)$, donc A est au-dessus de Δ. En faisant varier x, on voit que la partie de $\mathcal{C}$ à gauche de B est entièrement au-dessus de Δ. En faisant varier x et y, on obtient la conclusion annoncée pour les tangentes à droite, et l'argument vaut aussi pour les tangentes à gauche. ∎

Exercice 19.

Supposons que I soit ouvert. Montrer que $f : I \to \mathbb{R}$ est convexe si, et seulement si, elle a la propriété suivante : pour tout $x_0 \in I$, il existe une droite D passant par le point $M_0 := \big(x_0, f(x_0)\big) \in \mathcal{C}$ et telle que $\mathcal{C}$ soit entièrement située au-dessus de D.

Solution. Si f est convexe, le corollaire précédent montre que la tangente à droite à $\mathcal{C}$ en M_0 répond à la question. Supposons inversement que f vérifie la propriété de l'énoncé mais ne soit pas convexe. Il existe donc $x, y \in I$, $x < y$, et $\lambda \in]0,1[$ tels que, dans la figure II.1.4, le point M soit strictement au-dessus de la droite (AB). Autrement dit, $\text{pente}(MB) < \text{pente}(AB) < \text{pente}(AM)$. Par hypothèse, il existe une droite D passant par M et telle que $\mathcal{C}$ soit entièrement au-dessus de D. Puisque $A \in \mathcal{C}$ est à gauche de M, on a donc $\text{pente}(AM) \leqslant \text{pente}(D)$. Puisque $B \in \mathcal{C}$ est à droite de M, on a de même $\text{pente}(MB) \geqslant \text{pente}(D)$. D'où $\text{pente}(MB) \geqslant \text{pente}(AM)$. Nous avons obtenu une contradiction, ce qui montre que f est convexe.

Voici une caractérisation importante de la convexité d'une fonction dérivable :

Théorème 79. Si I est *ouvert*, une fonction dérivable $f : I \to \mathbb{R}$ est convexe (resp. strictement convexe) si, et seulement si, f' est croissante (resp. strictement croissante).

Démonstration. Si f est convexe, f' est croissante, vu le théorème précédent. Supposons inversement que f' soit croissante. Soient $x, y \in I$, $x < y$, et $\lambda \in]0,1[$. Reprenons les notations de la figure II.1.4 de la page 453 et posons $z := (1-\lambda)x + \lambda y$. D'après le théorème des accroissements finis, il existe $a \in]x, z[$ et $b \in]z, y[$ tels que $\text{pente}(AM) = f'(a)$ et $\text{pente}(MB) = f'(b)$. D'où $\text{pente}(AM) \leqslant \text{pente}(MB) = f'(b)$, ce qui montre que M est au-dessous de la corde AB. Il en résulte que f est convexe.

Si f est strictement convexe, les inégalités (31) sont toujours strictes. En examinant la figure II.1.4 de la page 453, on voit que $f'x) \leqslant \text{pente}(AM) < \text{pente}(AB) \leqslant f'(y)$, donc f' est strictement croissante. La réciproque est facile et laissée au lecteur. ∎

Une fonction dérivable sur un intervalle est croissante si, et seulement si, sa dérivée est positive. On en déduit un critère de convexité, c'est celui qu'on applique le plus souvent :

Théorème 80. Une fonction $f : I \to \mathbb{R}$ deux fois dérivable est convexe si, et seulement si, $f'' \geqslant 0$ sur I. L'intervalle I est ici quelconque.

Démonstration. Si I est ouvert, c'est une conséquence évidente du théorème précédent. Supposons par exemple que I soit de la forme $[a, b[$ ou $]a, b]$, avec $b \in \mathbb{R}$. Po-

sons $J :=]a, b[$. Puisque f est deux fois dérivable, f et f' sont continues sur I (à gauche en b). Supposons que f soit convexe sur I. Nous savons déjà que, sur J, f' est croissante et $f'' \geqslant 0$. Puisque f' est continue à gauche en b, elle est croissante sur $]a, b]$, et donc $f''(b) \geqslant 0$ (il s'agit de la dérivée à gauche de f' en b).

Inversement, supposons $f'' \geqslant 0$ sur I. Alors f est déjà convexe sur J. Dans l'inégalité (30), fixons $x \in J$ et faisons tendre y vers b^-. L'inégalité se conserve à la limite, puisque f est continue à gauche en b. Cela prouve que f est convexe sur $]a, b]$. On raisonne de même en a, le cas échéant. ∎

De même, f est strictement convexe dans I si, et seulement si, $f'' \geqslant 0$ et si l'ensemble des $x \in I$ tels que $f''(x) > 0$ est *dense* dans I (penser à $x \mapsto x^4$ sur $\mathbb{R}$).

Exemples.

1. La fonction exponentielle est strictement convexe sur $\mathbb{R}$, car elle est sa propre dérivée seconde et elle est strictement positive. La fonction $\ln$ est strictement concave sur $]0, +\infty[$, car sa dérivée seconde est $x \mapsto -1/x^2$. La fonction $-\ln$ est donc strictement convexe. La fonction sinus est strictement concave sur $]0, \pi[$ et strictement convexe sur $]-\pi, 0[$.
2. Soit $a \in \mathbb{R}$, $a \neq 0, 1$. la fonction puissance $x \mapsto x^a$ est indéfiniment dérivable sur $]0, +\infty[$, et sa dérivée seconde est $x \mapsto a(a-1)x^{a-2}$. Cette fonction est donc strictement convexe si $a > 1$ ou $a < 0$, et elle est strictement concave si $0 < a < 1$ (les cas $a = 0$ et $a = 1$ sont évidents).

5.4 Inégalités de convexité

En utilisant convenablement la convexité de fonctions simples, on peut établir de nombreuses inégalités (dites inégalités de convexité). Voici par exemple une conséquence de la concavité de la fonction $\ln$ sur $]0, +\infty[$:

Proposition 81. Soient $x_1, \ldots, x_n$ des réels strictement positifs. Pour tous réels positifs $\lambda_1, \ldots, \lambda_n$ de somme 1, on a :

$$x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_n^{\lambda_n} \leqslant \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n. \tag{34}$$

En particulier, on a l'inégalité *arithmético-géométrique* suivante :

$$\sqrt[n]{x_1 x_2 \cdots x_n} \leqslant \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{35}$$

Démonstration. Puisque la fonction $-\ln$ est convexe sur $]0, +\infty[$, on a :

$$-\ln(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n) \leqslant -\big[\lambda_1 \ln(x_1) + \lambda_2 \ln(x_2) + \cdots + \lambda_n \ln(x_n)\big]$$

en vertu de la proposition 75 de la page 454. En changeant les signes et en prenant les exponentielles des deux membres (l'exponentielle est croissante), on obtient l'inégalité (34). En prenant $\lambda_i := 1/n$ pour tout i, on obtient l'inégalité (35).

Le nombre $(x_1 + \cdots + x_n)/n$ est appelé *moyenne arithmétique* des x_i et $\sqrt[n]{x_1 \cdots x_n}$ est appelé *moyenne géométrique* des x_i. Cela explique le nom donné à cette inégalité. ∎

Construction de normes sur K^n.
On se donne un entier $n \geqslant 1$ et un réel $p > 0$. Si $\mathbf{x} := (x_1, \dots, x_n) \in K^n$, posons :

$$\|\mathbf{x}\|_p := \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}. \tag{36}$$

Alors $\|\cdot\|_p$ est une norme sur K^n dès que $p \geqslant 1$. Seule l'inégalité triangulaire est non triviale, elle sera prouvée plus loin. Commençons par établir une inégalité importante, qui généralise l'inégalité de Cauchy-Schwarz (obtenue lorsque $p := 2$) :

Théorème 82. Soient p, q deux réels strictement positifs tels que $\dfrac{1}{p} + \dfrac{1}{q} = 1$. Pour tous $\mathbf{a} := (a_1, \dots, a_n)$, $\mathbf{b} := (b_1, \dots, b_n) \in \mathbb{C}^n$, on a l'inégalité *de Hölder* suivante :

$$\left|\sum_{j=1}^{n} a_j b_j\right| \leqslant \left(\sum_{j=1}^{n} |a_j|^p\right)^{1/p} \left(\sum_{j=1}^{n} |b_j|^q\right)^{1/q}. \tag{37}$$

Démonstration. On peut supposer $\mathbf{a}, \mathbf{b}$ non nuls. Dans le second membre de cette inégalité, le premier facteur vaut $\|\mathbf{a}\|_p$ et le second vaut $\|\mathbf{b}\|_q$, avec la notation ci-dessus. Considérons deux réels $t, u > 0$. Appliquons la proposition précédente avec $n := 2$, $x_1 := t^p$, $x_2 = u^q$, $\lambda_1 := 1/p$ et $\lambda_2 := 1/q$. On obtient l'inégalité $tu \leqslant \dfrac{t^p}{p} + \dfrac{u^q}{q}$, encore vraie si $tu := 0$.
Soit maintenant $j \in [\![1, n]\!]$. En prenant $t := |a_j| / \|\mathbf{a}\|_p$ et $u := |b_j| / \|\mathbf{b}\|_q$, il vient :

$$\frac{|a_j b_j|}{\|\mathbf{a}\|_p \|\mathbf{b}\|_q} \leqslant \frac{|a_j|^p}{p\|\mathbf{a}\|_p^p} + \frac{|b_j|^q}{q\|\mathbf{b}\|_q^q}.$$

Sommons pour $j = 1, \dots, n$. Par définition de $\|a\|_p$ et $\|b\|_q$, il vient :

$$\sum_{j=1}^{n} \frac{|a_j b_j|}{\|\mathbf{a}\|_p \|\mathbf{b}\|_q} \leqslant \sum_{j=1}^{n} \frac{|a_j|^p}{p\|\mathbf{a}\|_p^p} + \sum_{j=1}^{n} \frac{|b_j|^q}{q\|\mathbf{b}\|_q^q} = \frac{1}{p} + \frac{1}{q} = 1.$$

L'inégalité de Hölder en résulte, puisque $\left|\sum_j a_j b_j\right| \leqslant \sum_j |a_j|\,|b_j|$. ∎

Nous pouvons en déduire que $\|\cdot\|_p$ est une norme lorsque $p \geqslant 1$:

Théorème 83. Considérons un réel $p \geqslant 1$. Pour tous vecteurs $\mathbf{x} := (x_1, \dots, x_n)$ et $\mathbf{y} := (y_1, \dots, y_n)$ de $\mathbb{C}^n$, on a l'inégalité *de Minkowski* suivante :

$$\left(\sum_{j=1}^{n} |x_j + y_j|^p\right)^{1/p} \leqslant \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p} + \left(\sum_{j=1}^{n} |y_j|^p\right)^{1/p}, \tag{38}$$

c'est-à-dire $\|\mathbf{x} + \mathbf{y}\|_p \leqslant \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$. Autrement dit, $\|\cdot\|_p$ est une norme sur K^n.

Démonstration. Le cas $p = 1$ est évident, c'est l'inégalité triangulaire. Supposons $p > 1$, et posons $q := \dfrac{p}{p-1}$, de sorte que $\dfrac{1}{p} + \dfrac{1}{q} = 1$. On applique l'inégalité de Hölder, en

choisissant judicieusement les a_j, b_j. Pour tout j, posons $a_j = |x_j|$ et $b_j := |x_j + y_j|^{p-1}$. L'inégalité de Hölder donne, en remarquant que $q(p-1) = p$:

$$\sum_{j=1}^{n} |x_j|\,|x_j+y_j|^{p-1} \leqslant \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p} \left(\sum_{j=1}^{n} |x_j+y_j|^p\right)^{1/q} = \|\mathbf{x}\|_p\,\|\mathbf{x}+\mathbf{y}\|_p^{p-1}.$$

De même $\sum\limits_{j=1}^{n} |y_j|\,|x_j+y_j|^{p-1} \leqslant \|\mathbf{y}\|_p\,\|\mathbf{x}+\mathbf{y}\|_p^{p-1}$. En sommant, il vient :

$$\|\mathbf{x}+\mathbf{y}\|_p^p = \sum_{j=1}^{n} |x_j+y_j|^p \leqslant \sum_{j=1}^{n} \big(|x_j|+|y_j|\big)\,|x_j+y_j|^{p-1} \leqslant \big(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p\big)\|\mathbf{x}+\mathbf{y}\|_p^{p-1},$$

grâce à l'inégalité triangulaire. Si $\mathbf{x}+\mathbf{y} := \mathbf{0}$, l'inégalité de Minkowski est vraie. Sinon, on l'obtient en divisant par $\|\mathbf{x}+\mathbf{y}\|_p^{p-1}$ les deux membres de l'inégalité précédente. ∎

Si $0 < p < 1$, $\|\cdot\|_p$ n'est pas une norme. On vérifie que, si $p = 1/2$, $\mathbf{x} := (1,2)$ et $\mathbf{y} := (3,4)$, on a $\|\mathbf{x}\|_{1/2} + \|\mathbf{y}\|_{1/2} = 19{,}75\ldots$, alors que $\|\mathbf{x}+\mathbf{y}\|_{1/2} = 19{,}79\ldots$
Dans les inégalités de Hölder et Minkowski, on peut remplacer les sommes finies par des intégrales. Pour simplifier, limitons-nous à des fonctions continues sur un segment $[a,b]$, $a < b$. On se donne un réel $p > 0$. Pour toute fonction $f \in \mathcal{C}\big([a,b], K\big)$, posons :

$$\|f\|_p := \left(\int_a^b |f(x)|^p\,\mathrm{d}x\right)^{1/p}. \tag{39}$$

Théorème 84. Soient p, q deux réels strictement positifs tels que $\dfrac{1}{p} + \dfrac{1}{q} = 1$.
Si $f, g : [a,b] \to \mathbb{C}$ sont continues, on a l'inégalité *de Hölder* suivante :

$$\left|\int_a^b fg\right| \leqslant \left(\int_a^b |f|^p\right)^{1/p} \left(\int_a^b |g|^q\right)^{1/q}. \tag{40}$$

Théorème 85. Considérons un réel $p \geqslant 1$. Alors $\|\cdot\|_p$ est une norme sur $\mathcal{C}\big([a,b], \mathbb{C}\big)$. Autrement dit, si $f, g : [a,b] \to \mathbb{C}$ sont continues, on a $\|f+g\|_p \leqslant \|f\|_p + \|g\|_p$, soit

$$\left(\int_a^b |f+g|^p\right)^{1/p} \leqslant \left(\int_a^b |f|^p\right)^{1/p} + \left(\int_a^b |g|^p\right)^{1/p} \quad \textit{(inégalité de Minkowski).} \tag{41}$$

Démonstration. La preuve de l'inégalité de Hölder est un simple passage à la limite. Si $h : [a,b] \to \mathbb{C}$ et si $n \in \mathbb{N}^*$, notons $S_n(h)$ la $n^{\text{ème}}$ somme de Riemann régulière à droite de h. En posant $x_k := a + k(b-a)/n$ pour $k = 1, \ldots, n$, $S_n(h)$ est la moyenne arithmétique des $h(x_k)$. La suite $\big(S_n(h)\big)$ converge vers $\displaystyle\int_a^b h$ (*cf.* le module II.3). C'est vrai, que h soit à valeurs réelles ou complexes.

Soit $n \in \mathbb{N}^*$. Appliquons l'inégalité (37) en prenant $a_k := f(x_k)$ et $b_k := g(x_k)$ pour $k = 1, \dots, n$. Nous obtenons ceci :

$$\left|\sum_{k=1}^{n} f(x_k)g(x_k\right| \leqslant \left(\sum_{k=1}^{n} |f(x_k)|^p\right)^{1/p} \left(\sum_{k=1}^{n} |g(x_k)|^q\right)^{1/q}.$$

En divisant par n, nous obtenons l'inégalité $|S_n(fg)| \leqslant S_n\big(|f|^p\big)^{1/p} S_n\big(|g|^q\big)^{1/q}$, à cause de l'égalité $\dfrac{1}{p} + \dfrac{1}{q} = 1$. Il suffit alors de faire tendre n vers $+\infty$.
La preuve de l'inégalité de Minkowski est entièrement analogue. ∎

Muni de la norme $\|\cdot\|_p$, l'espace $\mathcal{C}\big([a,b],\mathbb{C}\big)$ n'est pas complet (exercice II.1.42).

6 Espaces de Hilbert

Cette section doit être vue comme un pont entre d'une part le module I.4 et d'autre part le module II.4 ou le module « Polynômes orthogonaux » de [L2-L3]. Il s'agit essentiellement de fournir les généralisations « topologiques » de notions et résultats plus algébriques du module I.4, autant que ce sera nécessaire dans les modules mentionnés.

Définition 30. On appelle *espace de Hilbert* tout espace préhilbertien *complet*.

On parle d'espace de Hilbert réel (resp. complexe) si $K := \mathbb{R}$ (resp. $K := \mathbb{C}$). Voici des exemples qu'il convient de garder à l'esprit :

1. Tout espace euclidien (resp. hermitien) est un espace de Hilbert. En effet un tel espace est de dimension finie, il est donc complet (théorème 58 de la page 440). En particulier, l'espace euclidien usuel $\mathbb{R}^n$ et l'espace hermitien usuel $\mathbb{C}^n$ sont des espaces de Hilbert.
2. L'espace $\ell^2(K)$ est un espace de Hilbert, en vertu du théorème 33 de la page 424.
3. En revanche, muni de la norme $\|\cdot\|_2$ (*cf.* l'exemple 3b) de la page 394), l'espace préhilbertien $\mathcal{C}\big([a,b],K\big)$ (avec $a < b$) n'est pas complet, nous l'avons signalé. Cet espace (convenablement modifié) joue cependant un rôle important dans l'étude des séries de Fourier. En [L2-L3] et [MPA] nous verrons que, si l'on « agrandit » de manière appropriée un tel espace, on obtient un espace de Hilbert (« espace des fonctions de carré intégrable »), indispensable par exemple dans l'étude de la transformation de Fourier.

6.1 Théorème de projection

Dans le module I.4, nous avons vu que, si F est un sous-espace vectoriel *de dimension finie* d'un espace préhilbertien E, E est somme directe de F et de son orthogonal : $E = F \oplus F^\perp$. En outre, si $x \in E$, la distance $d(x,F)$ est atteinte en unique point, à savoir $p(x)$, où p est le projecteur orthogonal sur F. Ces résultats restent vrais lorsqu'on suppose seulement que F est complet, et l'on a en fait un théorème plus général :

Théorème et définition 86. Soient E un espace préhilbertien et C une partie convexe, complète (donc fermée) et non vide de E. Pour tout $x \in E$, la distance $d(x, C)$ de x à C est atteinte en un *unique* point c de C. Le point c est appelé *projection* de x sur le convexe C. De plus, pour tout $c' \in C$, on a :

$$\mathrm{Re}\,(x - c \mid c' - c) \leqslant 0. \tag{42}$$

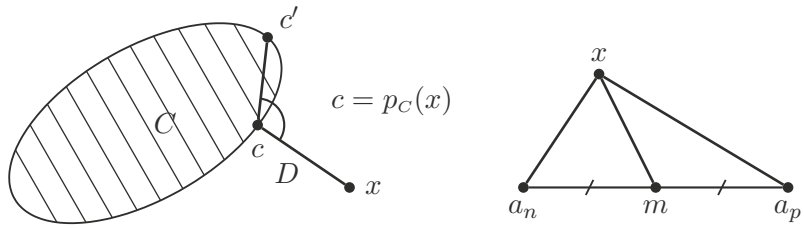


FIGURE II.1.6 – Projection sur un convexe complet.

Démonstration. Posons $D := d(x, C)$. Rappelons que $D := \inf_{a \in C} d(x, a)$. Pour tout entier $n \geqslant 1$, il existe donc $a_n \in C$ tel que $d(x, a_n) \leqslant D + 1/n$. Nous allons montrer que la suite (a_n) converge vers un point c ayant les propriétés requises. Commençons par prouver que (a_n) est une suite de Cauchy. Considérons deux entiers n, p tels que $1 \leqslant n < p$. Puisque C est convexe et $a_n, a_p \in C$, le point $m := (a_n + a_p)/2$ appartient à C, et donc $d(x, m) \geqslant D$. Écrivons l'égalité de la médiane dans le triangle formé par x, a_n, a_p :

$$\|a_n - a_p\|^2 + 4\|x - m\|^2 \;=\; 2\big[\|x - a_n\|^2 + \|x - a_p\|^2\big]. \tag{$*$}$$

De $\|x - a_n\| \leqslant D + 1/n$, $\|x - a_p\| \leqslant D + 1/p$ et $\|x - m\| \geqslant D$, on tire :

$$\|a_n - a_p\|^2 \leqslant 2\left(D + \frac{1}{n}\right)^2 + 2\left(D + \frac{1}{p}\right)^2 - 4D^2 \leqslant 4\left[\left(D + \frac{1}{n}\right)^2 - D^2\right] = \frac{4}{n}\left[2D + \frac{1}{n}\right].$$

On en déduit aisément que (a_n) est une suite de Cauchy de C. Puisque C est complet, cette suite converge vers un certain $c \in C$. La fonction d étant continue, les inégalités $D \leqslant d(x, a_n) \leqslant D + 1/n$ donnent, en faisant tendre n vers $+\infty$, $d(x, c) = D := d(x, C)$.

Prouvons l'unicité de c. Soit $c' \in C$ tel que $d(x, c') = D$. Dans l'égalité $(*)$, on peut remplacer a_n par c et a_p par c' : $\|c - c'\|^2 + 4\|x - m\|^2 = 2\big[\|x - c\|^2 + \|x - c'\|^2\big]$. Comme $\|x - c\| = \|x - c'\| = D$ et $\|x - m\| \geqslant D$, on a $\|c - c'\|^2 \leqslant 0$, d'où $c = c'$.

Il reste à vérifier l'inégalité (42) pour $c' \in C$ quelconque. Soit $t \in \,]0, 1]$. Puisque C est convexe, $z := (1 - t)c + tc' \in C$, donc $d(x, z) \geqslant d(x, c)$. Ainsi $\|x - c\|^2$ est majoré par :

$$\big\|x - \big[(1-t)c + tc'\big]\big\|^2 = \big\|(x - c) + t(c - c')\big\|^2 = \|x - c\|^2 + t^2\|c - c'\|^2 - 2t\,\mathrm{Re}\,(x - c \mid c' - c),$$

d'où $2t\,\mathrm{Re}\,(x - c \mid c' - c) \leqslant t^2\|c - c'\|^2$, soit $2\,\mathrm{Re}\,(x - c \mid c' - c) \leqslant t\|c - c'\|^2$.

En faisant tendre t vers 0, on obtient l'inégalité (42). ∎

Noter que, lorsque $x \neq c$ (autrement dit $x \notin C$) et $c' \neq c$, l'inégalité (42) signifie, lorsque $K := \mathbb{R}$, que l'angle des vecteurs $x - c$ et $c' - c$ est *obtus*, c'est-à-dire supérieur ou égal à $\pi/2$.
L'application de E dans C associant à tout $x \in E$ sa projection sur C sera notée p_C. L'exercice suivant en donne une caractérisation.

Exercice 20.
Conservons les hypothèses du théorème, et soit $x \in E$. Montrer que $c := p_C(x)$ est le seul point de C vérifiant les inégalités (42) (c' décrivant C). Montrer que p_C est 1-lipschitzienne.

Solution. Soient $x, \xi \in E$. Considérons *un* point c de C tel que $\mathrm{Re}\,(x - c \mid c' - c) \leqslant 0$ pour tout $c' \in C$. Considérons de même un point γ de C tel que $\mathrm{Re}\,(\xi - \gamma \mid c' - \gamma) \leqslant 0$ pour tout $c' \in C$. En prenant $c' := \gamma$ dans la première inégalité puis $c' := c$ dans la deuxième, on obtient $\mathrm{Re}\,(x - c \mid \gamma - c) \leqslant 0$ et $\mathrm{Re}\,(\xi - \gamma \mid c - \gamma) \leqslant 0$. Ajoutons :

$$0 \geqslant \mathrm{Re}\big[(x - c \mid \gamma - c) + (\xi - \gamma \mid c - \gamma)\big] = \mathrm{Re}\big((x - c) - (\xi - \gamma) \,\big|\, \gamma - c\big)$$
$$= \mathrm{Re}\big((x - \xi) - (c - \gamma) \,\big|\, \gamma - c\big) = \mathrm{Re}\,(x - \xi \mid \gamma - c) + \|\gamma - c\|^2.$$

D'où $\|\gamma - c\|^2 \leqslant |\mathrm{Re}\,(x - \xi \mid \gamma - c)| \leqslant \|x - \xi\|\,\|\gamma - c\|$, en vertu de l'inégalité de Cauchy-Schwarz. Ainsi $\|\gamma - c\|^2 \leqslant \|x - \xi\|\,\|\gamma - c\|$, d'où $\|\gamma - c\| \leqslant \|x - \xi\|$.
Appliquons d'abord ce qui précède en prenant $\xi := x$ et $\gamma := p_C(x)$. Alors $c = p_C(x)$, ce qui donne la caractérisation de $p_C(x)$ par les inégalités (42). Prenons ensuite x, ξ quelconques mais $c := p_C(x)$ et $\gamma := p_C(\xi)$. Alors $\big\|p_C(x) - p_C(\gamma\big\| \leqslant \|x - \xi\|$, ce qui montre que p_C est 1-lipschitzienne.

Dans le cas où C est un sous-espace vectoriel de E, voici la généralisation du théorème 14 de la page 152 mentionnée plus haut :

Théorème 87. Soient E un espace préhilbertien et F un sous–espace vectoriel complet (en particulier fermé) de E. Alors E est somme directe de F et de son orthogonal : $E = F \oplus F^\perp$; notons p le projecteur orthogonal de E sur F. Pour tout $x \in E$, $p(x)$ est le seul point de F en lequel la distance $d(x, F)$ est atteinte.

Démonstration. D'abord F est convexe et non vide (il contient 0). Appliquons le théorème précédent, en notant provisoirement p_F la projection de E sur F définie dans ce théorème. Soit $x \in E$, posons $c := p_F(x)$ et $y := x - c$. Nous savons que c est le seul point de F en lequel $d(x, F)$ soit atteinte. Puisque $x = c + y$, il suffit de montrer que y appartient à $F^\perp$. Soit $v \in F$. Puisque v et $-v$ sont dans F, l'inégalité (42) appliquée avec $c' = v + c$ puis $c' = -v + c$ montre que $\mathrm{Re}\,(y \mid v) = 0$ pour tout $v \in F$. Si $K := \mathbb{R}$, on a fini. Supposons que $K := \mathbb{C}$. On a aussi $\mathrm{Re}\,(y \mid iv) = 0$, soit $-\mathrm{Im}\,(y \mid v) = 0$ et donc $(y \mid v) = 0$ pour tout $v \in F$. Le vecteur y appartient donc à $F^\perp$. ∎

Dans un espace euclidien E, nous savons que l'application $x \longmapsto \big(y \mapsto (x \mid y)\big)$ est un isomorphisme de E sur son dual E^*, et l'on a une propriété analogue dans le cas hermitien. Voici la généralisation naturelle de ce résultat pour les espaces de Hilbert :

Théorème 88 (Riesz). Soit E un espace de Hilbert. À tout $a \in E$, associons la forme linéaire $\varphi_a : x \mapsto (\, a \mid x \,)$ sur E. Alors $\varphi : a \mapsto \varphi_a$ est une isométrie de E sur son dual topologique E_c^*. De plus, φ est linéaire si $K := \mathbb{R}$ et semi-linéaire si $K := \mathbb{C}$.

Démonstration. Rappelons que E_c^* est muni de la norme fonctionnelle $|||\cdot|||$. Pour tout $a \in E$, la forme linéaire $\varphi_a \in E^*$ est continue et $|||\varphi_a||| = \|a\|$, d'après l'exemple 2 de la page 417. Ainsi φ est une isométrie de E *dans* E_c^*, en particulier φ est injective.
Il reste la surjectivité de φ. Soit $f \neq 0$ une forme linéaire continue sur E. Son noyau H est un hyperplan de E, il est fermé car $H := f^{-1}\big(\{0\}\big)$. Puisque E est complet, H l'est aussi (théorème 34 de la page 425). D'après le théorème précédent, $E = H \oplus H^\perp$, donc $H^\perp$ est une droite vectorielle. Soit $a \in H^\perp$ un vecteur unitaire. Évidemment $H = a^\perp$, et ainsi les formes linéaires f et φ_a ont même noyau H. Il existe donc $\alpha \in K$ tel que $f = \alpha\varphi_a$. D'où $f = \alpha\varphi(a) = \varphi(\bar{\alpha}a) \in \operatorname{Im}\varphi$. ∎

6.2 Bases hilbertiennes

La notion de *base hilbertienne* généralise celle de *base orthonormée*. Il s'agit d'une notion importante, notamment dans l'étude des séries de Fourier du II.4 (voir aussi le module « Polynômes orthogonaux » de [L2-L3]). Il convient cependant de se méfier dès maintenant du vocabulaire : une base hilbertienne n'est pas toujours une base au sens algébrique, *i.e.* une famille libre et génératrice.

Définition 31. Soit E un espace préhilbertien. Une famille $(e_i)_{i\in I}$ de E est dite *totale* (ou *complète*) si le sous-espace vectoriel V qu'elle engendre est dense dans E, *i.e.* si tout $x \in E$ est limite d'une suite de V. Si de plus E est un espace de Hilbert et si $(e_i)_{i\in I}$ est orthonormée, on dit que c'est une *base hilbertienne* de E.

Dans cette définition, l'ensemble I est quelconque, fini ou non, dénombrable ou non. Il est clair que toute base orthonormée est une famille orthonormée totale (ici $V := E$). La réciproque est vraie lorsque E est de dimension finie (donc complet). En effet, soit $(x_i)_{i\in I}$ une base hilbertienne de E. C'est en particulier une famille libre. Puisque E est de dimension finie, I est fini. Mais alors V est de dimension finie, et par suite V est fermé dans E (théorème 59 de la page 441). Ainsi $V = E$ (V est à la fois fermé et dense dans E). Il en résulte que $(x_i)_{i\in I}$ est une base orthonormée de E.

Proposition 89. Soit $(e_i)_{i\in I}$ une famille orthonormée d'un espace préhilbertien E. Si cette famille est totale, c'est une famille orthonormée maximale (pour l'inclusion). La réciproque est vraie lorsque E est un espace de Hilbert.

Démonstration. Soit V le sous-espace vectoriel de E engendré par les e_i. Supposons que la famille $(e_i)_i$ ne soit pas une famille orthonormée maximale. Il existe alors un vecteur unitaire x orthogonal à tous les e_i, donc orthogonal à V. L'hyperplan $x^\perp$ contient alors V. Cet hyperplan étant fermé (exemple 2 de la page 417), il contient aussi $\overline{V}$, donc $\overline{V} \neq E$, *i.e.* la famille n'est pas totale.
Supposons inversement que la famille ne soit pas totale, c'est-à-dire $\overline{V} \neq E$, et que E soit complet. Dans ce cas, $\overline{V}$ est aussi complet (car il est fermé), et par conséquent $E = \overline{V} \oplus \overline{V}^\perp$

(théorème 87 de la page 462). L'espace $\overline{V}^{\perp}$ n'est alors pas nul, il contient un vecteur unitaire x, orthogonal à $\overline{V}$, donc à tous les e_i. Il en résulte que la famille orthonormée $(e_i)_i$ n'est pas maximale : en lui rajoutant x on obtient une famille orthonormée la contenant strictement. ∎

À l'aide de ce résultat et du lemme de Zorn on peut montrer, mais nous ne le ferons pas, que tout espace de Hilbert possède une base hilbertienne. Dans cet ouvrage, les espaces préhilbertiens les plus importants posséderont des familles orthonormées totales « naturelles », (*cf.* le module II.4 et celui de [L2-L3] sur les polynômes orthogonaux). Limitons-nous à un cas particulier significatif :

Exercice 21.

Soit E un espace préhilbertien possédant une famille totale finie ou dénombrable, ce qui est le cas s'il est séparable. Montrer que E possède une famille orthonormée totale finie ou dénombrable.

Solution. Le cas où E est de dimension finie est évident et laissé au lecteur. Supposons E de dimension infinie. Soient $(u_n)_{n\in\mathbb{N}}$ une famille totale de E, engendrant un sous-espace vectoriel V, de sorte que $\overline{V} = E$. Si cette famille n'est pas libre, on peut en extraire une famille libre engendrant encore V (sauriez-vous faire cette extraction explicitement ?). Supposons donc que $(u_n)_n$ soit libre, c'est donc une (vraie) base de V. Par le procédé de Gram-Schmidt, on en déduit une base orthonormée $(e_n)_n$ de V. Puisque $\overline{V} = E$, la famille $(e_n)_n$ est totale.

Si maintenant E est séparable (définition 10 de la page 407), il contient une partie dense D finie ou dénombrable, notons V le sous-espace vectoriel engendré par D. D'abord $\overline{V} = E$, car $E = \overline{D} \subset \overline{V} \subset E$. Il en résulte que V est de dimension infinie : sinon, il serait fermé dans E, donc égal à E, qui serait alors de dimension finie. Ainsi D est dénombrable : $D := \{u_n \mid n \in \mathbb{N}\}$. La suite $(u_n)_n$ est totale, et l'on conclut grâce à l'alinéa précédent.

Voici un exemple simple de base hilbertienne qui n'est pas une base au sens algébrique.

Exemple. Considérons l'espace de Hilbert $\ell^2(K)$, dont la norme est notée ici $\|\cdot\|$. Pour tout $k \in \mathbb{N}$, notons δ_k la suite dont tous les termes sont nuls sauf celui d'indice k, qui vaut 1. La famille $(\delta)_k$ est une famille orthonormée. Ce n'est pas une base de $\ell^2(K)$. Plus précisément, le sous-espace vectoriel V de $\ell^2(K)$ engendré par cette famille est formé des suites presque nulles (*i.e.* dont tous les termes sont nuls sauf un nombre fini). Évidemment $V \neq \ell^2(K)$: par exemple la suite (2^{-n}) appartient à $\ell^2(K) \setminus V$. Montrons que $(\delta_k)_{k\geqslant 0}$ est une base hilbertienne (dite *canonique*) de $\ell^2(K)$.

Soit $u := (u_n) \in \ell^2(K)$. La série $\sum_{n\geqslant 0} |u_n|^2$ converge, donc la suite (R_p) formée des restes de cette série converge vers 0. Rappelons que $R_p := \sum_{n=p+1}^{+\infty} |u_n|^2$. Soit $p \in \mathbb{N}$. La suite $r_p := u - \sum_{k=0}^{p} u_k\delta_k$ a les mêmes termes que u à partir du rang $p+1$, mais ses termes d'indices $0, 1, \ldots, p$ sont nuls. On en déduit que $\|r_p\|^2 = R_p$. Ainsi la suite (r_p) converge vers 0, ce qui montre que la suite des sommes partielles de la série $\sum_{k\geqslant 0} u_k\delta_k$ converge

vers u. Autrement dit, ladite série converge et l'on a : $u = \sum_{k=0}^{+\infty} u_k \delta_k$. Par ailleurs, chaque somme partielle $\sum_{k=0}^{p} u_k \delta_k$ appartient à V. Ainsi u est limite d'une suite de V, *i.e.* $u \in \overline{V}$. C'est vrai pour toute suite $u \in \ell^2(K)$, donc $\overline{V} = \ell^2(K)$. En d'autres termes, V est dense dans $\ell^2(K)$, donc $(\delta_k)_{k\geqslant 0}$ est une base hilbertienne de $\ell^2(K)$.

Voici un théorème capital, concernant les espaces préhilbertiens possédant une base hilbertienne *dénombrable* :

Théorème 90. Soient E un espace préhilbertien et $(e_n)_{n\geqslant 0}$ une famille orthonormée de E. Pour tous $x \in E$ et $n \in \mathbb{N}$, on pose $x_n := (\, e_n \mid x\,)$. Les conditions suivantes sont alors équivalentes :

(i) La famille $(e_n)_n$ est totale.

(ii) Pour tout $x \in E$, la série numérique $\sum_{n=0}^{+\infty} |x_n|^2$ est convergente, et l'on a :

$$\|x\|^2 = \sum_{n=0}^{+\infty} |x_n|^2 \qquad \textit{(égalité de Parseval)}.$$

(iii) Pour tous $x, y \in E$, la série numérique $\sum_{n\geqslant 0} \overline{x_n} y_n$ converge absolument, et l'on a :

$$(\, x \mid y\,) = \sum_{n=0}^{+\infty} \overline{x_n} y_n.$$

(iv) Pour tout $x \in E$, la série $\sum_{n\geqslant 0} x_n e_n$ converge (dans E) et sa somme est x.

Démonstration. Soit V le sous-espace vectoriel de E engendré par les e_n, $n \in \mathbb{N}$. Pour tout $n \in \mathbb{N}$, notons V_n le sous-espace vectoriel de V engendré par $(e_0, \ldots, e_n)$ (cette famille est une base orthonormée de V_n) et p_n le projecteur orthogonal de E sur V_n. La suite (V_n) est croissante pour l'inclusion, et la réunion des V_n est V.

$(i) \Longrightarrow (ii)$**.** Supposons (i).

Soit $x \in E$. Il existe une suite $(u_k)_{k\geqslant 0}$ de V convergeant vers x. Pour chaque k, il existe $n_k \in \mathbb{N}$ tel que $u_k \in V_{n_k}$, et l'on peut supposer que la suite $(n_k)_{k\geqslant 0}$ est strictement croissante. Si $k \in \mathbb{N}$, on a $d(x, V_{n_k}) \leqslant d(x, u_k)$. Puisque la suite $\big(d(x, u_k)\big)$ a pour limite 0, et puisque la suite $\big(d(x, V_n)\big)$ est décroissante, on en déduit que $\lim_{n\to+\infty} d(x, V_n) = 0$. Soit $n \in \mathbb{N}$. La proposition 13 de la page 151 ou son analogue sur $\mathbb{C}$ montre que $p_n(x) = \sum_{k=0}^{n} x_k e_k$, et le théorème 14 de la page 152 montre que $d(x, V_n) = \|x - p_n(x)\|$. La remarque suivant ce théorème donne :

$$\|x - p_n(x)\|^2 = \big(\, x - p_n(x) \;\big|\; x\,\big) = \|x\|^2 - \left(\sum_{k=0}^{n} x_k e_k \;\middle|\; x \right) = \|x\|^2 - \sum_{k=0}^{n} |x_k|^2. \qquad (*)$$

Il en résulte que la suite des sommes partielles de la série $\sum_{n\geqslant 0} |x_n|^2$ converge vers $\|x\|^2$, d'où l'assertion (ii).

$(ii) \iff (iii)$**.** L'assertion (ii) est un cas particulier $(x = y)$ de l'assertion (iii). L'assertion (iii) se déduit de l'assertion (ii) par polarisation. Traitons le cas $K := \mathbb{C}$ en laissant le cas réel (plus simple) au lecteur. Les identités de polarisation donnent :

$$4(\,x \mid y\,) \;=\; \|x+y\|^2 - \|x-y\|^2 + i\big[\|x-iy\|^2 - \|x+iy\|^2\big]$$

$$4\bar{x}_n y_n \;=\; |x_n+y_n|^2 - |x_n-y_n|^2 + i\big[|x_n-iy_n|^2 - |x_n+iy_n|^2\big].$$

L'assertion (ii) appliquée à $x+y$, $x-y$, $x+iy$ et $x-iy$ montre que la série $\sum_{n\geqslant 0} \bar{x}_n y_n$ est absolument convergente (comme somme de quatre séries absolument convergente), et que la somme de cette série est $(\,x \mid y\,)$.

$(ii) \Longrightarrow (iv)$**.** Soit $x \in E$. Pour tout $n \in \mathbb{N}$, $p_n(x)$ n'est autre que la somme partielle d'indice n de la série $\sum_{m\geqslant 0} x_m e_m$. Vu l'hypothèse et l'égalité $(*)$, on a $\lim_{n\to+\infty} \|x - p_n(x)\| = 0$, ce qui montre que la série $\sum_{m\geqslant 0} x_m e_m$ converge et a pour somme x.

$(iv) \Longrightarrow (i)$**.** Soit $x \in E$. Montrons que $x \in \overline{V}$. Par hypothèse, la suite $\big(p_n(x)\big)_n$ converge vers x. Pour tout n, $p_n(x) \in V_n \subset V$. Ainsi x est limite d'une suite de V, d'où la conclusion. ∎

La série $\sum_{n\geqslant 0} x_n e_n$ (dans E) ne converge pas toujours *absolument*.

Posons ainsi $a_n := 1/(n+1)$ pour tout $n \in \mathbb{N}$. La série $\sum_{n\geqslant 0} 1/(n+1)^2$ converge, donc $x := (a_n) \in \ell^2(\mathbb{R})$. Avec les notations de l'exemple de la page 464, $x_k := (\,\delta_k \mid x\,) = a_k$ pour tout k. La série $\sum_{k\geqslant 0} a_k \delta_k$ converge dans $\ell^2(\mathbb{R})$, et sa somme est x. Cette série ne converge pas absolument : pour tout k, $\|a_k\delta_k\| = 1/(k+1)$, et $\sum_{k\geqslant 0} 1/(k+1)$ diverge.

Remarque. Dans le module II.4 intervient un espace préhilbertien naturellement muni d'une famille orthonormée totale indexée non pas par $\mathbb{N}$ mais par $\mathbb{Z}$. Supposons plus généralement que $(e_i)_{i\in I}$ soit une famille orthonormée totale de l'espace préhilbertien E, l'ensemble I étant *dénombrable*, c'est-à-dire en bijection avec $\mathbb{N}$. Le sous-espace vectoriel V de E engendré par $(e_i)_{i\in I}$ est par hypothèse *dense* dans E. Lorsque $x \in E$ et $i \in I$, posons $x_i := (\,e_i \mid x\,)$.

Soit s une bijection de $\mathbb{N}$ sur I. La famille orthonormée $(e_{s(n)})_{n\geqslant 0}$ est totale (car c'est une base de V), appliquons-lui le théorème précédent. Soient $x, y \in E$. D'abord, $\|x\|^2 = \sum_{n=0}^{+\infty} \big|x_{s(n)}\big|^2$. De même, si t est une autre bijection de $\mathbb{N}$ sur I, $\|x\|^2 = \sum_{n=0}^{+\infty} \big|x_{t(n)}\big|^2$. Ainsi les séries $\sum_{n\geqslant 0} \big|x_{s(n)}\big|^2$ et $\sum_{n\geqslant 0} \big|x_{t(n)}\big|^2$ convergent et ont même

somme. On pourrait déduire l'égalité de ces deux sommes du théorème 42 de la page 431, mais c'est inutile, puisque ces deux sommes valent $\|x\|^2$! En tous cas, puisque la somme $\sum_{n\geqslant 0} \left|x_{s(n)}\right|^2$ ne dépend pas de s, il est naturel et commode de la noter $\sum_{i\in I} |x_i|^2$. En procédant de même pour les assertions (ii) et (iii) du théorème, nous aboutissons aux « égalités » suivantes :

$$\|x\|^2 \;=\; \sum_{i\in I} |x_i|^2\,, \qquad (\,x \mid y\,) \;=\; \sum_{i\in I} \bar{x}_i y_i, \qquad x \;=\; \sum_{i\in I} x_i e_i. \tag{43}$$

La troisième « égalité », par exemple, signifie simplement ceci : *quelle que soit* la bijection s de $\mathbb{N}$ sur I, la série $\sum_{n\geqslant 0} x_{s(n)} e_{s(n)}$ converge dans E et sa somme est x.

Le corollaire suivant souligne l'importance de l'espace $\ell^2(K)$:

Corollaire 91. Soient E un espace préhilbertien et $(e_n)_{n\in\mathbb{N}}$ une famille orthonormée totale de E. Soit $\varphi : E \to \ell^2(K)$ l'application associant à tout $x \in E$ la suite $(x_n)_n$, où $x_n := (\,e_n \mid x\,)$ pour tout n. Alors :

1. l'application φ est linéaire injective et conserve la norme,
2. si de plus E est un espace de Hilbert, l'application φ est surjective, c'est donc une isométrie linéaire de E sur $\ell^2(K)$.

Démonstration. Observons d'abord que, si E n'est pas complet, φ ne peut pas être une isométrie de E sur $\ell^2(K)$, car $\ell^2(K)$ est complet (théorème 33 de la page 424). L'application φ est clairement linéaire. Notons $\|\cdot\|_2$ la norme de $\ell^2(K)$. Soit $x \in E$. D'après le théorème précédent, la série $\sum_{n\geqslant 0} |x_n|^2$ converge, autrement dit $\varphi(x) := (x_n)$ appartient à $\ell^2(K)$. De plus, la somme $\|\varphi(x)\|_2^2$ de cette série est égale à $\|x\|^2$ (égalité de Parseval), donc $\|\varphi(x)\|_2 = \|x\|$ pour tout $x \in E$: φ conserve la norme. Le noyau de φ est donc nul, *i.e.* φ est injective.

Prouvons maintenant la surjectivité de φ lorsque E est complet. Soit $u := (u_n) \in \ell^2(K)$. La série $\sum_{n\geqslant 0} |u_n|^2$ est donc convergente. L'idée est de montrer que la série $\sum_{n\geqslant 0} u_n e_n$ converge (dans E) et que, si l'on note x sa somme, on a $\varphi(x) = u$. Soit $\varepsilon > 0$. Puisque $\sum_{n\geqslant 0} |u_n|^2$ converge, elle vérifie le critère de Cauchy. Il existe donc un N tel que $\sum_{k=n}^{p} |u_k|^2 \leqslant \varepsilon$ dès que $N \leqslant n < p$. Puisque les e_k sont deux à deux orthogonaux, on a $\left\| \sum_{k=n}^{p} u_k e_k \right\|^2 = \sum_{k=n}^{p} |u_k|^2 \leqslant \varepsilon$, par Pythagore. La série $\sum_{n\geqslant 0} u_n e_n$ vérifie ainsi le critère de Cauchy. Puisque E est complet, cette série converge ; notons x sa somme.

Soit $k \in \mathbb{N}$. La forme linéaire $y \mapsto (\,e_k \mid y\,)$ sur E est continue. En vertu du théorème 10 de la page 409, $(\,e_k \mid x\,)$ est somme de la série $\sum_{n\geqslant 0} (\,e_k \mid u_n e_n\,)$, somme qui vaut évidemment u_k. Ainsi $(\,e_k \mid x\,) = u_k$ pour tout k, et par suite $u = \big((\,e_k \mid x\,)\big)_{k\geqslant 0} = \varphi(x)$. ∎

Exercice type corrigé

II.1.0 Soient $(E, \|\cdot\|)$ un e.v. n. réel de dimension finie $n \geqslant 2$, S la sphère unité de E et $\mathcal{B}$ une base de E fixée.
Le dual topologique E_c^* de E est muni de la norme $|||\cdot||| : \forall f \in E^*,\ |||f||| := \sup_{x\in S} \big|f(x)\big|$.

1. Montrer qu'il existe une base $\mathcal{C} := (e_1, \dots, e_n)$ de E telle que, en notant $(e_1^*, \dots, e_n^*)$ la base duale de $\mathcal{C}$, on ait $\|e_i\| = |||e_i^*||| = 1$ pour $i = 1, \dots, n$. On étudiera l'application $g : E^n \to \mathbb{R}$ associant à toute famille $(y_1, \dots, y_n) \in E^n$ son déterminant dans la base $\mathcal{B}$.

2. Une telle base étant donnée, montrer que, pour tout $x := \sum_{i=1}^{n} x_i e_i \in E$, on a :

$$\max\big(|x_1|, \dots, |x_n|\big) \ \leqslant\ \|x\| \ \leqslant\ \sum_{i=1}^{n} |x_i|\,.$$

Solution.
1. L'application g est n-linéaire (propriété du déterminant de n vecteurs dans une base). Puisque E est de dimension finie, g est donc continue. Soit $h : S^n \to \mathbb{R}$ la restriction de g à S^n, elle est continue puisque g l'est. La sphère S est compacte (parce que E est de dimension finie), donc S^n est compact. Il en résulte que h atteint son maximum absolu M en un point $\mathcal{C} := (e_1, \dots, e_n)$ de S^n. Nous allons montrer que $\mathcal{C}$ répond à la question.
Pour tout $i \in [\![1, n]\!]$, on a $\|e_i\| = 1$ puisque $e_i \in S$. Il faut vérifier que $\mathcal{C}$ est une base de E ou, c'est équivalent, que $\det_{\mathcal{B}}(e_1, \dots, e_n) := M$ n'est pas nul. Écrivons $\mathcal{B} := (\varepsilon_1, \dots, \varepsilon_n)$ et posons $u_i := \varepsilon_i/\|\varepsilon_i\| \in S$ pour tout i. On a $\det_{\mathcal{B}}(\varepsilon_1, \dots, \varepsilon_n) = 1$, par définition, d'où par multilinéarité :

$$h(u_1, \dots, u_n) \ =\ \left(\prod_{i=1}^{n} \|\varepsilon_i\|\right)^{-1} > 0.$$

D'où $M > 0$, puisque $h(u_1, \dots, u_n) \leqslant M$. Ainsi $\mathcal{C}$ est une base de E, ce qui permet de considérer sa base duale $(e_1^*, \dots, e_n^*)$.
Soit $i \in [\![1, n]\!]$. L'égalité $e_i^*(e_i) = 1$ donne : $1 \leqslant |||e_i^*|||\|e_i\| = |||e_i^*|||$, *cf.* l'inégalité (18) de la page 416, d'où $|||e_i^*||| \geqslant 1$. Pour l'inégalité inverse, raisonnons par l'absurde, en supposant $|||e_i^*||| > 1$. D'après l'exercice 13 de la page 442, il existe $a \in S$ tel que $|e_i^*(a)| = |||e_i^*||| > 1$. Quitte à remplacer a par $-a$, on peut supposer $e_i^*(a) > 1$. Écrivons $a = \sum_{j=1}^{n} a_j e_j$, de sorte que $a_i = e_i^*(a) > 1$. Le caractère multilinéaire alterné du déterminant dans la base $\mathcal{B}$ donne, en remplaçant e_i par a :

$$h(e_1, \dots, a, \dots, e_n) = g(e_1, \dots, a_i e_i, \dots, e_n) = a_i g(e_1, \dots, e_n) = a_i M > M,$$

ce qui est absurde puisque M est le maximum absolu de h. Ainsi $|||e_i^*||| = 1$

2. Soit $x := \sum_{i=1}^{n} x_i e_i \in E$.

Puisque les e_i sont unitaires, l'inégalité triangulaire donne : $\|x\| \leqslant \sum_{i=1}^{n} |x_i|$. Pour tout $i \in [\![1, n]\!]$, $x_i = e_i^*(x)$, d'où $|x_i| \leqslant |||e_i^*|||\|x\| = \|x\|$. Le maximum des $|x_i|$ est donc majoré par $\|x\|$.

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 (u_n) une suite de l'espace métrique X et A l'ensemble des valeurs de cette suite. Montrer que $\overline{A}$ est la réunion de A et de l'ensemble des valeurs d'adhérence de (u_n).

II.1.2 Soient x, y deux vecteurs non nuls d'un e.v. n. E. Établir l'inégalité :

$$\|x-y\| \geqslant \frac{1}{2}\max\big(\|x\|,\|y\|\big)\left\|\frac{x}{\|x\|}-\frac{y}{\|y\|}\right\|.$$

Peut-on améliorer le facteur $1/2$?

II.1.3 Soit A une partie non vide de $\mathbb{R}$.
Pour tout polynôme $P \in \mathbb{R}[X]$, on pose $\|P\|_A := \sup_{x\in A}|P(x)|$.
Quelles conditions A doit-elle satisfaire pour que l'on obtienne ainsi une norme $\|\cdot\|_A$ sur $\mathbb{R}[X]$?

II.1.4 Pour tout $(x,y) \in \mathbb{R}^2$, on pose $\|(x,y)\| := \sup_{t\in\mathbb{R}}\left|\dfrac{x+ty}{1+t^2}\right|$.
Montrer que l'on définit ainsi une norme sur $\mathbb{R}^2$, et dessiner la boule unité.

II.1.5 Soit K un corps quelconque. Si $P = \sum_{n\geqslant 0} a_n X^n \in K[X]$ est non nul, l'*ordre* $\omega(P)$ de P est le plus *petit* entier $n \geqslant 0$ tel que $a_n \neq 0$. Pour tous $P, Q \in K[X]$, on pose $d(P,Q) := 0$ si $P = Q$ et $d(P,Q) := \exp\big(-\omega(P-Q)\big)$ si $P \neq Q$.

1. Montrer que d est une distance sur $K[X]$, vérifiant la propriété suivante :
$\forall P, Q, R \in K[X]\,,\ d(P,R) \leqslant \max\big(d(P,Q), d(Q,R)\big)$ (« inégalité ultramétrique »).

2. Montrer que le résultat de l'exercice 2 de la page 399 est faux dans l'espace métrique $\big(K[X], d\big)$.
Donner dans cet espace un exemple de sphère d'intérieur non vide.

II.1.6 Si $P \in \mathbb{C}[X]$, on note $\|P\|$ le plus grand des modules des coefficients de P.

1. Montrer que $\|\cdot\|$ est une norme sur $\mathbb{C}[X]$.

2. On fixe $t \in \mathbb{C}$, et l'on note L l'application $P \mapsto P(t)$ de $\mathbb{C}[X]$ dans $\mathbb{C}$.
À quelle condition $L : \big(\mathbb{C}[X], \|\cdot\|\big) \to \mathbb{C}$ est-elle continue ?

II.1.7 Soient a, b deux réels, $a < b$.
On munit $\mathbb{R}[X]$ de la norme associée au produit scalaire $(\,P \mid Q\,) = \displaystyle\int_a^b P(t)Q(t)\,\mathrm{d}t$.
La forme linéaire L sur $\mathbb{R}[X]$ définie par $L(X^n) = 1$ pour tout n est-elle continue ?

II.1.8 Soit A un sous-espace de l'espace métrique X. Montrer que les fermés (resp. ouverts) de A sont exactement les $A \cap T$, où T est un fermé (resp. un ouvert) de X.

II.1.9 Soient A, B deux parties de l'espace métrique X.
Montrer que $\overline{A \cup B} = \overline{A} \cup \overline{B}$ et $\overline{A \cap B} \subset \overline{A} \cap \overline{B}$. Donner un exemple dans lequel $\overline{A \cap B} \neq \overline{A} \cap \overline{B}$.

II.1.10 Dans un espace métrique X, soient U, V deux ouverts disjoints.
Montrer que les intérieurs de $\overline{U}$ et $\overline{V}$ sont disjoints.

II.1.11 Soient E un e.v.n. non nul et $u, v \in \mathcal{L}_c(E) := \mathcal{L}_c(E, E)$.
Montrer que l'égalité $u \circ v - v \circ u = \mathrm{Id}_E$ est impossible.

II.1.12 Soit X un espace métrique. On suppose que X est réunion d'un nombre fini de fermés $F_1, \dots, F_n$. Montrer qu'une application f de X dans un espace métrique Y est continue si, et seulement si, sa restriction à chaque F_i est continue.

II.1.13 On munit K^n de la norme $\|\cdot\|_1$ (exemple 4 de la page 395).
On munit ensuite $M_n(K)$ de la norme subordonnée $|||\cdot|||_1$ correspondante.
Si $A := (a_{i,j}) \in M_n(K)$, expliciter $|||A|||_1$ en fonction des $a_{i,j}$.

II.1.14 Soient E un e.v.n. de dimension finie et $u \in \mathcal{L}(E)$ un endomorphisme tel que $|||u||| \leqslant 1$.
Montrer que E est somme directe du noyau et de l'image de $u - \mathrm{Id}_E$.

II.1.15 Soient E un e.v.n., f une forme linéaire continue non nulle sur E et H le noyau de f.
Pour tout $x \in E$, établir l'égalité $|f(x)| = d(x, H)|||f|||$.

Exercices sur la section 2

II.1.16 Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction continue.
Pour tous $x, y \in \mathbb{R}$, on pose : $d(x, y) := |f(x) - f(y)|$. À quelle condition (sur f) obtient-on une distance sur $\mathbb{R}$? Si c'est le cas, à quelle condition l'espace métrique $(\mathbb{R}, d)$ est-il complet ?

II.1.17 Montrer que l'espace métrique de l'exercice II.1.5 n'est pas complet.

II.1.18 Soient $f, g : \mathbb{R} \to \mathbb{R}$ deux fonctions de classe $\mathcal{C}^1$.
On suppose qu'il existe $k \in [0, 1[$ tel que $|f'(x)| \leqslant k$ et $|g'(x)| \leqslant k$ pour tout $x \in \mathbb{R}$.
Montrer que l'application $F : (x, y) \mapsto \big(x + f(y), y + g(x)\big)$ est un difféomorphisme de classe $\mathcal{C}^1$ de $\mathbb{R}^2$ sur $\mathbb{R}^2$.

II.1.19 $*$ Soit $f : \mathbb{R}^n \to \mathbb{R}^n$ une application de classe $\mathcal{C}^1$ vérifiant la condition suivante :

$$\|f(x) - f(y)\| \geqslant \|x - y\| \qquad \text{pour tous } x, y \in \mathbb{R}^n.$$

Montrer que F est un difféomorphisme de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ sur $\mathbb{R}^n$.

II.1.20 Soient X un espace métrique et A une partie dense de X. Par ailleurs, soit f une application *uniformément continue* de A dans un espace métrique complet Y. Montrer que f se prolonge d'une seule manière en une application continue $g : X \to Y$, et que g est uniformément continue.

II.1.21 Soit $\mathcal{S}$ la $\mathbb{Q}$-algèbre formée des suites de $\mathbb{Q}$, le produit de deux suites (u_n) et (v_n) étant $(u_n v_n)$.

1. Montrer que les suites de Cauchy de $\mathbb{Q}$ forment une sous-algèbre $\mathcal{A}$ de $\mathcal{S}$ et que les suites convergeant vers 0 forment un idéal *maximal* I de $\mathcal{A}$.

2. Établir un isomorphisme de $\mathbb{Q}$-algèbres de $\mathcal{A}/I$ sur $\mathbb{R}$. Cela suggère-t-il une construction de $\mathbb{R}$ à partir de $\mathbb{Q}$?

II.1.22 Soient X un espace métrique complet et f une application de X dans X. Soit $n \in \mathbb{N}^*$. On suppose que $f^n = f \circ f \circ \circ \cdots \circ f$ (il y a n facteurs) est contractante. Montrer alors que f possède un unique point fixe.

Exercices sur la section 3

II.1.23 Soit X un espace métrique. Montrer que X est compact et discret si, et seulement s'il est fini.

II.1.24 Soit X l'ensemble des points $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ tels que $x_i \geqslant 0$ pour tout i et $x_1 + \cdots + x_n = 1$.

1. Montrer que X est compact.

2. Soit $f : X \to \mathbb{R}$ la fonction $(x_1, \ldots, x_n) \mapsto x_1 x_2 \cdots x_n$. Montrer que f a un maximum absolu M et que, si $x := (x_1, \ldots, x_n) \in X$ et $f(x) = M$, tous les x_i sont égaux.

3. En déduire l'inégalité arithmético-géométrique (*cf.* la proposition 81 de la page 457).

II.1.25 Soient (X, d) un espace métrique compact et $f : X \to X$ une application telle que $d\big(f(x), f(y)\big) < d(x, y)$ pour tous $x, y \in X$, $x \neq y$. Montrer que f a un unique point fixe.

II.1.26 $**$ Le but de cet exercice est de montrer que l'espace métrique $(\mathcal{K}, \delta)$ défini dans l'exercice 11 de la page 439 est complet. On considère donc une suite de Cauchy (A_k) de $\mathcal{K}$. Pour tout $k \in \mathbb{N}$, on note F_k l'adhérence de la réunion des A_p pour $p \geqslant k$. On note enfin l'intersection des F_k, $k \in \mathbb{N}$. Montrer alors que F est limite de la suite (A_k).

II.1.27 $**$ On reprend les notations de l'exercice précédent. Soit h (resp. h') l'homothétie de centre 0 (resp. 1) et de rapport $1/3$ dans $\mathbb{R}$.
Pour tout $x \in \mathbb{R}$, on a donc $h(x) = x/3$ et $h'(x) = (x+2)/3$.
Soit $f : \mathcal{K} \to \mathcal{K}$ l'application $A \mapsto h(A) \cup h'(A)$.

1. Montrer que f est contractante avec comme rapport de contraction $1/3$.

2. Montrer que f possède dans $\mathcal{K}$ un unique point fixe, à savoir l'ensemble triadique de Cantor K_3 (se reporter à la proposition 53 de la page 436 et à sa preuve).

II.1.28 $*$ On reprend les hypothèses et notations de l'exercice 10 de la page 437. On suppose que $\mathcal{T} := M_1M_2M_3$ est un triangle de périmètre minimum. Montrer que la droite O_1M_1 est la bissectrice intérieure de $\mathcal{T}$ en M_1, c'est-à-dire que les deux angles indiqués sur la figure sont égaux.

Exercices sur la section 4

II.1.29 $*$ Soient E un e.v.n. de dimension finie $n \geqslant 1$ et A une partie bornée non vide de E. Montrer que A et sa frontière $\partial(A)$ ont le même diamètre. On pourra se ramener au cas où A est compacte et commencer par le cas $n = 1$.

II.1.30 Soient E un K-espace vectoriel de dimension finie et $\|\cdot\|, \|\cdot\|'$ deux normes sur E. Définissons $f : E \to E$ par $f(0) := 0$ et $f(x) := (\|x\|'/\|x\|)x$ si $x \neq 0$. Montrer que f est bijective et expliciter la fonction réciproque. Montrer que f est un homéomorphisme de E sur lui-même, appliquant $\overline{B}(0,1)$, boule unité fermée pour $\|\cdot\|$ sur $\overline{B}'(0,1)$, boule unité fermée pour $\|\cdot\|'$, et de même pour les sphères $S(0,1)$ et $S'(0,1)$.

II.1.31 Montrer que tout e.v. n. de dimension finie est séparable (définition 10 de la page 407).

II.1.32 $**$ Soit $A \in M_n(\mathbb{C})$, notons $\rho(A)$ le *rayon spectral* de A, *i.e.* le plus grand des modules des valeurs propres de A.

1. Montrer que $\rho(A) \leqslant \|A\|$ pour toute norme matricielle $\|\cdot\|$ sur $M_n(\mathbb{C})$.

2. Soit $\varepsilon > 0$.
Montrer qu'il existe une norme matricielle subordonnée $|||\cdot|||$ telle que $|||A||| \leqslant \rho(A) + \varepsilon$.

3. En déduire l'équivalence des conditions suivantes :

a) La suite $(A^k)_{k\geqslant 0}$ converge vers 0.

b) On a $\rho(A) < 1$.

c) Il existe une norme matricielle subordonnée $|||\cdot|||$ telle que $|||A||| < 1$.

d) Pour tous vecteurs $C, X_0 \in \mathbb{C}^n$, la suite $(X_k)_k$ de $\mathbb{C}^n$ définie par son premier terme X_0 et la relation de récurrence $X_{k+1} := AX_k + C$ converge. La limite de cette suite est de plus $(I_n - A)^{-1}C$.

4. Soit $\|\cdot\|$ une norme matricielle quelconque sur $M_n(\mathbb{C})$. Prouver la formule :

$$\rho(A) \;=\; \lim_{k\to+\infty} \|A^k\|^{1/k}.$$

Exercices sur la section 5

II.1.33 Soient C_1, C_2 deux parties convexes d'un espace vectoriel réel et $\lambda_1, \lambda_2 \in \mathbb{R}$. Montrer que $\lambda_1 C_1 + \lambda_2 C_2 := \{\lambda_1 x_1 + \lambda_2 x_2 \mid (x_1, x_2) \in C_1 \times C_2\}$ est convexe.

II.1.34 Soit $f : \mathbb{R}_+ \to \mathbb{R}$ une fonction concave telle que $f(0) = 0$ et $f(t) > 0$ pour tout $t > 0$. Montrer que f est croissante et $f(a + b) \leqslant f(a) + f(b)$ pour tous $a, b \geqslant 0$. En déduire que, si d est une distance sur un ensemble X, $f \circ d$ est aussi une distance sur X.
Exemples : $f(t) := \min(1, t)$, $f(t) := t/(1 + t)$.

II.1.35 Soit C une partie fermée d'un e.v. n. E. Montrer que C est convexe si, et seulement si, C contient le milieu de tout segment joignant deux points de C.

II.1.36 Montrer que les composantes connexes d'un ouvert de $\mathbb{R}^n$ sont ouvertes. En déduire que tout ouvert de $\mathbb{R}$ est réunion d'une famille finie ou dénombrable d'intervalles ouverts deux à deux disjoints.

II.1.37 Dans $\mathbb{R}^2$, soit E la réunion des deux axes de coordonnées. En considérant les composantes connexes de $E \setminus \{(0, 0)\}$, montrer que E n'est pas homéomorphe à $\mathbb{R}$.

II.1.38 Montrer que $\mathbb{R}^2 \setminus \{0\}$ est connexe par arcs.
En déduire que $\mathbb{R}$ et $\mathbb{R}^2$ ne sont pas homéomorphes.

II.1.39 $*$ On considère un entier $n \geqslant 2$ et une partie finie S de $\mathbb{R}^n$. Montrer que $X := \mathbb{R}^n \setminus S$ est connexe par arcs.

II.1.40 $*$ Soit $n \in \mathbb{N}^*$. Montrer que l'ensemble $\mathbf{P}^{n-1}(\mathbb{R})$ des droites vectorielles de $\mathbb{R}^n$, muni de la distance δ définie page 398, est compact et connexe par arcs.

II.1.41 $*$ Soit $f :]0, +\infty[\to \mathbb{R}$ la fonction $x \mapsto \sin(1/x)$ et $C \subset \mathbb{R}^2$ son graphe.

1. Montrer que $\overline{C}$ est la réunion de C et du segment $T := \{(0, y) \mid y \in [-1, 1]\}$.

2. Montrer que $\overline{C}$ est connexe, mais n'est pas connexe par arcs.

II.1.42 Considérons un réel $p \geqslant 1$. Le but de cet exercice est de montrer que l'espace $E = \mathcal{C}\big([0, 1], \mathbb{R}\big)$, muni de la norme $\|\cdot\|_p$ (*cf.* la section 5.4), n'est pas complet.

1. Soit $\alpha := 1/(p + 1)$. Pour tout $n \in \mathbb{N}^*$, soit $f_n : [0, 1] \to \mathbb{R}$ la fonction définie par $f_n(t) = \min\big(n, t^{-\alpha}\big)$ $\big(f_n(0) = n\big)$. Montrer que $(f_n)_n$ est une suite de Cauchy de E.

2. On suppose que la suite $(f_n)_n$ possède dans E une limite f. En se référant au module II.3, montrer que $\displaystyle\int_0^1 \big|f(t) - t^{-\alpha}\big|^p \, \mathrm{d}t = 0$. Conclusion ?

II.1.43 $**$ Soit K_3 l'ensemble triadique de Cantor (*cf.* page 436).

1. Montrer que K_3 a la puissance du continu, *i.e.* est en bijection avec $\mathbb{R}$ (se reporter au module I.1 du cours de [L1]).

2. En se référant au module II.3, montrer que K_3 est de longueur nulle, en ce sens que la fonction caractéristique $\chi : \mathbb{R} \to \mathbb{R}$ de K_3 (valant 1 sur K_3 et 0 en dehors) est intégrable et $\displaystyle\int_{\mathbb{R}} \chi(t) \, \mathrm{d}t = 0$.

3. Montrer que K_3 est d'intérieur vide et que, plus précisément, il ne contient aucun intervalle $]a, b[$, $a < b$. En déduire que les composantes connexes de K_3 sont des points. On dit que K_3 est « totalement discontinu ».

Exercices sur la section 6

II.1.44 Soit V un sous-espace vectoriel de l'espace de Hilbert E.
Montrer que l'orthogonal $V^{\perp\perp}$ de $V^{\perp}$ est égal à $\overline{V}$.

II.1.45 Montrer qu'un espace de Hilbert possédant une base hilbertienne dénombrable est séparable.

II.1.46 $*$ Soient E un espace de Hilbert réel et $f : E \times E \to \mathbb{R}$ une forme bilinéaire continue. On suppose qu'il existe un réel $k > 0$ tel que $\forall x \in E : f(x, x) \geqslant k\|x\|^2$ (on dit alors que f est *coercive*). Soit φ une forme linéaire continue sur E. On veut montrer qu'il existe un unique $a \in E$ tel que $\forall x \in E : \varphi(x) = f(a, x)$ (théorème de Lax-Milgram).

1. Montrer qu'il existe un unique $b \in E$ tel que $\forall x \in E : \varphi(x) = (\, b \mid x \,)$.

2. Montrer qu'il existe un unique $u \in \mathcal{L}(E)$ tel que $\forall x, y \in E : f(x, y) = \big(\, u(x) \;\big|\; y \,\big)$. Montrer ensuite que u est continu.

3. Si r est un nombre réel strictement positif suffisamment petit, montrer que l'application $x \mapsto x + r\big(b - u(x)\big)$ de E dans lui-même est contractante. Conclure.

4. Donner une preuve directe du théorème lorsque f est symétrique. On observera que f est alors un produit scalaire et que la norme associée est équivalente à la norme $\|\cdot\|$.

Suites et séries de fonctions

II.2

Les séries de fonctions constituent un outil fondamental, que ce soit pour obtenir des approximations d'une fonction donnée ou pour définir de nouvelles fonctions et en étudier les propriétés. Deux types de séries de fonctions ont une importance particulière : les séries entières, qui font l'objet de la section 3, et les séries de Fourier, étudiées dans le module II.4.

1 Suites de fonctions

En passant des suites et séries numériques aux suites et séries de fonctions, on ne manipule plus des nombres mais des fonctions, *i.e.* on ne travaille plus dans $\mathbb{R}$ (ou $\mathbb{R}^p$), mais dans des espaces vectoriels (de fonctions) de dimension infinie. Il s'agit donc d'un saut qualitatif, dont la difficulté ne doit pas être minimisée (ni exagérée). Il va de soi que l'étude des suites et séries numériques faite dans [L1] sert de façon permanente. En outre, il sera fréquemment fait appel aux notions et résultats « topologiques » du module II.1.

1.1 Convergence simple, convergence uniforme

Soient X un ensemble non vide, $(E, \|\cdot\|)$ un espace vectoriel normé réel (en abrégé e.v. n., *cf.* le module II.1), et (f_n) une suite d'*applications* de X dans E. Lorsque E est égal à $\mathbb{R}$ ou $\mathbb{C}$, comme c'est le plus souvent le cas dans les applications, on dit plutôt que (f_n) est une suite de *fonctions* (réelles ou complexes). Peut-on parler de « convergence de la suite (f_n) », de la même façon qu'on parle de convergence d'une suite numérique ? Nous allons voir que l'on peut donner au terme « convergence » plusieurs sens ; on dit qu'il y a plusieurs « modes de convergence » de la suite (f_n).

Le premier mode de convergence, naturel mais peu utile, est la « convergence simple ». Voici de quoi il s'agit. Pour tout $x \in X$ fixé, $\big(f_n(x)\big)$ est une suite de E, et l'on peut exiger que cette suite converge (dans E), ceci pour tout $x \in X$.

Cela conduit à la définition suivante :

Définition 1. On dit que la suite d'applications (f_n) *converge simplement* si, pour tout $x \in X$, la suite $\big(f_n(x)\big)$ de E converge. Dans ce cas, pour tout $x \in X$, on peut poser $f(x) := \lim\limits_{n\to+\infty} f_n(x)$, ce qui définit une application $f : X \to E$. On dit que f est *limite simple* de la suite (f_n), ou que la suite (f_n) *converge simplement* vers f.

Le fait que (f_n) converge simplement vers f se traduit ainsi :

$$\forall x \in X,\ \forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \geqslant N,\ \|f_n(x) - f(x)\| \leqslant \varepsilon. \tag{1}$$

Il est important de remarquer que l'entier N dépend ici à la fois de x et de ε. Par ailleurs, si (f_n) converge simplement vers f et aussi vers $g : X \to E$, on a $g = f$ (*unicité de la limite simple éventuelle*). En effet, pour tout $x \in X$, la suite $\big(f_n(x)\big)$ converge à la fois vers $f(x)$ et vers $g(x)$, donc $f(x) = g(x)$ (unicité de la limite d'une suite de E).

Exemples.

1. Prenons $X := \mathbb{R}$, $E := \mathbb{R}$ et $f_n(x) := x^n$ pour tout $(n, x) \in \mathbb{N} \times \mathbb{R}$. Soit $x \in \mathbb{R}$ fixé. La suite réelle (x^n) converge si, et seulement si, $x \in]-1, 1]$. On dit alors que le domaine de convergence simple de la suite (f_n) est $]-1, 1]$. De plus, (x^n) converge vers 0 si $|x| < 1$ et vers 1 si $x := 1$. Nous dirons donc que la suite (f_n) converge simplement *sur* $]-1, 1]$ vers la fonction $f :]-1, 1] \to \mathbb{R}$ définie par $f(1) := 1$ et $f(x) := 0$ si $x \neq 1$.

   Pour tout n, notons g_n la restriction de f_n à $]-1, 1]$. La suite (g_n) converge donc simplement vers f. Noter que la fonction f n'est pas continue (à gauche) en 1, bien que les g_n soient continues. La convergence simple n'est donc pas un mode de convergence assez fort pour conserver la continuité. La « convergence uniforme » corrigera ce défaut.

   Soient $x \in]-1, 1[$ et $\varepsilon > 0$. Soit $N \in \mathbb{N}$. L'implication $(n \geqslant N) \Longrightarrow \big(|x^n| \leqslant \varepsilon\big)$ est vraie si, et seulement si, $N \ln|x| \leqslant \ln \varepsilon$, ce qui équivaut à $N \geqslant \ln \varepsilon / \ln|x|$ puisque $\ln|x| < 0$. Le fait que N dépende de ε et de x est ici bien visible.

2. Prenons $X := \mathbb{C}$, $E := \mathbb{C}$ et $f_n(z) := \Big(1 + \dfrac{z}{n}\Big)^n$ pour tous $z \in \mathbb{C}$ et $n \in \mathbb{N}^*$.

   Dans l'exercice 11 de la page 519, nous démontrerons la formule suivante, concernant l'exponentielle $\exp(z)$ de z :

   $$\exp z \ = \ \lim_{n \to +\infty} \Big(1 + \frac{z}{n}\Big)^n. \tag{2}$$

   Ainsi la fonction $\exp : \mathbb{C} \to \mathbb{C}$ est limite simple de la suite $(f_n)_{n \geqslant 1}$.

3. Définissons une suite de fonctions $\big(f_n : [0, 1] \to \mathbb{R}\big)$ par $f_0 := 0$ et, pour tout $n \in \mathbb{N}$:

   $$f_{n+1}(x) \ := \ f_n(x) + \frac{1}{2}\big(x - f_n(x)^2\big) \quad \text{pour tout} \quad x \in I := [0, 1]. \tag{3}$$

   Étudions la convergence simple de la suite (f_n). Fixons $x \in I$ et soit $\varphi : I \to \mathbb{R}$ la fonction $t \mapsto t + (x - t^2)/2$. Elle croît de $x/2$ à $(x + 1)/2$, donc $\varphi(I) \subset I$. Posons $u_n := f_n(x)$ pour tout n. Ainsi $u_0 := 0$ et $u_{n+1} := \varphi(u_n)$ pour tout $n \in \mathbb{N}$. Il en résulte que (u_n) est une suite croissante de I, elle converge donc vers $\ell \in I$. Puisque φ est continue, ℓ est un point fixe de φ, d'où aussitôt $\ell := \sqrt{x}$. Ainsi la suite $\big(f_n(x)\big)$ converge vers $\sqrt{x}$. Conclusion : la suite (f_n) converge simplement vers la fonction $x \mapsto \sqrt{x}$.

Définition 2. On dit que la suite d'applications $(f_n : X \to E)$ converge *uniformément* vers une application $f : X \to E$ si la condition suivante est vérifiée :

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \geqslant N,\ \forall x \in X,\ \|f_n(x) - f(x)\| \leqslant \varepsilon. \tag{4}$$

On dit alors aussi que f est *limite uniforme* de la suite (f_n).

Par rapport à la condition (1), il est *essentiel* de noter que $(\forall x \in X)$ a changé de place : dans la condition (4), N *doit être indépendant de* x. Cette comparaison entre les conditions (1) et (4) montre en tous cas que, si (f_n) converge uniformément vers f, alors (f_n) converge simplement vers f : *la convergence uniforme entraîne la convergence simple*.
On peut interpréter cette définition à l'aide de la « norme de la convergence uniforme », qui généralise l'exemple 6 de la page 395. Pour toute application $f : X \to E$, posons :

$$\|f\|_\infty \ :\ \sup_{x \in X} \|f(x)\|. \tag{5}$$

Cette borne supérieure peut être égale à $+\infty$, et elle est finie si, et seulement si, f est *bornée*. Notons alors $\mathcal{B}(X, E)$ l'ensemble des applications bornées de X dans E, c'est bien sûr un sous-espace vectoriel de l'ensemble $\mathcal{F}(X, E)$ des applications de X dans E.

Proposition et définition 1. L'application $f \mapsto \|f\|_\infty$ de $\mathcal{B}(X, E)$ dans $\mathbb{R}$ est une norme sur $\mathcal{B}(X, E)$, on l'appelle *norme de la convergence uniforme*.

Démonstration. Seule l'inégalité triangulaire mérite d'être vérifiée.
Soient $f, g \in \mathcal{B}(X, E)$. Pour tout $x \in E$, on a $\|f(x) + g(x)\| \leqslant \|f(x)\| + \|g(x)\|$. Par définition d'une borne supérieure, on a $\|f(x)\| \leqslant \|f\|_\infty$ et $\|g(x)\| \leqslant \|g\|_\infty$, d'où $\|f(x) + g(x)\| \leqslant \|f\|_\infty + \|g\|_\infty$. Il suffit alors de « passer à la borne supérieure pour $x \in X$ » : $\|f\|_\infty + \|g\|_\infty$ est un majorant de $\big\{\|f(x) + g(x)\| \mid x \in X\big\}$, et le plus petit des majorants de cet ensemble est $\|f + g\|_\infty$, donc $\|f + g\|_\infty \leqslant \|f\|_\infty + \|g\|_\infty$. ∎

La convergence uniforme de (f_n) vers f signifie ceci : pour tout $\varepsilon > 0$, il existe un $N \in \mathbb{N}$ tel que l'inégalité $n \geqslant N$ entraîne $\|f_n - f\|_\infty \leqslant \varepsilon$. Autrement dit, $f_n - f$ est bornée pour n assez grand, et $\|f_n - f\|_\infty$ tend vers 0 quand n tend vers $+\infty$. D'où le :

Corollaire 2. La suite $(f_n : X \to E)$ converge uniformément vers $f : X \to E$ si, et seulement si, $\lim\limits_{n \to +\infty} \|f_n - f\|_\infty = 0$.

En particulier, si chaque f_n est bornée, la suite (f_n) converge uniformément si, et seulement si, c'est une suite convergente de l'e.v.n. $\big(\mathcal{B}(X, E), \|\cdot\|_\infty\big)$. Pour la nécessité, remarquer que la limite f de (f_n) est bornée car, si n est assez grand, $f_n - f$ et f_n sont bornées, donc f est bornée. Cela justifie la terminologie de « norme de la convergence uniforme ».

Exemples.

1. Reprenons l'exemple 1 de la page ci-contre, avec $X := [0, 1[$ et $f_n(x) := x^n$ pour tout $(n, x) \in \mathbb{N} \times X$. La suite (f_n) converge *simplement* vers la fonction nulle. La

convergence de (f_n) vers 0 *n'est pas uniforme* car $\|f_n - 0\|_\infty = \sup_{x\in X} x^n = 1$ pour tout n.

Considérons un réel $a \in]0,1[$ fixé, et soit $X_a := [0,a]$. On a $\sup_{x\in X_a} x^n = a^n$ pour tout n, ce qui montre que la convergence de (f_n) vers 0 est *uniforme sur* X_a. En conclusion, la convergence de (f_n) vers 0 *est uniforme sur* X_a *pour tout* a*, mais n'est pas uniforme sur* X*, qui est pourtant la réunion des* X_a. La notion de convergence uniforme est donc *relative à la partie sur laquelle on se place*.

2. Montrons que la suite de fonctions (f_n) définie sur $\mathbb{R}$ par $f_n(x) := \cos^n x \sin x$ converge uniformément vers 0. La convergence simple est évidente. En effet, soit $x \in \mathbb{R}$. Si $\sin x = 0$, on a $f_n(x) = 0$ pour tout n. Si $\sin x \neq 0$, on a $|\cos x| < 1$, donc $\lim_{n\to+\infty} \cos^n x = 0$. Dans les deux cas, la suite $\big(f_n(x)\big)$ converge vers 0.

   Pour la convergence uniforme, il faut travailler un peu plus. Soit $n \in \mathbb{N}^*$, calculons $\|f_n\|_\infty$ (il suffirait d'ailleurs de majorer cette norme). La fonction $|f_n|$ est paire, π-périodique et continue. Elle atteint donc son maximum absolu en un point $a_n \in \big]0,\pi/2\big[$, en lequel f_n' s'annule.

   Puisque $f_n'(x) = \cos^{n-1} x(\cos^2 x - n\sin^2 x)$, on a $\cos^2 a_n - n(1-\cos^2 a_n) = 0$, d'où $\cos a_n = \sqrt{n/(n+1)}$, puis $\sin a_n = 1/\sqrt{n+1}$. Dans ces conditions, il vient :

$$\|f_n\|_\infty \;=\; \cos^n a_n \sin a_n \;=\; \left(1+\frac{1}{n}\right)^{-n/2} \frac{1}{\sqrt{n+1}} \;\leqslant\; \frac{1}{\sqrt{n}}.$$

   Lorsque n tend vers $+\infty$, $\|f_n\|_\infty$ tend donc vers 0, en fait $\|f_n\|_\infty \sim 1/\sqrt{en}$. Le corollaire précédent prouve en tous cas que la suite de fonctions (f_n) converge uniformément vers 0.

Exercice 1.

Dans l'exemple 3 de la page 476, montrer que la suite (f_n) converge uniformément vers la fonction $f : x \mapsto \sqrt{x}$.

Solution. Montrons que (f_n) converge vers f dans $\big(\mathcal{C}(I,\mathbb{R}), \|\cdot\|_\infty\big)$, *i.e.* que la suite réelle $\big(\|f_n - f\|_\infty\big)$ converge vers 0. Si $n \in \mathbb{N}$ et $x \in I$, la formule (3) de la page 476 donne :

$$\sqrt{x} - f_{n+1}(x) = \sqrt{x} - f_n(x) - \big(x - f_n(x)^2\big)/2 = \big(\sqrt{x} - f_n(x)\big)\big[1 - \big(\sqrt{x} + f_n(x)\big)/2\big]$$

L'expression entre crochets est positive et majorée par $1 - \sqrt{x}$. Comme $\sqrt{x} - f_0(x) := \sqrt{x}$ on en déduit, par récurrence sur n, l'encadrement $0 \leqslant \sqrt{x} - f_n(x) \leqslant \sqrt{x}(1-\sqrt{x})^n$. Fixons $n \in \mathbb{N}$. Un calcul immédiat montre que, sur I, la fonction $t \mapsto t(1-t)^n$ atteint son maximum absolu en $t := 1/(n+1)$, ce maximum absolu valant $M_n := \dfrac{1}{n+1}\left(1 - \dfrac{1}{n+1}\right)^n$. L'encadrement ci-dessus montre que $\|f - f_n\|_\infty \leqslant M_n$. Lorsque n tend vers $+\infty$, M_n tend vers 0, en fait $M_n \sim 1/en$. *A fortiori* $\|f - f_n\|_\infty$ tend vers 0, d'où la conclusion.

Le théorème suivant est très important, il permet de voir si une suite d'applications converge uniformément, *lorsqu'on ne connaît pas à l'avance la limite simple de cette suite*. Nous donnons des exemples d'applications de ce théorème dans la section suivante.

Théorème 3 (critère de Cauchy uniforme). Supposons que E soit *complet*. Pour qu'une suite d'applications $(f_n : X \to E)$ converge uniformément, il faut, et il suffit, que pour tout $\varepsilon > 0$ il existe $N \in \mathbb{N}$ tel que les inégalités $p \geqslant N$ et $q \geqslant N$ entraînent :

$$\forall x \in X, \quad \big\|f_p(x) - f_q(x)\big\| \leqslant \varepsilon. \tag{6}$$

Démonstration. Supposons que (f_n) converge uniformément vers $f : X \to E$, et soit $\varepsilon > 0$. Il existe $N \in \mathbb{N}$ tel que $\|f_n(x) - f(x)\| \leqslant \varepsilon$ pour tout $n \geqslant N$ et tout $x \in X$. Soient p, q deux entiers tels que $p \geqslant N$ et $q \geqslant N$. Pour tout $x \in X$, on a $\|f_p(x) - f(x)\| \leqslant \varepsilon$ et $\|f_q(x) - f(x)\| \leqslant \varepsilon$, d'où $\|f_p(x) - f_q(x)\| \leqslant 2\varepsilon$ par inégalité triangulaire. La suite (f_n) vérifie donc le critère de Cauchy uniforme (*i.e.* la condition de l'énoncé).

Supposons inversement que (f_n) vérifie le critère de Cauchy uniforme. Soit x un point de X fixé. La suite $\big(f_n(x)\big)$ de E est de Cauchy. Puisque E est complet, cette suite converge, notons $f(x)$ sa limite. Nous définissons ainsi une application $f : X \to E$, qui est limite simple de la suite (f_n). Il reste à montrer que la convergence de (f_n) vers f est *uniforme*.

Soit $\varepsilon > 0$. Par hypothèse, il existe $N \in \mathbb{N}$ tel que la condition (6) soit vérifiée dès que $p \geqslant N$ et $q \geqslant N$. Fixons un entier $p \geqslant N$ et un point $x \in X$. Dans l'inégalité (6), faisons tendre q vers $+\infty$, de sorte que $f_q(x)$ tend vers $f(x)$. Compte tenu de la continuité de la norme, il vient : $\|f_p(x) - f(x)\| \leqslant \varepsilon$. Cette inégalité étant vraie pour tout $p \geqslant N$ et tout $x \in X$, la convergence uniforme de (f_n) vers f est établie. ∎

Avec la norme $\|\cdot\|_\infty$, le critère de Cauchy uniforme se traduit ainsi : pour tout $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que les inégalités $p \geqslant N$ et $q \geqslant N$ entraînent $\|f_p - f_q\|_\infty \leqslant \varepsilon$. Si les f_n sont bornées, cela signifie que (f_n) est une suite de Cauchy de l'e.v. n. $\big(\mathcal{B}(X, E), \|\cdot\|_\infty\big)$.

Corollaire 4. Si l'e.v. n. E est complet, l'e.v. n. $\big(\mathcal{B}(X, E), \|\cdot\|_\infty\big)$ l'est aussi.

Démonstration. Soit (f_n) une suite de Cauchy de $\big(\mathcal{B}(X, E), \|\cdot\|_\infty\big)$. Chaque f_n étant bornée, la remarque ci-dessus montre que (f_n) vérifie le critère de Cauchy uniforme. Puisque E est complet, le théorème 3 montre que la suite (f_n) converge uniformément, vers une application $f : X \to E$, elle aussi bornée. Ce qui précède montre alors que (f_n) converge vers f dans l'e.v. n. $\big(\mathcal{B}(X, E), \|\cdot\|_\infty\big)$. ∎

1.2 Convergence uniforme et continuité

Dans cette section, (X, d) est un espace métrique non vide et $(E, \|\cdot\|)$ est un e.v. n.. L'intérêt de la convergence uniforme est qu'elle conserve la continuité, nous allons le voir.

Théorème 5. Soit $(f_n : X \to E)$ une suite d'applications convergeant uniformément vers $f : X \to E$. Si chaque f_n est continue en un point $a \in X$, f est continue en a.

Démonstration. Soit $\varepsilon > 0$. Il existe $N \in \mathbb{N}$ tel que $\|f_n - f\|_\infty \leqslant \varepsilon$ dès que $n \geqslant N$. En particulier $\|f_N - f\|_\infty \leqslant \varepsilon$, c'est-à-dire $\|f_N(x) - f(x)\| \leqslant \varepsilon$ quel que soit $x \in X$. Puisque f_N est continue en a, il existe $\alpha > 0$ tel que $\big(d(x,a) \leqslant \alpha\big) \Longrightarrow \big(\|f_N(x) - f_N(a)\| \leqslant \varepsilon\big)$. Soit $x \in X$ tel que $d(x,a) \leqslant \alpha$. L'inégalité triangulaire donne :

$$\|f(x) - f(a)\| \leqslant \|f(x) - f_N(x)\| + \|f_N(x) - f_N(a)\| + \|f_N(a) - f(a)\| \leqslant \varepsilon + \varepsilon + \varepsilon = 3\varepsilon,$$

d'où la continuité de f en a. Il faut observer que c'est le fait que N ne dépende pas de x qui nous a permis, dans le second membre, de majorer les deux termes extrêmes par ε. ∎

Ce résultat peut être utilisé de façon « négative ». Ainsi, dans l'exemple 1 de la page 476, la limite simple f de la suite f_n n'est pas continue en 1, bien que les f_n soient continues sur $\mathbb{R}$. La convergence de (f_n) vers f n'est donc pas uniforme, que ce soit sur $\mathbb{R}$ ou sur $[0,1]$.

Dans la pratique, il arrive souvent que la convergence de (f_n) vers f soit uniforme non pas sur X, mais sur tout compact de X. C'est en fait suffisant pour assurer la continuité :

Théorème 6. Soit $(f_n : X \to E)$ une suite d'applications convergeant vers $f : X \to E$, *uniformément sur tout compact de* X. Si chaque f_n est continue en un point $a \in X$, f est continue en a.

Ce théorème résulte aussitôt du théorème précédent et du lemme que voici :

Lemme 7. Soient $f : X \to Y$ une application entre deux espaces métriques et a un point de X. Si la restriction de f à tout compact de X contenant a est continue en a, la fonction f est continue en a.

Démonstration. Soit (a_n) une suite de X convergeant vers a. Il suffit de montrer que $\big(f(a_n)\big)$ converge vers $f(a)$. Posons $Z := \big\{a_n \mid n \in \mathbb{N}\big\} \cup \{a\}$. Si nous montrons que Z est compact, nous aurons gagné, car alors, par hypothèse, la restriction g de f à Z sera continue, et donc $\big(g(a_n)\big) = \big(f(a_n)\big)$ convergera vers $g(a) = f(a)$.

Soit donc $(u_k)_{k\geqslant 0}$ une suite de Z. Il s'agit de montrer que cette suite possède une valeur d'adhérence ℓ dans Z. Si a est valeur d'adhérence de (u_k), $\ell := a$ convient. Si a n'est pas valeur d'adhérence de (u_k), il existe un réel $r > 0$ tel que l'ensemble $\big\{k \in \mathbb{N} \mid u_k \in B(a,r)\big\}$ soit fini (*cf.* la propriété suivant la définition 7 de la page 402). Il existe donc $A \in \mathbb{N}$ tel que $(k \geqslant A) \Longrightarrow \big(u_k \notin B(a,r)\big)$. D'un autre côté, la suite (a_n) convergeant vers a, il existe $N \in \mathbb{N}$ tel que $(n > N) \Longrightarrow \big(a_n \in B(a,r)\big)$. Pour tout entier $k \geqslant A$, u_k appartient donc à l'ensemble fini $S := \{a_0, \ldots, a_N\}$. Puisque S est compact, la suite $(u_k)_{k\geqslant A}$ possède une valeur d'adhérence $\ell \in S \subset Z$, et alors $\ell \in Z$ convient. ∎

Une application étant continue sur X si, et seulement si, elle est continue en chaque point de X, nous en déduisons une conséquence *capitale* du théorème 5 :

Théorème 8. Toute limite uniforme d'applications continues est continue.

Vu le théorème 6, il suffit que la limite soit *uniforme sur tout compact*. Voici deux résultats constituant des réciproques partielles du théorème précédent, ils permettent dans certains cas d'établir la convergence uniforme d'une suite de fonctions :

Théorème 9 (Premier théorème de Dini). Soient X un espace métrique compact (non vide) et $(f_n : X \to \mathbb{R})$ une suite de fonctions continues convergeant simplement vers une fonction continue f. On suppose que, pour tout $x \in X$, la suite $\big(f_n(x)\big)$ est décroissante. La suite (f_n) converge alors uniformément vers f.

Démonstration. La fonction f étant continue, on peut se ramener au cas où $f = 0$, en remplaçant $(f_n)_n$ par $(f_n - f)_n$. Supposons donc que $f = 0$. Soit $\varepsilon > 0$. Pour tout $n \in \mathbb{N}$, $F_n := f_n^{-1}\big([\varepsilon, +\infty[\big)$ est un fermé, donc un compact, de X. La suite $(F_n)_n$ est décroissante pour l'inclusion, car $(n \geqslant p) \implies (f_n \leqslant f_p)$. De plus, l'intersection des F_n est vide car, si x était un point de cette intersection, on aurait $f_n(x) \geqslant \varepsilon$ pour tout n, d'où $0 = \lim\limits_{n\to+\infty} f_n(x) \geqslant \varepsilon$, ce qui est absurde. D'après la proposition 49 de la page 435, il existe un $N \in \mathbb{N}$ tel que F_N soit vide. Considérons alors un entier $n \geqslant N$ et un point x de X. D'abord $x \notin F_n$, c'est-à-dire $f_n(x) < \varepsilon$. Ensuite, puisque la suite $\big(f_k(x)\big)_k$ converge vers 0 en décroissant, tous ses termes sont positifs, donc $0 \leqslant f_n(x) < \varepsilon$. Finalement $|f_n(x)| < \varepsilon$, ceci dès que $n \geqslant N$ et $x \in X$, ce qui prouve la convergence uniforme de $(f_n)_n$ vers 0. ∎

Noter que l'hypothèse de continuité de f est nécessaire, comme le montre l'exemple 1 de la page 476, avec $X := [0, 1]$.

Théorème 10 (Deuxième théorème de Dini). Soit (f_n) une suite de fonctions *croissantes* de $I := [a, b]$ dans $\mathbb{R}$ $(a < b)$ convergeant simplement vers une fonction f. Si f est *continue*, la convergence de (f_n) vers f est uniforme.

Démonstration. La fonction f est croissante : si $a \leqslant x \leqslant y \leqslant b$, on a $f_n(x) \leqslant f_n(y)$ pour tout n, d'où $f(x) \leqslant f(y)$, en faisant tendre n vers $+\infty$. Puisque f est continue et croissante, $f(I)$ est le segment $J := \big[f(a), f(b)\big]$. Soit maintenant $\varepsilon > 0$. Choisissons $p \in \mathbb{N}^*$ tel que $\big(f(b) - f(a)\big)/p \leqslant \varepsilon$. Partageons J en p intervalles égaux, à l'aide des points de subdivision $f(a) = y_0 < y_1 < \cdots < y_p := f(b)$. Pour tout $k \in [\![0, p]\!]$, il existe $x_k \in I$ tel que $y_k := f(x_k)$, et l'on a $a = x_0 < x_1 < \cdots < x_p = b$, car f est croissante. Pour chaque $k \in [\![0, p]\!]$, la suite $\big(f_n(x_k)\big)_{n\geqslant 0}$ converge vers y_k. Il existe donc N tel que $|f_n(x_k) - y_k| \leqslant \varepsilon$ pour tout $n \geqslant N$ et tout $k \in [\![0, p]\!]$. Soit alors $x \in I$. Il existe $k \in [\![0, p-1]\!]$ tel que $x_k \leqslant x \leqslant x_{k+1}$, de sorte que $y_k \leqslant f(x) \leqslant y_{k+1}$. Considérons un entier $n \geqslant N$. Si $h \in \{k, k+1\}$, on a $|f(x) - y_h| \leqslant y_{k+1} - y_k \leqslant \varepsilon$ et $|f_n(x_h) - y_h| \leqslant \varepsilon$, d'où $|f_n(x_h) - f(x)| \leqslant 2\varepsilon$. Ainsi $f_n(x_k)$ et $f_n(x_{k+1})$ appartiennent à $\big[f(x) - 2\varepsilon, f(x) + 2\varepsilon\big]$. Comme $f_n(x_k) \leqslant f_n(x) \leqslant f_n(x_{k+1})$, il en résulte que $f_n(x) \in \big[f(x)-2\varepsilon, f(x)+2\varepsilon\big]$, *i.e.* $|f_n(x) - f(x)| \leqslant 2\varepsilon$. Cette inégalité est vraie pour tout $x \in I$ et tout entier $n \geqslant N$, ce qui prouve que (f_n) converge uniformément vers f. ∎

Exemple. Soit $I := [0,1]$. Reprenons la suite de fonctions (f_n) définie dans l'exemple 3 de la page 476. Nous avons vu que $f_n(I) \subset I$ pour tout n. Chaque f_n est dérivable, et $f'_{n+1} = 1/2 + f'_n(1 - f_n)$. Comme $f_0 := 0$, f_n est croissante ($f'_n \geqslant 0$) pour tout $n \in \mathbb{N}$. D'après le deuxième théorème de Dini, la suite (f_n) converge uniformément vers $x \mapsto \sqrt{x}$. Nous retrouvons ainsi le résultat de l'exercice 1 de la page 478.

Théorème 11. Soient X un espace métrique compact (non vide) et E un e.v. n. complet. L'espace $\mathcal{C}(X,E)$ des applications continues de X dans E est un sous-espace vectoriel fermé de $\mathcal{B}(X,E)$. Ainsi $\mathcal{C}(X,E)$, muni de la norme $\|\cdot\|_\infty$, est un e.v. n. *complet*.

Démonstration. Si $f : X \to E$ est continue, l'application $x \mapsto \|f(x)\|$ de X dans $\mathbb{R}$ l'est aussi, puisque la norme est continue. Puisque X est compact, cette application est bornée, *i.e.* f est bornée. Ainsi $\mathcal{C}(X,E)$ est un sous-espace vectoriel de $\mathcal{B}(X,E)$.

D'après le corollaire 4 de la page 479, l'e.v. n. $\big(\mathcal{B}(X,E), \|\cdot\|_\infty\big)$ est complet. Pour conclure, en vertu du théorème 34 de la page 425, il suffit de montrer que $\mathcal{C}(X,E)$ est fermé dans $\mathcal{B}(X,E)$. Soit donc (f_n) une suite de $\mathcal{C}(X,E)$ convergeant vers $f \in \mathcal{B}(X,E)$. D'après le corollaire 2 de la page 477, (f_n) est une suite d'applications continues convergeant uniformément, sa limite f est donc continue, *i.e.* $f \in \mathcal{C}(X,E)$. D'où la conclusion. ∎

Ainsi, pour tout segment $[a,b]$ de $\mathbb{R}$, les espaces vectoriels $\mathcal{C}([a,b],\mathbb{R})$ et $\mathcal{C}([a,b],\mathbb{C})$ sont complets, lorsqu'on les munit de la norme $\|\cdot\|_\infty$.

Théorème 12 (de la double limite). On suppose que E est *complet*. Soient I un intervalle non trivial de $\mathbb{R}$ et c une extrémité (finie ou non) de I. Soit $(f_n : I \to E)$ une suite d'applications convergeant uniformément vers $f : I \to E$. On suppose que, pour tout $n \in \mathbb{N}$, $f_n(x)$ a une limite $\ell_n \in E$ lorsque $x \in I$ tend vers c. Alors la suite (ℓ_n) de E a une limite ℓ, et $f(x)$ tend vers ℓ lorsque $x \in I$ tend vers c. Ainsi :

$$\lim_{n\to+\infty} \left(\lim_{x\to c} f_n(x) \right) = \lim_{x\to c} \left(\lim_{n\to+\infty} f_n(x) \right). \tag{7}$$

Démonstration. Il faut bien comprendre cette formule d'interversion des limites, que l'on peut écrire $\lim\limits_{n\to+\infty} \ell_n = \lim\limits_{x\to c} f(x)$. Le théorème affirme *d'abord* que chacune de ces deux limites existe, ce qui n'est pas évident, et *ensuite* que ces deux limites sont égales.

Si $c \in I$, la conclusion vient du théorème 5 de la page 480. En effet, pour tout n, la limite ℓ_n de f_n en c est nécessairement $f_n(c)$, et la suite $\big(f_n(c)\big)$ converge vers $f(c)$. Supposons par exemple que $I :=]a,c[$, où $c \in \mathbb{R} \cup \{+\infty\}$. Montrons que (ℓ_n) est une suite de Cauchy de E. Soit $\varepsilon > 0$. Il existe $N \in \mathbb{N}$ tel que $\|f_p(x) - f_q(x)\| \leqslant \varepsilon$ pour tous $p \geqslant N$, $q \geqslant N$ et $x \in I$. Fixons deux entiers p, q tels que $p \geqslant N$, $q \geqslant N$. Dans l'inégalité $\|f_p(x) - f_q(x)\| \leqslant \varepsilon$, faisons tendre $x \in I$ vers c. Puisque $f_p(x)$ et $f_q(x)$ tendent vers ℓ_p et ℓ_q respectivement, il vient : $\|\ell_p - \ell_q\| \leqslant \varepsilon$. La suite (ℓ_n) de E est donc de Cauchy. Puisque E est complet, cette suite converge, soit $\ell \in E$ sa limite.

Soient ε et N comme ci-dessus. D'abord $\|\ell_N - \ell\| \leqslant \varepsilon$ et $\|f_N(x) - f(x)\| \leqslant \varepsilon$ pour tout $x \in I$. Ensuite, dès que $x \in I$ est assez voisin de c, $\|f_N(x) - \ell_N\| \leqslant \varepsilon$, d'où $\|f(x) - \ell\| \leqslant 3\varepsilon$, par inégalité triangulaire. Cela montre que ℓ est la limite de $f(x)$ lorsque $x \in I$ tend vers c. ∎

Exemple. Posons $f_n(x) := \dfrac{n}{n+e^x}$ pour tout $(n,x) \in \mathbb{N}\times\mathbb{R}$. Si x est fixé, $\big(f_n(x)\big)$ converge vers 1. Si n est fixé, $f_n(x)$ tend vers 0 lorsque x tend vers $+\infty$. La convergence de (f_n) vers 1 n'est donc pas uniforme. Le lecteur pourra le vérifier directement et montrer que (f_n) converge uniformément vers 1 sur $]-\infty,a[$, ceci pour tout $a\in\mathbb{R}$ fixé.

1.3 Intégration et dérivation

Voici deux théorèmes fondamentaux, portant sur l'intégration ou la dérivation d'une suite de fonctions. Ces théorèmes seront illustrés dans la section suivante.

Théorème 13. Si $(g_n : [a,b] \to \mathbb{R})$ est une suite de fonctions intégrables qui converge uniformément vers une fonction g, alors g est intégrable sur $[a,b]$, et l'on a :

$$\int_a^b g \;=\; \lim_{n\to+\infty}\int_a^b g_n. \tag{8}$$

Ce théorème sera démontré dans le module II.3, nous l'admettrons donc ici. Observons cependant que, si chaque g_n est continue, le résultat est évident. D'abord, g est continue, en vertu du théorème 8 de la page 480. Soit $\varepsilon > 0$. À partir d'un certain rang N, on a $\|g-g_n\|_\infty \leqslant \varepsilon$, c'est-à-dire $|g(x)-g_n(x)| \leqslant \varepsilon$ pour tout $x\in[a,b]$. Dès que $n \geqslant N$, il vient :

$$\left|\int_a^b g(x)\,\mathrm{d}x - \int_a^b g_n(x)\,\mathrm{d}x\right| = \left|\int_a^b \big(g(x)-g_n(x)\big)\,\mathrm{d}x\right| \leqslant \int_a^b |g(x)-g_n(x)|\,\mathrm{d}x \leqslant \varepsilon(b-a),$$

ce qui établit l'assertion (8) dans ce cas.
La dérivation étant « inverse » de l'intégration, on en déduit un théorème de dérivation :

Théorème 14. Soient I un intervalle de $\mathbb{R}$ et $(f_n : I\to\mathbb{R})$ une suite de fonctions de classe $\mathcal{C}^1$. On fait les hypothèses suivantes :
a) Il existe un point $a\in I$ tel que la suite $\big(f_n(a)\big)$ converge.
b) La suite (f_n') converge, uniformément sur I, vers une fonction $g : I\to\mathbb{R}$.
La suite (f_n) converge alors, uniformément *sur tout intervalle borné* $J\subset I$, vers une fonction $f : I\to\mathbb{R}$. De plus, f est de classe $\mathcal{C}^1$ sur I et $f'=g$.

Démonstration. Puisque chaque f_n' est continue sur I, g est continue sur I, comme limite uniforme de fonctions continues. Soit $G : I\to\mathbb{R}$ la primitive de g nulle en a. Par ailleurs, notons ℓ la limite de la suite $\big(f_n(a)\big)$. Soit $x\in I$. Si $x > a$, appliquons le théorème précédent, sur le segment $[a,x]$, à la suite (f_n'). Il vient :

$$G(x) \;:=\; \int_a^x g \;=\; \lim_{n\to+\infty}\int_a^x f_n' \;=\; \lim_{n\to+\infty}\big[f_n(x)-f_n(a)\big].$$

Il en résulte que la suite $\big(f_n(x)\big)$ converge vers $\ell+G(x)$. On voit immédiatement que ce résultat est encore vrai lorsque $x\leqslant a$. Ainsi la suite de fonctions (f_n) converge simplement vers $f := \ell+G$. En particulier, f est de classe $\mathcal{C}^1$ sur I, et $f'=G'=g$.

Soient $J \subset I$ un intervalle borné de longueur L et $x_0 \in J$. Étant donné un réel $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que, pour tout entier $n \geqslant N$, on ait d'une part $|f(x_0) - f_n(x_0)| \leqslant \varepsilon$ et d'autre part $|g - f_n'| \leqslant \varepsilon$ sur I. Pour tout $n \geqslant N$ et tout $x \in J$, on a :

$$\begin{aligned}\big|\big[f(x) - f_n(x)\big] - \big[f(x_0) - f_n(x_0)\big]\big| &= \left|\int_{x_0}^{x} \big(f'(t) - f_n'(t)\big)\,\mathrm{d}t\right| \\ &= \left|\int_{x_0}^{x} \big(g(t) - f_n'(t)\big)\,\mathrm{d}t\right| \\ &\leqslant |x - x_0| \sup_{t\in I} \big|g(t) - f_n'(t)\big| \leqslant \varepsilon L.\end{aligned}$$

D'où $|f(x) - f_n(x)| \leqslant \varepsilon L + |f(x_0) - f_n(x_0)| \leqslant \varepsilon(L+1)$, ceci pour tout $n \geqslant N$ et tout $x \in J$. La suite $(f_n)_n$ converge donc vers f uniformément sur J. ∎

Il est clair que l'hypothèse b) peut être remplacée par la suivante, parfois plus facile à vérifier : la suite (f_n') converge vers g *uniformément sur tout compact de* I. Dans la conclusion, la convergence de (f_n) vers f est alors *uniforme sur tout compact de* I.

Il est essentiel de noter que l'hypothèse b) porte sur la suite (f_n'). Même si (f_n) converge vers f uniformément et si chaque f_n est de classe $\mathcal{C}^1$, f n'est pas forcément de classe $\mathcal{C}^1$. Posons ainsi $f_n(x) := \sqrt{x^2 + 2^{-n}}$ pour tout $(n, x) \in \mathbb{N} \times \mathbb{R}$. Chaque f_n est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$. Si x est fixé, la suite $\big(f_n(x)\big)$ converge vers $\sqrt{x^2} = |x|$. En fait, (f_n) converge vers $f : x \mapsto |x|$ uniformément car, pour tous n, x, on a :

$$0 \leqslant f_n(x) - |x| = \sqrt{x^2 + 2^{-n}} - |x| = \frac{2^{-n}}{\sqrt{x^2 + 2^{-n}} + |x|} \leqslant \frac{2^{-n}}{\sqrt{x^2 + 2^{-n}}} \leqslant 2^{-n/2}.$$

Cependant, la fonction f n'est pas dérivable en 0.

Remarque. Ces deux théorèmes restent valables pour des fonctions à valeurs dans $\mathbb{R}^p$ (on raisonne composante par composante). Le théorème 14 reste vrai pour des fonctions à valeurs dans un $\mathbb{R}$-espace vectoriel E de dimension finie $p \geqslant 1$, voyons pourquoi.

Qu'entend-on d'abord par dérivée de $f : I \to E$ en $x_0 \in I$? C'est bien sûr la limite, lorsque $x \in I$ tend vers x_0 (et $x \neq x_0$) de $\dfrac{1}{x - x_0}\big(f(x) - f(x_0)\big)$ (les scalaires sont à gauche des vecteurs). Pour parler de limite (et de continuité), il faut une norme sur E. Mais, toutes les normes sur E étant équivalentes, l'existence (et la valeur) de cette limite ne dépendent pas de la norme choisie. De même, le fait que la suite (f_n') converge uniformément vers g ne dépend pas de la norme choisie sur E. L'énoncé du théorème 14 garde donc un sens.

Pour *prouver* ce théorème ainsi généralisé, on se ramène au cas $E := \mathbb{R}^p$ en choisissant une base $(e_1, \ldots, e_p)$ de E et en posant $\|x\| := (x_1^2 + \cdots + x_p^2)^{1/2}$ si $x := x_1 e_1 + \cdots + x_p e_p \in E$.

Terminons cette section par un résultat important sur les « séries doubles » :

Théorème 15 (Fubini pour les séries absolument convergentes).
Soit $(a_{n,p})_{(n,p)\in\mathbb{N}^2}$ une famille de nombres complexes. On suppose que :

a) pour tout $n \in \mathbb{N}$ fixé, la série $\sum_{p\geqslant 0} |a_{n,p}|$ converge, on note s_n sa somme ;

b) la série $\sum_{n\geqslant 0} s_n$ converge.

Sous ces hypothèses, on dit que la série double est *absolument convergente*. Alors :

1. Pour tout $p \in \mathbb{N}$ fixé, la série $\sum_{n\geqslant 0} |a_{n,p}|$ converge, on note t_p sa somme.
2. La série $\sum_{p\geqslant 0} t_p$ converge.
3. On a l'égalité suivante (avec convergence de toutes les séries qui interviennent) :

$$\sum_{n=0}^{+\infty}\left(\sum_{p=0}^{+\infty} a_{n,p}\right) = \sum_{p=0}^{+\infty}\left(\sum_{n=0}^{+\infty} a_{n,p}\right). \tag{9}$$

Démonstration. Commençons par une remarque générale. Soit $\sum_{n\geqslant 0} u_n$ une série à termes réels. Soit $f : I := [0,+\infty[\to \mathbb{R}$ la fonction valant u_n sur chaque intervalle $[n, n+1[$, $n \in \mathbb{N}$.
On montre, en appliquant le corollaire 103 de la page 614 (à $|f|$ et aux intervalles $J_n := [0,n]$) que la fonction f est *absolument* intégrable au sens de Riemann généralisé sur I si, et seulement si, la série donnée converge absolument. Dans ce cas, f est intégrable au sens de Riemann généralisé sur I et $\sum_{n=0}^{+\infty} u_n = \int_0^{+\infty} f$.

Soit $M := \sum_{n=0}^{+\infty} s_n$. Pour tous entiers $N, P \in \mathbb{N}$, on a d'abord :

$$\sum_{p=0}^{P}\left(\sum_{n=0}^{N}|a_{n,p}|\right) = \sum_{n=0}^{N}\left(\sum_{p=0}^{P}|a_{n,p}|\right) \leqslant M. \tag{$*$}$$

En effet, puisque les s_n sont positifs, on a $\sum_{n=0}^{N} s_n \leqslant M$. Le même argument donne $\sum_{p=0}^{P}|a_{n,p}| \leqslant s_n$ pour tout $n \in \mathbb{N}$. On en déduit l'inégalité $(*)$, l'égalité étant banale puisqu'il s'agit de sommes finies.
Soit $p \in \mathbb{N}$ fixé. Pour tout $N \in \mathbb{N}$, $(*)$ implique en particulier $\sum_{n=0}^{N}|a_{n,p}| \leqslant M$, ce qui prouve la convergence de la série $\sum_{n\geqslant 0}|a_{n,p}|$. Ensuite, dans l'inégalité $\sum_{p=0}^{P}\left(\sum_{n=0}^{N}|a_{n,p}|\right) \leqslant M$, fixons P et faisons tendre N vers $+\infty$. Il vient $\sum_{p=0}^{P} t_p \leqslant M$, d'où l'assertion 2. Il reste à prouver l'égalité (9).
En séparant parties réelles et parties imaginaires, on ramène au cas où les $a_{n,p}$ sont réels. Si $n \in \mathbb{N}$, notons $f_n : [0,+\infty[\to \mathbb{R}$ la fonction valant $a_{n,p}$ sur chaque intervalle $[p, p+1[$, $p \in \mathbb{N}$. D'après la remarque du début, f_n est absolument intégrable ; de plus $\int_0^{+\infty}|f_n| = s_n$ et $\int_0^{+\infty} f_n = \sum_{p=0}^{+\infty} a_{n,p}$.
Les séries de fonctions $\sum_{n\geqslant 0} f_n$ et $\sum_{n\geqslant 0}|f_n|$ convergent simplement. En effet soit $x \in [0,+\infty[$; on a $x \in [p, p+1[$ pour un unique $p \in \mathbb{N}$. Alors $\sum_{n\geqslant 0}|f_n(x)| = \sum_{n\geqslant 0}|a_{n,p}|$, qui converge, vu l'assertion 1.

On a $F(x) := \sum_{n\geqslant 0} f_n(x) = \sum_{n\geqslant 0} a_{n,p}$ et la fonction F est *constante* sur les intervalles $[p, p+1[$, donc *localement intégrable* sur I. Puisque la série $\sum_{n\geqslant 0} s_n$ converge, $\sum_{n=0}^{+\infty}\left(\sum_{p=0}^{+\infty} a_{n,p}\right)$ converge et la proposition 125 de la page 639 donne :

$$\sum_{n=0}^{+\infty}\left(\sum_{p=0}^{+\infty} a_{n,p}\right) = \sum_{n=0}^{+\infty}\left(\int_0^{+\infty} f_n\right) = \int_0^{+\infty}\left(\sum_{n=0}^{+\infty} f_n\right).$$

Si x et p sont comme ci-dessus, $\sum_{n=0}^{+\infty} f_n(x) = \sum_{n=0}^{+\infty} a_{n,p}$. La remarque du début appliquée à la fonction F montre alors que $\int_0^{+\infty}\left(\sum_{n=0}^{+\infty} f_n\right) = \int_0^{+\infty} F = \sum_{p=0}^{+\infty}\left(\sum_{n=0}^{+\infty} a_{n,p}\right)$, et l'égalité (9) en résulte. ∎

Théorème 16 (Fubini pour les séries à termes positifs). Supposons que les termes de la famille $(a_{n,p})_{(n,p)\in\mathbb{N}^2}$ soient des réels positifs ou nuls. Alors l'égalité (9) est encore vraie, que la série double converge ou diverge vers $+\infty$.

Ce théorème signifie que, ou bien les deux membres de (9) sont finis ou égaux, ou bien les deux membres sont infinis.

Démonstration. Si la série double est absolument convergente, il suffit d'appliquer le théorème précédent.

Si la série double n'est pas absolument convergente, le membre de gauche de (9) diverge par hypothèse, puisque les termes sont positifs. Si le membre de droite était convergent, alors, en échangeant les rôles de n et p dans le théorème 15 de la page précédente, on verrait que le membre de gauche est convergent, ce qui est faux. Le membre de droite est donc lui aussi divergent. ∎

Exercice 2.

Soit $z \in \mathbb{C}$ tel que $|z| < 1$. Établir l'égalité $\displaystyle\sum_{n=1}^{+\infty} \frac{z^n}{1-z^{2n}} = \sum_{n=0}^{+\infty} \frac{z^{2n+1}}{1-z^{2n+1}}$.

Solution. Si $n \geqslant 1$ est fixé, $(1-z^{2n})^{-1} = \sum_{p=0}^{+\infty} z^{2np}$. Appliquons le théorème précédent en posant $a_{n,p} := z^{n(2p+1)}$ pour tout $(n,p) \in \mathbb{N}^* \times \mathbb{N}$. Si $n \geqslant 1$ est fixé, la série $\sum_{p\geqslant 0} |a_{n,p}|$ converge absolument (c'est une série géométrique de raison $|z|^{2n}$), et sa somme vaut :

$$s_n = \sum_{p=0}^{+\infty} |z|^{n(2p+1)} = \frac{|z|^n}{1-|z|^{2n}}.$$

Lorsque n tend vers $+\infty$, $s_n \sim |z|^n$, ce qui montre que la série $\sum_{n\geqslant 1} s_n$ converge. Nous sommes donc en droit d'appliquer le théorème.

Si $n \geqslant 1$ est fixé, $\sum_{p=0}^{+\infty} a_{n,p} = z^n(1-z^{2n})^{-1}$. Ainsi $\sum_{n=1}^{+\infty}\left(\sum_{p=0}^{+\infty} a_{n,p}\right)$ est le premier membre de l'égalité à établir. Si $p \geqslant 0$ est fixé, la série $\sum_{n\geqslant 1} a_{n,p}$ est aussi une série géométrique, dont la raison et le premier terme valent z^{2p+1}. La somme de cette série est donc $z^{2p+1}(1-z^{2p+1})^{-1}$. Ainsi $\sum_{p=0}^{+\infty}\left(\sum_{n=1}^{+\infty} a_{n,p}\right)$ est le deuxième membre de l'égalité à établir. D'où la conclusion.

1.4 Théorème d'approximation de Weierstraß

Il est important de pouvoir approcher une fonction donnée par des fonctions plus simples, ou suffisamment régulières. Le théorème suivant montre que toute fonction réelle continue sur un segment peut être approchée *uniformément*, d'aussi près qu'on veut, par une fonction polynôme. Une autre démonstration de ce théorème sera donnée dans le module II.4.

Théorème 17 (Weierstraß). Toute fonction réelle continue sur un segment $I := [a,b]$ $(a<b)$ est limite uniforme d'une suite de fonctions polynômes.

Le théorème affirme que l'ensemble $\mathcal{P}$ des fonctions polynômes sur I est *dense* dans l'e.v. n. $\big(\mathcal{C}(I,\mathbb{R}), \|\cdot\|_\infty\big)$, c'est-à-dire que l'adhérence $\overline{\mathcal{P}}$ de $\mathcal{P}$ dans $\mathcal{C}(I,\mathbb{R})$ est égale à $\mathcal{C}(I,\mathbb{R})$. Puisque $\mathcal{P}$ est un sous-espace vectoriel de $\mathcal{C}(I,\mathbb{R})$, c'est vrai aussi pour $\overline{\mathcal{P}}$ (*cf.* la proposition 2 de la page 401). En tous cas, voici ce que nous devons prouver :

$$\forall f \in \mathcal{C}(I,\mathbb{R}),\ \forall \varepsilon > 0,\ \exists P \in \mathcal{P},\ \|f-P\|_\infty \leqslant \varepsilon.$$

Démonstration. Nous allons procéder en trois étapes.

1. Soit $\mathcal{A}$ l'ensemble des fonctions de I dans $\mathbb{R}$ qui sont continues et *affines par morceaux*. Nous allons prouver l'inclusion $\mathcal{A} \subset \overline{\mathcal{P}}$. Considérons d'abord une subdivision de I, $\sigma : a = x_0 < x_1 < \cdots < x_p = b$. Si $\mathbf{y} := (y_0,\dots,y_p) \in \mathbb{R}^{p+1}$, soit $g_{\mathbf{y}} : I \to \mathbb{R}$ la fonction dont le graphe est la réunion des segments $[A_k, A_{k+1}]$, $k = 0,\dots,p-1$, en posant $A_h := (x_h, y_h)$ pour tout h. Lorsque σ est fixée, l'application $\mathbf{y} \mapsto g_{\mathbf{y}}$ est un isomorphisme de $\mathbb{R}^{p+1}$ sur un sous-espace vectoriel $\mathcal{A}_\sigma$ de $\mathcal{C}(I,\mathbb{R})$. Par définition, $\mathcal{A}$ est la réunion des $\mathcal{A}_\sigma$ pour toutes les subdivisions σ de I. Il suffit donc de prouver l'inclusion $\mathcal{A}_\sigma \subset \overline{\mathcal{P}}$ lorsque la subdivision σ (notée comme ci-dessus) est fixée.

   Si $k \in [\![0,p]\!]$, la restriction φ_k de la fonction $x \mapsto |x-x_k|$ à I appartient à $\mathcal{A}_\sigma$. De plus la famille $(\varphi_0,\dots,\varphi_p)$ est libre. Cela résulte du fait que, si k est donné, φ_k est dérivable en tout point de I sauf en x_k. Comme $\dim \mathcal{A}_\sigma = p+1$, $(\varphi_0,\dots,\varphi_p)$ est une base de $\mathcal{A}_\sigma$. Puisque $\overline{\mathcal{P}}$ est un sous-espace vectoriel de $\mathcal{C}(I,\mathbb{R})$, il suffit, pour prouver l'inclusion $\mathcal{A} \subset \overline{\mathcal{P}}$, de montrer que, si $c \in I$, la restriction ψ de $x \mapsto |x-c|$ à I appartient à $\overline{\mathcal{P}}$.

2. Soit $\varepsilon > 0$. Choisissons $\lambda > 0$ tel que $I := [a,b] \subset [c-\lambda^{-1}, c+\lambda^{-1}]$.

D'après l'exemple de la page 482, il existe $P \in \mathbb{R}[X]$ tel que $\left|\sqrt{t} - P(t)\right| \leqslant \varepsilon\lambda$ pour tout $t \in [0,1]$. En remplaçant t par t^2, on a $\left||t| - P(t^2)\right| \leqslant \varepsilon\lambda$ pour tout $t \in [-1,1]$. Remplaçons t par $\lambda(x-c)$. Dès que x vérifie $|x-c| \leqslant \lambda^{-1}$, on a $||x-c| - Q(x)| \leqslant \varepsilon$, où Q est le polynôme $\lambda^{-1}P\big(\lambda^2(X-c)^2\big)$. En particulier, $||x-c| - Q(x)| \leqslant \varepsilon$ pour tout $x \in I$: nous avons approché uniformément ψ, à ε près, par une fonction polynôme. D'où $\mathcal{A} \subset \overline{\mathcal{P}}$.

3. Soient $f \in \mathcal{C}(I,\mathbb{R})$ et $\varepsilon > 0$. La fonction f étant *uniformément continue* (théorème 56 de la page 439), il existe $\alpha > 0$ tel que, pour tous $x, y \in I$, $\big(|x-y| \leqslant \alpha\big)$ implique $\big(|f(x) - f(y)| \leqslant \varepsilon\big)$. Soit $p \in \mathbb{N}^*$ tel que $(b-a)/p \leqslant \alpha$. Soient σ la subdivision régulière de I, $\sigma : a = x_0 < x_1 < \cdots < x_p = b$, et $\mathbf{y} := \big(f(x_0), \ldots, f(x_p)\big)$. Soit $x \in I$. Il existe $k \in [\![0, p-1]\!]$ tel que $x \in [x_k, x_{k+1}]$. Le segment $\big[f(x) - \varepsilon, f(x) + \varepsilon\big]$ contient $f(x_k)$ et $f(x_{k+1})$. Puisque $g_{\mathbf{y}}(x)$ est par construction compris entre $f(x_k)$ et $f(x_{k+1})$, on en déduit $|g_{\mathbf{y}}(x) - f(x)| \leqslant \varepsilon$. Ainsi $\|f - g_{\mathbf{y}}\|_\infty \leqslant \varepsilon$.

   Puisque $\mathcal{A} \subset \overline{\mathcal{P}}$, il existe $g \in \mathcal{P}$ telle que $\|g_{\mathbf{y}} - g\|_\infty \leqslant \varepsilon$. Alors $\|f - g\|_\infty \leqslant 2\varepsilon$. Pour toute fonction $f \in \mathcal{C}(I,\mathbb{R})$ et tout $\varepsilon > 0$, nous avons donc montré qu'on pouvait approcher uniformément f, à 2ε près, par une fonction polynôme. Le théorème est ainsi établi. ∎

Dans le théorème de Weierstraß, il *faut* que l'intervalle soit compact. Par exemple, soit f la fonction $x \mapsto 1/x$, définie sur $I :=]0,1]$. Supposons qu'il existe une suite (f_n) de fonctions polynômes sur $\mathbb{R}$ dont les restrictions à I convergent uniformément vers f. Chaque f_n est bornée sur $[0,1]$, donc sur I. Puisque f est limite uniforme de (f_n) sur I, f est bornée sur I, ce qui est évidemment absurde.

Exercice 3.

Soit $I := [0,1]$.

Montrer que $f \mapsto \|f\| := \sup_{n\in\mathbb{N}} \left|\int_0^1 f(x)x^n \,\mathrm{d}x\right|$ est une norme sur $\mathcal{C}(I,\mathbb{R})$.

Solution. Si $f \in \mathcal{C}(I,\mathbb{R})$, $\|f\|$ est un réel. En effet, pour tout $n \in \mathbb{N}$, $\left|\int_0^1 f(x)x^n \,\mathrm{d}x\right|$ est majorée par $\|f\|_\infty \int_0^1 x^n \,\mathrm{d}x = \|f\|_\infty/(n+1)$, donc par $\|f\|_\infty$. L'application $\|\cdot\|$ vérifie clairement la propriété d'homogénéité et l'inégalité triangulaire. Le seul point non évident est la séparation : $(\|f\| = 0) \Longrightarrow (f = 0)$. Soit donc $f \in \mathcal{C}(I,\mathbb{R})$ telle que $\|f\| = 0$. On a donc $\int_0^1 f(x)x^n \,\mathrm{d}x = 0$ pour tout $n \in \mathbb{N}$, d'où $\int_0^1 f(x)P(x) \,\mathrm{d}x = 0$ pour tout $P \in \mathbb{R}[X]$. D'après le théorème de Weierstraß, f est limite uniforme d'une suite (g_k) de fonctions polynômes. Puisque f est bornée, f^2 est alors limite uniforme de (fg_k). Pour chaque k, $\int_0^1 fg_k = 0$. D'après le théorème 13 de la page 483, $\int_0^1 f^2 = 0$, d'où $f^2 = 0$ puisque f^2 est continue et positive. En conclusion, $f = 0$.

2 Séries de fonctions

Dans cette section, X est un ensemble non vide et E est un e.v. n.. L'étude d'une série de fonctions se ramène en principe à celle d'une suite de fonctions (ses sommes partielles). Il est cependant utile de formuler séparément les résultats portant sur ces séries, d'autant plus que l'on dispose pour elles d'un outil supplémentaire puissant, la « convergence normale ».

2.1 Passage des suites de fonctions aux séries de fonctions

Par définition, une série $\sum_{n\geqslant 0} u_n$, dans laquelle les u_n appartiennent à E, est convergente si, et seulement si, la suite (S_n) de ses sommes partielles converge. Dans ce cas, la somme de la série est la limite de la suite (S_n), ce que l'on écrit : $\sum_{n=0}^{+\infty} u_n := \lim_{n\to+\infty} S_n$.

Une série $\sum_{n\geqslant 0} f_n$, dans laquelle chaque f_n est une application de X dans E, est appelée une *série d'applications*, ou *série de fonctions* lorsque E est égal à $\mathbb{R}$ ou $\mathbb{C}$. Chaque somme partielle $S_n := \sum_{k=0}^{n} f_k$ est une application de X dans E. Pour la suite (S_n), il y a plusieurs modes de convergence : convergence simple, uniforme, ou uniforme sur tout compact.

Nous dirons que la série d'applications $\sum_{n\geqslant 0} f_n$ *converge simplement* si la suite d'applications (S_n) converge simplement. Cela revient à dire que, pour tout $x \in X$, la série $\sum_{n\geqslant 0} f_n(x)$ converge. Si tel est le cas, la limite de la suite (S_n) est appelée *somme* de la série $\sum_{n\geqslant 0} f_n$, c'est l'application $x \mapsto \sum_{n=0}^{+\infty} f_n(x)$ de X dans E.

De même, $\sum_{n\geqslant 0} f_n$ *converge uniformément* si la suite d'applications (S_n) converge uniformément, et pareillement pour la convergence uniforme sur tout compact. Voici par exemple la traduction, pour une série de fonctions, du théorème 6 de la page 480 :

Théorème 18. Soit X un espace métrique. La somme d'une série d'applications continues convergeant uniformément sur tout compact de X est continue.

Proposition 19. Pour qu'une série $\sum_{n\geqslant 0} f_n$ d'applications de X dans E converge uniformément, il faut, et il suffit, qu'elle converge simplement et que la suite (R_n) de ses restes converge uniformément vers 0.
Par définition, si $n \in \mathbb{N}$ et $x \in X$, $R_n(x) := \sum_{k=n+1}^{+\infty} f_k(x)$.

Démonstration. On peut supposer que $\sum_{n\geqslant 0} f_n$ converge simplement, notons f sa somme. Pour tout $x \in X$, on a $f(x) := \sum_{n=0}^{+\infty} f_n(x) = S_n(x) + R_n(x)$. Vu la définition 2 de la page 477, (S_n) converge uniformément, nécessairement vers f, si, et seulement si, la suite $(f - S_n) = (R_n)$ converge uniformément vers 0. ∎

Voici une application simple de ce résultat :

Corollaire 20 (critère spécial « uniforme » des séries alternées). Soit $(f_n : X \to \mathbb{R}_+)$ une suite d'applications convergeant uniformément vers 0. On suppose que, pour tout $x \in X$, la suite $\big(f_n(x)\big)$ est décroissante. Alors la série de fonctions $\sum_{n\geqslant 0}(-1)^n f_n$ converge uniformément.

Démonstration. Soit $x \in X$. D'après le critère spécial « ordinaire » des séries alternées, la série $\sum_{n\geqslant 0}(-1)^n f_n(x)$ converge et, si $n \in \mathbb{N}$, le reste $R_n(x) := \sum_{k=n+1}^{+\infty}(-1)^k f_k(x)$ est majoré en valeur absolue par $f_{n+1}(x)$. Il suffit alors d'appliquer la proposition. ∎

Prenons par exemple $X := [0,+\infty]$ et $f_n(x) := 1/(n+x)$ si $n \in \mathbb{N}^*$ et $x \geqslant 0$. Si x est fixé, la suite $\big((n+x)^{-1}\big)_{n\geqslant 1}$ décroît. Pour tout $n \in \mathbb{N}^*$, on a évidemment $\|f_n\|_\infty = 1/n$, de sorte que la suite de fonctions (f_n) converge uniformément vers 0. D'après le corollaire, la série de fonctions $\displaystyle\sum_{n\geqslant 1}\frac{(-1)^n}{n+x}$ converge donc uniformément sur $[0,+\infty[$.

Lorsqu'on applique le critère de Cauchy uniforme à la suite (S_n) des sommes partielles d'une série d'applications $\sum_{n\geqslant 0} f_n$, on obtient le critère suivant :

Théorème 21 (critère de Cauchy uniforme). Supposons que l'e.v.n. E soit *complet*. Une série $\sum_{n\geqslant 0} f_n$ d'applications de X dans E converge uniformément si, et seulement si, pour tout $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que les inégalités $q > p \geqslant N$ entraînent :

$$\forall x \in X, \quad \left\| \sum_{n=p}^{q} f_n(x) \right\| \leqslant \varepsilon. \tag{10}$$

Nous donnerons dans la section suivante un outil très puissant (la convergence normale) pour établir la convergence uniforme d'une série de fonctions. Lorsque cet outil n'est pas applicable, le théorème ci-après, généralisant le corollaire 20, permet parfois de montrer malgré tout qu'une série de fonctions converge uniformément. La preuve de ce théorème est basée sur le critère de Cauchy uniforme et sur la « transformation d'Abel ». Il s'agit d'une formule très simple permettant de transformer une somme finie. Cette formule est en un sens l'analogue pour les sommes finies de la formule d'intégration par parties.

Considérons un espace vectoriel E sur un corps K et une somme finie $S := \sum_{n=p}^{q} \lambda_n a_n$, où $(\lambda_n, a_n) \in K \times E$ pour tout n et $p, q \in \mathbb{Z}$, $p < q$. Posons $A_k := \sum_{n=p}^{k} a_n$ si $k \in [\![p,q]\!]$ et $A_k := 0$ lorsque $k \in \mathbb{Z}$ et $k < p$. Ainsi $a_n = A_n - A_{n-1}$ pour tout $n \in [\![p,q]\!]$, d'où :

$$S = \sum_{n=p}^{q} \lambda_n (A_n - A_{n-1}) = \sum_{n=p}^{q} \lambda_n A_n - \sum_{n=p}^{q} \lambda_n A_{n-1} = \sum_{n=p}^{q} \lambda_n A_n - \sum_{n=p-1}^{q-1} \lambda_{n+1} A_n.$$

Comme $A_{p-1} := 0$, on obtient la formule d'Abel suivante :

$$S \;=\; \sum_{n=p}^{q} \lambda_n a_n \;=\; \lambda_q A_q + \sum_{n=p}^{q-1} (\lambda_n - \lambda_{n+1}) A_n. \tag{11}$$

Théorème 22. Soient X un ensemble non vide et E un e.v. n. complet. Considérons deux suites d'applications $(f_n : X \to \mathbb{R}_+)$ et $(g_n : X \to E)$ telles que :

1. Pour tout $x \in X$, la suite $\big(f_n(x)\big)$ est décroissante.
2. La suite de fonctions $(f_n : X \to \mathbb{R}_+)$ converge uniformément vers 0.
3. Il existe un réel M tel que $\left\| \sum_{k=0}^{n} g_k(x) \right\| \leqslant M$ pour tout $n \in \mathbb{N}$ et tout $x \in X$.

Sous ces hypothèses, la série de fonctions $\sum_{n \geqslant 0} f_n g_n$ converge uniformément.

Démonstration. Montrons que la série de fonctions donnée satisfait le critère de Cauchy uniforme. Soit $\varepsilon > 0$. D'après l'hypothèse 2, il existe $N \in \mathbb{N}^*$ tel que $f_n(x) \leqslant \varepsilon$ pour tout $n \geqslant N$ et tout $x \in X$. Soient $x \in X$ et p, q tels que $N \leqslant p < q$. Appliquons la formule d'Abel, en posant $\lambda_n := f_n(x)$ et $a_n := g_n(x)$ pour tout $n \in [\![p, q]\!]$. Pour tout $n \in [\![p, q]\!]$, l'hypothèse 3 donne :

$$\|A_n\| \;=\; \left\| \sum_{k=p}^{n} g_k(x) \right\| \;=\; \left\| \sum_{k=0}^{n} g_k(x) - \sum_{k=0}^{p-1} g_k(x) \right\| \;\leqslant\; 2M \quad \text{(inégalité triangulaire).}$$

Toujours grâce à l'inégalité triangulaire, la formule d'Abel entraîne donc :

$$\left\| \sum_{n=p}^{q} f_n(x) g_n(x) \right\| \;\leqslant\; 2M \left[f_q(x) + \sum_{n=p}^{q-1} \big| f_n(x) - f_{n+1}(x) \big| \right].$$

L'hypothèse 1 intervient ici : elle permet de supprimer les valeurs absolues dans le second membre. Le crochet vaut alors simplement $f_p(x)$ (les termes s'éliminent). D'où $\big\| \sum_{n=p}^{q} f_n(x) g_n(x) \big\| \leqslant 2M\varepsilon$, ceci pour tous p, q tels que $N \leqslant p < q$ et tout $x \in X$. Le critère de Cauchy uniforme étant ainsi vérifié, la série de fonctions donnée converge uniformément, car E est complet. ∎

Exercice 4.

Montrer que la formule $f(t) := \sum_{n=1}^{+\infty} e^{int}/n$ définit une fonction continue f sur $]0, 2\pi[$.

Solution. Appliquons le théorème avec $E := \mathbb{C}$, $f_n(t) := 1/n$, $g_n(t) = e^{int}$ pour tous $n \in \mathbb{N}^*$ et $t \in]0, 2\pi[$. Les hypothèses 1 et 2 sont évidemment satisfaites. Si $n \in \mathbb{N}^*$ et $t \in]0, 2\pi[$, on a $e^{it} \neq 1$, d'où :

$$\sum_{k=1}^{n} e^{ikt} = e^{it} \frac{e^{int} - 1}{e^{it} - 1} = e^{it} \frac{e^{int/2} \big[2i \sin(nt/2) \big]}{e^{it/2} \big[2i \sin(t/2) \big]} = e^{it(n+1)/2} \frac{\sin(nt/2)}{\sin(t/2)}.$$

On en déduit la majoration $\left| \sum_{k=1}^{n} e^{ikt} \right| \leqslant \dfrac{1}{\sin(t/2)}$.

Pour avoir une majoration uniforme en t, fixons $a \in]0,\pi[$. Lorsque t décrit $[a, 2\pi - a]$, $t/2$ décrit $[a/2, \pi - a/2]$, donc $\sin(t/2) \geqslant \sin(a/2)$ et par suite $\left|\sum_{k=1}^{n} e^{ikt}\right| \leqslant \dfrac{1}{\sin(a/2)}$.
Le théorème montre alors que la série de fonctions donnée converge uniformément sur $[a, 2\pi - a]$, d'où la continuité de f sur cet intervalle. Puisque c'est vrai pour tout $a \in]0,\pi[$, f est continue sur $]0, 2\pi[$.

Traduisons pour des séries de fonctions les théorèmes d'intégration et de dérivation terme à terme d'une suite de fonctions.

Théorème 23. Soit $(f_n : [a,b] \to \mathbb{R})$ une suite de fonctions intégrables. Si la série $\sum_{n\geqslant 0} f_n$ converge uniformément, sa somme f est intégrable sur $[a,b]$, et l'on a :

$$\int_a^b f := \int_a^b \left(\sum_{n=0}^{+\infty} f_n\right) = \sum_{n=0}^{+\infty} \left(\int_a^b f_n\right). \tag{12}$$

Démonstration. Il suffit d'appliquer à la suite (S_n) des sommes partielles de la série de fonctions donnée le théorème 13 de la page 483, en notant que chaque S_n est intégrable. ∎

Théorème 24. Soient I un intervalle de $\mathbb{R}$, E un e.v. n. réel de dimension finie et $(f_n : I \to E)$ une suite de fonctions de classe $\mathcal{C}^1$. On fait les hypothèses suivantes :
a) Il existe un point $a \in I$ tel que la série $\sum_{n\geqslant 0} f_n(a)$ converge.
b) La série $\sum_{n\geqslant 0} f_n'$ converge, uniformément sur I, vers une fonction $g : I \to E$.
La série de fonctions $\sum_{n\geqslant 0} f_n$ converge alors, uniformément sur tout compact de I, vers une fonction $f : I \to E$. De plus, f est de classe $\mathcal{C}^1$ sur I et $f' = g$.

Démonstration. Il suffit d'appliquer à la suite des sommes partielles de la série de fonctions donnée le théorème 14 de la page 483 et la remarque de la page 484. ∎

Dans l'hypothèse b), la convergence uniforme sur tout compact de I suffirait évidemment.

2.2 Convergence normale

Pour une série d'applications, il y a un autre mode de convergence, à la fois plus fort que la convergence uniforme et souvent plus facile à établir, c'est la « convergence normale ». Soient toujours X un ensemble non vide et E un e.v. n..

Définition 3. On dit qu'une série $\sum_{n\geqslant 0} f_n$ d'applications de X dans E converge *normalement* si les f_n sont bornées et si la série numérique $\sum_{n\geqslant 0} \|f_n\|_\infty$ converge.

Traduisons : les f_n sont bornées et, dans l'e.v. n. $\left(\mathcal{B}(X,E),\|\cdot\|_\infty\right)$, la série donnée converge absolument au sens de la définition 24 de la page 431.

Théorème 25. Supposons que l'e.v. n. E soit complet. Pour une série d'applications de X dans E, la convergence normale entraîne la convergence uniforme.

Démonstration. Soit $\sum_{n\geqslant 0} f_n$ une série normalement convergente d'applications de X dans E. C'est donc une série absolument convergente de l'e.v. n. $\left(\mathcal{B}(X,E),\|\cdot\|_\infty\right)$. Puisque E est complet, $\left(\mathcal{B}(X,E),\|\cdot\|_\infty\right)$ l'est aussi (corollaire 4 de la page 479). D'après le théorème 41 de la page 431, la série $\sum_{n\geqslant 0} f_n$ de $\left(\mathcal{B}(X,E),\|\cdot\|_\infty\right)$ converge, soit f sa somme. Vu le corollaire 2 de la page 477, (f_n) converge uniformément vers f. ∎

Pour établir la convergence normale de $\sum_{n\geqslant 0} f_n$, il n'est pas indispensable de calculer les normes $\|f_n\|_\infty$, il suffit de les *majorer* efficacement, à cause du résultat suivant, très utile :

Proposition 26. Pour que la série d'applications $\sum_{n\geqslant 0} f_n$ converge normalement il faut, et il suffit, qu'il existe une série numérique *convergente* $\sum_{n\geqslant 0} \alpha_n$ telle que, pour tout $n\in\mathbb{N}$ et tout $x\in X$, on ait : $\|f_n(x)\|\leqslant\alpha_n$ (la norme est ici prise dans E).

Démonstration. La nécessité est immédiate : s'il y a convergence normale, il suffit de poser $\alpha_n := \|f_n\|_\infty$ pour tout n. Supposons inversement que $\sum_{n\geqslant 0} \alpha_n$ vérifie la condition de l'énoncé. Si n est fixé, f_n est donc bornée et $\|f_n\|_\infty\leqslant\alpha_n$. Par comparaison, la convergence de $\sum_{n\geqslant 0}\|f_n\|_\infty$ résulte de celle de $\sum_{n\geqslant 0}\alpha_n$. ∎

Comme exemple (avec $X := \mathbb{R}$ et $E := \mathbb{C}$), posons $f_n(x) := e^{inx}/n^2$ pour tout $(n,x)\in\mathbb{N}^*\times\mathbb{R}$. Si $n\in\mathbb{N}^*$, $|f_n(x)| = 1/n^2$ pour tout x, d'où $\|f_n\|_\infty = 1/n^2$. La série $\sum_{n\geqslant 1} f_n$ converge donc normalement. Comme autre exemple, nous invitons le lecteur à se reporter à l'exercice 5 de la page 495, dans lequel est construite (comme somme d'une série de fonctions) une fonction continue de $\mathbb{R}$ dans $\mathbb{R}$ qui n'est dérivable en aucun point. Pour une série de fonctions $\sum_{n\geqslant 0} f_n$ de X dans $\mathbb{C}$, deux remarques s'imposent ici.

1. Supposons que cette série converge normalement. Soit $x\in X$. Pour tout $n\in\mathbb{N}$, on a $|f_n(x)|\leqslant\|f_n\|_\infty$, donc la série $\sum_{n\geqslant 0}|f_n(x)|$ converge, par comparaison. Ainsi *la convergence normale entraîne la convergence absolue simple*.
2. En combinant le théorème précédent et le théorème 6 de la page 480, on obtient ceci (X étant un espace métrique) : *si chaque f_n est continue et si la série de fonctions $\sum_{n\geqslant 0} f_n$ converge normalement sur tout compact de X, sa somme est continue sur X.*

Pour prouver que la somme d'une série de fonctions est de classe $\mathcal{C}^k$ ou $\mathcal{C}^\infty$, les résultats suivants sont très commodes et fréquemment employés (lorsque c'est possible !) :

Théorème 27. Soient $k \in \mathbb{N}^*$, I un intervalle de $\mathbb{R}$, x_0 un point de I, E un e.v. n. réel de dimension finie et $(f_n : I \to E)$ une suite de fonctions de classe $\mathcal{C}^k$, tels que :

a) Pour $p = 0, \dots, k-1$, la série $\sum_{n\geqslant 0} f_n^{(p)}(x_0)$ converge.

b) La série $\sum_{n\geqslant 0} f_n^{(k)}$ converge uniformément (par exemple normalement) sur tout compact de I.

Alors :

1. Pour $p = 0, \dots, k-1$, la série $\sum_{n\geqslant 0} f_n^{(p)}$ converge uniformément sur tout compact de I.
2. La fonction $f := \sum_{n=0}^{+\infty} f_n$ est de classe $\mathcal{C}^k$ sur I, et ses dérivées s'obtiennent en dérivant terme à terme :

$$f^{(p)}(x) = \sum_{n=0}^{+\infty} f_n^{(p)}(x) \quad \text{pour tout} \quad (p,x) \in [\![0,k]\!] \times I. \tag{13}$$

Démonstration. Comme pour le théorème 14 de la page 483, on peut supposer que $E := \mathbb{R}$. Par ailleurs, il suffit évidemment de traiter le cas où I est un segment $I := [a,b]$, $a < b$ (tout compact de $\mathbb{R}$ est contenu dans un segment). L'hypothèse b) signifie donc que la série $\sum_{n\geqslant 0} f_n^{(k)}$ converge uniformément sur I. Montrons alors d'une part que chaque série $\sum_{n\geqslant 0} f_n^{(p)}$ converge uniformément sur I ($p = 0, \dots, k-1$), et d'autre part que l'assertion 2 est vraie.

Raisonnons par récurrence sur k. Le cas $k := 1$ résulte du théorème 14 de la page 483. Supposons $k \geqslant 2$, le résultat étant vrai à l'ordre $k-1$. Soient $(f_n)_n$, $I = [a,b]$ et $x_0 \in I$ vérifiant la condition a) et tels que la série $\sum_{n\geqslant 0} f_n^{(k)}$ converge uniformément sur I. Appliquons l'hypothèse de récurrence à la série $\sum_{n\geqslant 0} f_n'$, formée de fonctions de classe $\mathcal{C}^{k-1}$. D'abord, pour $p = 0, \dots, k-2$, la série $\sum_{n\geqslant 0} f_n^{(p+1)}$ converge uniformément sur I, en particulier $\sum_{n\geqslant 0} f_n'$ converge uniformément sur I, notons g sa somme. Ensuite, g est de classe $\mathcal{C}^{k-1}$ et, pour tout $(p,x) \in [\![0,k-1]\!] \times I$, on a $g^{(p)}(x) = \sum_{n=0}^{+\infty} f_n^{(p+1)}(x)$.

Maintenant, la série $\sum_{n\geqslant 0} f_n'$ converge uniformément sur I et la série $\sum_{n\geqslant 0} f_n(x_0)$ converge.

D'après le théorème 14 de la page 483, la série $\sum_{n\geqslant 0} f_n$ converge uniformément sur I, sa somme f est de classe $\mathcal{C}^1$, et $f' = g$. Puisque g est de classe $\mathcal{C}^{k-1}$, f est de classe $\mathcal{C}^k$, et les assertions 1 et 2 sont clairement satisfaites pour la série $\sum_{n\geqslant 0} f_n$. ∎

Théorème 28. Soient I un intervalle de $\mathbb{R}$, E un e.v. n. réel de dimension finie et $(f_n : I \to E)$ une suite de fonctions de classe $\mathcal{C}^\infty$ telle que, pour tout $p \geqslant 0$, la série des dérivées $\sum_{n\geqslant 0} f_n^{(p)}$ converge uniformément (par exemple normalement) sur tout compact de I. Alors $f := \sum_{n=0}^{+\infty} f_n$ est de classe $\mathcal{C}^\infty$ et ses dérivées s'obtiennent en dérivant terme à terme : la formule (13) est valable pour tout $p \in \mathbb{N}$.

Démonstration. Il suffit manifestement d'appliquer le théorème précédent pour chaque $k \in \mathbb{N}^*$, puisque, par définition, une fonction est de classe $\mathcal{C}^\infty$ si, et seulement si, elle est de classe $\mathcal{C}^k$ pour tout $k \in \mathbb{N}$. ∎

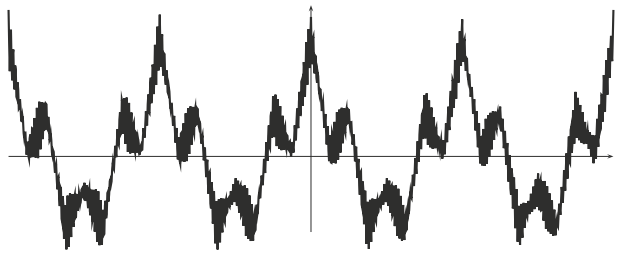
Il existe des fonctions « pathologiques »qui sont la somme d'une série de fonctions continues sur $\mathbb{R}$, normalement convergente sur $\mathbb{R}$, mais qui ne sont dérivables en aucun point de $\mathbb{R}$. L'exercice suivant en donne un exemple simple dû, en 1953, au mathématicien américain J. MacCarthy (un autre est proposé dans l'exercice 9 de la page 697).

Exercice 5 (La fonction de MacCarthy).

Soit $g : \mathbb{R} \to \mathbb{R}$ la fonction de période 4 telle que : $g(x) = \begin{cases} 1+x & \text{si } -2 \leqslant x \leqslant 0 \\ 1-x & \text{si } 0 \leqslant x \leqslant 2. \end{cases}$

1. Pour tout $k \in \mathbb{N}^*$, on définit la fonction $g_k : x \in \mathbb{R} \mapsto g\big(2^{2^k} x\big)$. Montrer que pour tout $x \in \mathbb{R}$, la série $\sum_{k\geqslant 1} \frac{1}{2^k} g_k(x)$ est convergente et que sa somme M est continue sur $\mathbb{R}$.
2. Avec les notations usuelles, pour tout $x \in \mathbb{R}$ et tout $n \in \mathbb{N}^*$, on a :
$$M(x) = S_{n-1}(x) + \frac{1}{2^n} g_n(x) + R_n(x).$$
   a) Soient x un réel donné et $n \in \mathbb{N}^*$. Montrer qu'il existe $\varepsilon = \pm 1$ tel que les points $\big(x, g_n(x)\big)$ et $\Big(x + \frac{\varepsilon}{2^{2^n}}, g_n\big(x + \frac{\varepsilon}{2^{2^n}}\big)\Big)$ soient sur un même segment du graphe de g_n.
   b) Soit $h_n = \frac{\varepsilon}{2^{2^n}}$, montrer que $R_n(x) = R_n(x + h_n)$.
   c) Montrer que $\big|g_n(x + h_n) - g_n(x)\big| = 1$.
   d) Montrer que $\sup_{1\leqslant k\leqslant n-1} \big|g_k(x + h_n) - g_k(x)\big| \leqslant 2^{-2^{n-1}}$. En déduire :
$$\big|S_{n-1}(x + h_n) - S_{n-1}(x)\big| \leqslant \frac{1}{2^{2^{n-1}}}.$$

3. Montrer que M n'est dérivable en aucun point.



Solution.

1. Pour tous $k \in \mathbb{N}^*, x \in \mathbb{R}$, $\left|\frac{1}{2^k} g_k(x)\right| \leqslant \frac{1}{2^k}$, la série $\sum_{k\in\mathbb{N}^*} g_k(t)$ est donc normalement convergente sur $\mathbb{R}$. Sa somme M est une fonction continue sur $\mathbb{R}$ puisque les g_k sont continues sur $\mathbb{R}$.
2. a) Soit x un réel donné. Soit $n \in \mathbb{N}^*$, il existe (au moins) un segment I de longueur $2l := \dfrac{2}{2^{2^n}}$ contenant x sur lequel g_n est affine. Soit m l'abscisse du milieu de I. Si $x \geqslant m$, $x - l \in I$ et si $x \leqslant m$, $x + l \in I$. D'où le résultat.
   b) Soit $k \in \mathbb{N}^*$, $k > n$, on a : $g_k(x + h_n) - g_k(x) = g\big(2^{2^k}(x + h_n)\big) - g(2^{2^k} x)$. L'entier $2^{2^k} h_n$ est multiple de 4, 4 est période de g, donc $g_k(x+h_n)-g_k(x) = 0$, par suite $R_n(x) = R_n(x + h_n)$.
   c) On a $g_n(x+h_n)-g_n(x) = g\big(2^{2^n}(x+h_n)\big)-g(2^{2^n} x)$. D'après la définition de h_n, les points $\Big(2^{2^n}(x+h_n), g\big(2^{2^n}(x+h_n)\big)\Big)$ et $\Big(2^{2^n} x, g\big(2^{2^n} x\big)\Big)$ sont sur un même segment du graphe de g dont la pente est ± 1. Puisque $2^{2^n}(x + h_n) - 2^{2^n} x = \varepsilon$, on a $\left|g_n(x + h_n) - g_n(x)\right| = \left|g\big(2^{2^n}(x + h_n)\big) - g(2^{2^n} x)\right| = 1$.
   d) Avec le même argument qu'au c) on peut affirmer :

$$\forall k \in [\![1, n-1]\!],\ \left|g\big(2^{2^k}(x + h_n)\big) - g(2^{2^k} x)\right| = \frac{2^{2^k}}{2^{2^n}} \leqslant \frac{2^{2^{n-1}}}{2^{2^n}} = \frac{1}{2^{2^{n-1}}}.$$

   Par suite : $\left|S_{n-1}(x + h_n) - S_{n-1}(x)\right| \leqslant \dfrac{1}{2^{2^{n-1}}} \displaystyle\sum_{k=1}^{n-1} \frac{1}{2^k} < \frac{1}{2^{2^{n-1}}}$.
3. Pour tout $x \in \mathbb{R}$ on a :

$$\frac{M(x + h_n) - M(x)}{h_n} = \frac{S_{n-1}(x + h_n) - S_{n-1}(x)}{h_n} + \frac{g_n(x + h_n) - g_n(x)}{h_n};$$

   on sait que $\left|S_{n-1}(x + h_n) - S_{n-1}(x)\right| \leqslant \dfrac{1}{2^{2^{n-1}}}$ et $\left|g_n(x + h_n) - g_n(x)\right| = 1$, par suite, dès que $n \geqslant 3$ et donc $2^{n-1} \geqslant n + 1$:

$$\left| \frac{M(x + h_n) - M(x)}{h_n} \right| \geqslant 2^{2^n} \Big(\frac{1}{2^n} - \frac{1}{2^{2^{n-1}}} \Big) \geqslant 2^{2^n - n - 1}.$$

   Ainsi $\lim\limits_{n\to+\infty} h_n = 0$ et $\lim\limits_{n\to+\infty} \left| \frac{M(x+h_n)-M(x)}{h_n} \right| = +\infty$, M n'est donc pas dérivable en x.

Exemple : fonction zêta.
Comme exemple d'utilisation de la convergence normale, donnons la définition de la *fonction zêta de Riemann*, ainsi que quelques unes de ses propriétés les plus simples. Cette fonction a de nombreuses applications (elle joue un grand rôle, par exemple, dans l'étude de la répartition des nombres premiers). Il s'agit de la fonction définie par la formule suivante, s étant une variable complexe :

$$\zeta(s) := \sum_{n=1}^{+\infty} \frac{1}{n^s}. \tag{14}$$

Pour tout $n \in \mathbb{N}^*$, notons f_n la fonction (continue) $s \mapsto n^{-s} := \exp(-s \ln n)$ de $\mathbb{C}$ dans $\mathbb{C}$. Il s'agit donc d'étudier la série de fonctions $\sum_{n\geqslant 1} f_n$. Commençons par la convergence simple, afin de déterminer le domaine de définition $\mathcal{D}$ de ζ. Soit $s \in \mathbb{C}$, écrivons $s := u + iv$, où $u, v \in \mathbb{R}$. Alors $n^{-s} = \exp\bigl(-(u+iv)\ln n\bigr) = n^{-u}\exp(-iv \ln n)$. On en déduit que $|n^{-s}| = n^{-u}$, car $\exp(-iv \ln n)$ est de module 1. Si $u \leqslant 0$, n^{-u} ne tend pas vers 0 quand n tend vers $+\infty$; la série $\sum_{n\geqslant 1} f_n(s)$ diverge donc, puisque son terme général ne tend pas vers 0. Si $u > 1$, la série $\sum_{n\geqslant 1} n^{-u}$ converge (critère de Riemann), donc la série $\sum_{n\geqslant 1} f_n(s)$ converge absolument. Si $0 < u \leqslant 1$, la série $\sum_{n\geqslant 1} f_n(s)$ ne converge pas absolument (critère de Riemann), en fait cette série diverge (*cf.* les exercices II.2.20 et II.2.21). Le domaine de définition de ζ est donc le demi-plan ouvert $\mathcal{D} := \{s \in \mathbb{C} \mid \operatorname{Re} s > 1\}$.
Voyons s'il y a convergence uniforme. On peut montrer que $\sum_{n\geqslant 1} f_n$ ne converge pas uniformément sur $\mathcal{D}$. Par contre, un réel $a > 1$ étant fixé, considérons le demi-plan $H_a := \{s \in \mathbb{C} \mid \operatorname{Re} s \geqslant a\}$. Pour tout $s := u + iv \in H_a$, on a $u \geqslant a$, donc $|n^{-s}| = n^{-u} \leqslant n^{-a}$. Puisque $a > 1$, la série $\sum_{n\geqslant 1} n^{-a}$ converge, et donc la série $\sum_{n\geqslant 1} f_n$ converge *normalement sur* H_a. En particulier $\sum_{n\geqslant 1} f_n$ converge uniformément sur H_a, donc ζ est continue sur H_a. C'est vrai pour tout $a > 1$, d'où la continuité de ζ sur $\mathcal{D}$.
Restreignons-nous à l'intervalle $I :=]1, +\infty[$, et étudions la dérivabilité de ζ. Chaque f_n est de classe $\mathcal{C}^\infty$ sur I. Si $n, p \in \mathbb{N}^*$ et $s \in I$, on a de plus $f_n^{(p)}(s) = (-\ln n)^p n^{-s}$. Un entier $p \geqslant 1$ et un réel $a > 1$ étant fixés, soit $I_a := [a, +\infty[$. Pour tout entier $n \geqslant 1$, posons $u_n := (\ln n)^p n^{-a}$. Soit $\alpha \in]0, a-1[$. Lorsque n tend vers $+\infty$, $\ln n$ est négligeable devant $n^{\alpha/p}$, donc $(\ln n)^p$ est négligeable devant n^α, et par suite u_n est négligeable devant $1/n^{a-\alpha}$. Puisque $a - \alpha > 1$, le critère de Riemann montre que la série $\sum_{n\geqslant 1} u_n$ converge. Puisque $|n^{-s}| \leqslant n^{-a}$ pour tout $s \in I_a$, on a $\left|f_n^{(p)}(s)\right| \leqslant u_n$ pour tous $n \in \mathbb{N}^*$ et $s \in I_a$. Ainsi la série $\sum_{n\geqslant 1} f_n^{(p)}$ converge normalement, en particulier uniformément, sur I_a.
Le théorème 28 de la page 495 montre alors que ζ est de classe $\mathcal{C}^\infty$ sur I_a. C'est vrai pour tout $a > 1$, donc ζ est de classe $\mathcal{C}^\infty$ sur I. Voici la conclusion obtenue :
La fonction ζ est continue sur $\{s \in \mathbb{C} \mid \operatorname{Re} s > 1\}$, et elle est de classe $\mathcal{C}^\infty$ sur $]1, +\infty[$.

Exercice 6.

Montrer que la formule $f(s) := \sum_{n=1}^{+\infty}(-1)^n n^{-s}$ définit une fonction continue f sur $]0,+\infty[$. Quel lien y-a-t-il avec la fonction zêta ? Limite de $(s-1)\zeta(s)$ lorsque s tend vers 1^+ ?

Solution. Soit $a>0$. Le critère spécial uniforme des séries alternées montre que la série de fonctions $s\mapsto \sum_{n=1}^{+\infty}(-1)^n n^{-s}$ converge uniformément sur $[a,+\infty[$, donc f est continue sur $[a,+\infty[$. C'est vrai pour tout $a>0$, d'où la continuité de f sur $]0,+\infty[$.

Considérons un réel $s>1$ fixé, comparons $f(s)$ à $\zeta(s)$. Soit $n\in\mathbb{N}^*$. La somme $1+(-1)^n$ vaut 0 si n est impair et 2 si n est pair ($n:=2p$, où $p\in\mathbb{N}^*$). D'où :

$$\zeta(s)+f(s) \;=\; \sum_{n=1}^{+\infty}\frac{1+(-1)^n}{n^s} \;=\; 2\sum_{p=1}^{+\infty}\frac{1}{(2p)^s} \;=\; 2^{1-s}\sum_{p=1}^{+\infty}p^{-s} \;=\; 2^{1-s}\zeta(s).$$

Ainsi $f(s)=\zeta(s)\big[2^{1-s}-1\big]$. D'où $\zeta(s)=\big[2^{1-s}-1\big]^{-1}f(s)$ pour tout $s>1$. Cette formule permet si l'on veut de *définir* $\zeta(s)$ à partir de $f(s)$, ceci pour tout réel $s>0$, $s\neq 1$. On a ainsi prolongé la fonction zêta sur $]0,1[$.

Supposons que $s>1$ tende vers 1. D'abord $f(s)$ tend vers $f(1)$, car f est continue en 1. Ensuite $f(1)=-\sum_{n=1}^{+\infty}(-1)^{n-1}/n=-\ln 2$, *cf.* le corollaire 44 de la page 516. La dérivée de $s\mapsto 2^{1-s}=\exp\big((1-s)\ln 2\big)$ est $s\mapsto -(\ln 2)2^{1-s}$, elle vaut $-\ln 2$ en 1. D'où :

$$\lim_{s\to 1}\frac{2^{1-s}-1}{s-1}=-\ln 2,\quad \text{puis}\quad \lim_{s\to 1}(s-1)\big[2^{1-s}-1\big]^{-1}=-(\ln 2)^{-1}.$$

Il en résulte que $(s-1)\zeta(s)$ tend vers 1 lorsque s tend vers 1^+.

L'exercice suivant porte sur la fonction Γ, il propose une preuve du fait que Γ soit de classe $\mathcal{C}^\infty$ sur $]0,+\infty[$.

Exercice 7.

1. À l'aide d'une série de fonctions appropriée, montrer que la formule :

$$\Gamma(x) \;:=\; \lim_{n\to+\infty}\frac{n!n^x}{x(x+1)\cdots(x+n)} \tag{15}$$

définit une fonction $\Gamma:]0,+\infty[\to\mathbb{R}$ de classe $\mathcal{C}^\infty$ (fonction Gamma d'Euler).

2. Soit C la *constante d'Euler* (*cf.* le module IV.4 de [L1]). Établir la formule :

$$(\ln\Gamma)'(x) \;=\; \frac{\Gamma'(x)}{\Gamma(x)} \;=\; -C-\frac{1}{x}+\sum_{k=1}^{+\infty}\left(\frac{1}{k}-\frac{1}{x+k}\right)\quad\text{pour tout}\quad x>0. \tag{16}$$

En déduire que la fonction $\ln\Gamma$ est strictement convexe sur I.

Solution.

1. Posons $I :=]0, +\infty[$. Notons $(g_n)_{n\geqslant 1}$ la suite de fonctions apparaissant dans la formule. Pour tout n, posons $\varphi_n := \ln g_n$. Pour la première assertion de l'énoncé, il suffit de montrer que la suite (φ_n) converge simplement vers une fonction φ de classe $\mathcal{C}^\infty$, car (g_n) convergera alors simplement vers $\exp\circ\varphi$, qui sera de classe $\mathcal{C}^\infty$ comme composée de fonctions de classe $\mathcal{C}^\infty$. Si $n \in \mathbb{N}^*$ et $x \in I$, on a :

$$\varphi_n(x) \;:=\; x\ln n + \ln(n!) - \sum_{k=0}^{n} \ln(x+k) \;=\; x\ln n + \sum_{k=1}^{n} \ln k - \sum_{k=0}^{n} \ln(x+k).$$

Cette expression suggère de poser pour tout n : $f_n := \varphi_{n+1} - \varphi_n$. D'abord parce que la suite (φ_n) converge simplement vers une fonction φ si, et seulement si, la *série* de fonctions $\sum\limits_{n\geqslant 1} f_n$ converge simplement et a pour somme $\varphi - \varphi_1$ (les sommes partielles de cette série sont les $\varphi_{n+1} - \varphi_1$). Une autre raison est que f_n a une expression plus simple que φ_n :

$$f_n(x) := x\big[\ln(n+1) - \ln n\big] + \ln(n+1) - \ln(x+n+1) = x\ln\left(1+\frac{1}{n}\right) - \ln\left(1+\frac{x}{n+1}\right).$$

Chaque f_n est en fait de classe $\mathcal{C}^\infty$ sur $J :=]-1, +\infty[$. Pour tout $x \in J$, on a :

$$f_n'(x) = \ln\left(1+\frac{1}{n}\right) - \frac{1}{x+n+1} \quad\text{et}\quad f_n^{(p)}(x) = \frac{(-1)^p(p-1)!}{(x+n+1)^p} \quad\text{dès que}\quad p \geqslant 2.$$

Appliquons le théorème 27 de la page 494 à la série $\sum\limits_{n\geqslant 1} f_n$, en prenant $x_0 := 0$ et $k \geqslant 2$ arbitraire. La série $\sum\limits_{n\geqslant 1} f_n(0)$ converge trivialement, car tous ses termes sont nuls. Lorsque n tend vers $+\infty$, on a :

$$f_n'(0) = \ln\left(1+\frac{1}{n}\right) - \frac{1}{n+1} = \frac{1}{n} - \frac{1}{n+1} + \mathrm{O}\left(\frac{1}{n^2}\right) = \mathrm{O}\left(\frac{1}{n^2}\right).$$

Il en résulte que la série $\sum\limits_{n\geqslant 1} f_n'(0)$ converge absolument. Pour tout $p \geqslant 2$, la série $\sum\limits_{n\geqslant 1} f_n^{(p)}$ converge normalement sur J. En effet, pour tout $x \in J$, on a :

$$\left|f_n^{(p)}(x)\right| \;=\; \frac{(p-1)!}{(x+n+1)^p} \;\leqslant\; \frac{(p-1)!}{n^p},$$

et la série $\sum\limits_{n\geqslant 1} 1/n^p$ converge.

Le théorème 27 de la page 494 montre ainsi que la série $\sum\limits_{n\geqslant 1} f_n$ converge uniformément sur tout compact de J et que sa somme f est une fonction de classe $\mathcal{C}^\infty$ sur J. Revenons alors à la suite $(\varphi_n)_n$. Nous savons maintenant qu'elle converge simplement vers la fonction $\varphi := f + \varphi_1$. Comme f et $\varphi_1 : x \mapsto -\big[\ln x + \ln(x+1)\big]$ sont de classe $\mathcal{C}^\infty$ sur I, φ est de classe $\mathcal{C}^\infty$ sur I. Enfin, sur l'intervalle I, la suite $(\varphi_n)_n$ converge simplement vers φ, donc la suite $(g_n)_n$ converge simplement vers $\exp\varphi$. Cela justifie la définition de Γ, et montre de plus que $\Gamma := \exp\varphi$ est de classe $\mathcal{C}^\infty$ sur I, comme composée de fonctions de classe $\mathcal{C}^\infty$.

2. Le théorème 27 de la page 494 montre aussi que la série $\sum_{n\geqslant 1} f_n'$ converge uniformément sur tout compact de J et que sa somme est $f' = \varphi' - \varphi_1'$. Calculons les sommes partielles de cette série. Pour tout entier $n \geqslant 2$ et tout $x \in J$, on a :

$$\begin{aligned}\sum_{k=1}^{n-1} f_k'(x) &= \sum_{k=1}^{n-1}\left\{\ln\left(1+\frac{1}{k}\right) - \frac{1}{x+k+1}\right\} = \sum_{k=1}^{n-1}\left\{\ln(k+1) - \ln k - \frac{1}{x+k+1}\right\}\\ &= \ln n - \sum_{k=1}^{n-1}\frac{1}{x+k+1} = \ln n + \frac{1}{x+1} - \sum_{k=1}^{n}\frac{1}{x+k}\\ &= \frac{1}{x+1} + \sum_{k=1}^{n}\left(\frac{1}{k} - \frac{1}{x+k}\right) - \underbrace{\left(1+\frac{1}{2}+\cdots+\frac{1}{n}-\ln n\right)}_{u_n}.\end{aligned}$$

Par définition, la constante d'Euler C est la limite de u_n lorsque n tend vers $+\infty$. En faisant tendre n vers $+\infty$ dans l'égalité précédente, il vient, pour tout $x \in I$:

$$\varphi'(x) = \varphi_1'(x) + f'(x) = \varphi_1'(x) + \frac{1}{x+1} - C + \sum_{k=1}^{+\infty}\left(\frac{1}{k} - \frac{1}{x+k}\right),$$

et la formule (16) en résulte, puisque $\varphi(x) = \ln\Gamma(x)$ et $\varphi_1'(x) = -1/x - 1/(x+1)$.
Le théorème 27 de la page 494 permet encore de dériver l'égalité (16) terme à terme, pour $x > 0$:

$$(\ln\Gamma)''(x) \;=\; \sum_{k=0}^{+\infty}\frac{1}{(x+k)^2}\cdot$$

La dérivée seconde de $\ln\Gamma$ est donc strictement positive sur I, ce qui montre que $\ln\Gamma$ est strictement convexe sur I.

---

3 Séries entières

On appelle *série entière* toute série de fonctions $\sum_{n\geqslant 0} f_n$ dont le terme général f_n est de la forme $z \mapsto a_n z^n$ (c'est une fonction définie et continue sur $\mathbb{C}$), où (a_n) est une suite complexe donnée. Cette série entière sera notée simplement $\sum_{n\geqslant 0} a_n z^n$. Si $n \in \mathbb{N}$, a_n est appelé *coefficient d'indice* n de la série entière. Les fonctions polynômes s'identifient à des séries entières particulières : celles dont tous les coefficients sont nuls, sauf un nombre fini. Dans cette section, nous utiliserons la notation suivante. Soit $R \in]0,+\infty[$. On notera $D(0,R)$ le disque ouvert $\{z \in \mathbb{C} \mid |z| < R\}$ et $\overline{D}(0,R)$ le disque fermé $\{z \in \mathbb{C} \mid |z| \leqslant R\}$. On pose aussi $D(0,0) := \varnothing$, $D(0,+\infty) := \mathbb{C}$, $\overline{D}(0,0) := \{0\}$ et $\overline{D}(0,+\infty) := \mathbb{C}$.

3.1 Domaine de convergence

Soit $\sum_{n\geqslant 0} a_n z^n$ une série entière. Comme pour toute série de fonctions, on s'intéresse en premier lieu à la convergence simple : il s'agit de déterminer le *domaine de convergence* de cette série, c'est-à-dire l'ensemble $\mathcal{D}$ des $z \in \mathbb{C}$ tels que la série numérique $\sum_{n\geqslant 0} a_n z^n$ converge. Si $z \in \mathcal{D}$, rappelons que la *somme* de la série $\sum_{n\geqslant 0} a_n z^n$ est notée $\sum_{n=0}^{+\infty} a_n z^n$. D'où une fonction $f : z \mapsto \sum_{n=0}^{+\infty} a_n z^n$ de $\mathcal{D}$ dans $\mathbb{C}$, appelée *somme* de la série entière $\sum_{n\geqslant 0} a_n z^n$. Montrons d'abord que $\mathcal{D}$ dépend essentiellement de la croissance de la suite $(|a_n|)$.

Définition 4. On appelle *rayon de convergence* de la série entière $F := \sum_{n\geqslant 0} a_n z^n$ la borne supérieure R_F de l'ensemble des réels $r \geqslant 0$ tels que la suite $(a_n r^n)$ soit bornée.

Un commentaire s'impose. Soit S l'ensemble des réels $r \geqslant 0$ tels que la suite $(a_n r^n)$ soit bornée. D'abord S n'est pas vide, car $0 \in S$. La borne supérieure $R_F := \sup S$ existe donc toujours, elle est soit finie soit égale à $+\infty$. Ensuite, il est clair que, si r, r' sont deux réels tels que $0 \leqslant r' \leqslant r$, $\big(r \in S\big) \Longrightarrow \big(r' \in S\big)$. D'où trois possibilités pour S :

- ou bien $R_F := 0$, auquel cas $S := \{0\}$,
- ou bien $R_F := +\infty$, auquel cas $S := [0, +\infty[$,
- ou bien $0 < R_F < +\infty$, auquel cas S est égal soit à $[0, R_F[$ soit à $[0, R_F]$.

La terminologie de « rayon de convergence » est justifié par le théorème suivant :

Théorème 29. Soit R le rayon de convergence de la série entière $\sum_{n\geqslant 0} a_n z^n$. Alors :

1. Pour tout $z \in \mathbb{C}$ tel que $|z| < R$, la série $\sum_{n\geqslant 0} a_n z^n$ converge absolument.
2. Pour tout $z \in \mathbb{C}$ tel que $|z| > R$, la série $\sum_{n\geqslant 0} a_n z^n$ diverge.

Démonstration. Soit $z \in \mathbb{C}$. Si $|z| < R$, choisissons un réel r tel que $|z| < r < R$. Alors $r \in S$, *i.e.* il existe un réel M tel que $|a_n|\, r^n \leqslant M$ pour tout n. Pour tout $n \in \mathbb{N}$, on a : $|a_n z^n| = \big(|a_n|\, r^n\big)(|z|/r)^n \leqslant M(|z|/r)^n$. Puisque $|z|/r < 1$, la série géométrique $\sum_{n\geqslant 0} (|z|/r)^n$ converge. Par comparaison, la série $\sum_{n\geqslant 0} |a_n z^n|$ converge, d'où l'assertion 1.

Supposons que $|z| > R$. Alors $|z| \notin S$, *i.e.* la suite $\big(|a_n z^n|\big)$ n'est pas bornée. En particulier la suite $(a_n z^n)$ ne converge pas vers 0, donc la série $\sum_{n\geqslant 0} a_n z^n$ diverge. ∎

Le domaine de convergence $\mathcal{D}$ de $F := \sum_{n\geqslant 0} a_n z^n$ est *presque* déterminé par le théorème précédent. Il y a deux cas évidents : $R_F := 0$, auquel cas $\mathcal{D} := \{0\}$, et $R_F := +\infty$, auquel cas $\mathcal{D} := \mathbb{C}$. Si $0 < R_F < +\infty$, le théorème donne : $D(0, R_F) \subset \mathcal{D} \subset \overline{D}(0, R_F)$. En

revanche, lorsque $|z| := R_F$, le théorème ne dit rien sur la convergence ou la divergence de $\sum_{n\geqslant 0} a_n z^n$ (nous reviendrons sur ce point ultérieurement, *cf.* le théorème radial d'Abel de la page 506). En tous cas, ce théorème fournit un premier moyen pour calculer R_F, c'est de déterminer $\mathcal{D}$ directement.

Exemples.

1. Soit $z \in \mathbb{C}$. La série géométrique $\sum_{n\geqslant 0} z^n$ converge si $|z| < 1$ et diverge si $|z| \geqslant 1$. La série entière $\sum_{n\geqslant 0} z^n$ a donc pour domaine de convergence $\mathcal{D} := D(0,1)$ et pour rayon de convergence $R := 1$. Nous connaissons aussi la somme de cette série : $\sum_{n=0}^{+\infty} z^n = 1/(1-z)$ dès que $|z| < 1$. Noter que $\sum_{n\geqslant 0} z^n$ diverge lorsque $|z| = 1$.

2. Considérons la série entière $\sum_{n\geqslant 1} z^n/n$. Pour $z := 1$, on obtient la série harmonique $\sum_{n\geqslant 1} 1/n$, qui diverge ; ainsi $1 \notin \mathcal{D}$. Pour $z := -1$, on obtient la série harmonique alternée $\sum_{n\geqslant 1} (-1)^n/n$, qui converge ; ainsi $-1 \in \mathcal{D}$. Ces deux renseignements suffisent pour voir que $R := 1$. En effet, $R \geqslant 1$ parce que $-1 \in \mathcal{D}$ (assertion 2 du théorème). De même $R \leqslant 1$ parce que $1 \notin \mathcal{D}$ (assertion 1 du théorème). En fait $\mathcal{D} := \overline{D}(0,1) \setminus \{1\}$ (*cf.* l'exercice 4 de la page 491).

3. Soient R le rayon de convergence et f la somme de $\sum_{n\geqslant 0} a_n z^n$. Posons $Z := az^p$, où $a \in \mathbb{C}^*$ et $p \in \mathbb{N}^*$ sont donnés. La série $\sum_{n\geqslant 0} a_n Z^n$ converge si $|Z| < R$, c'est-à-dire si $|a|\,|z|^p < R$, elle diverge si $|Z| > R$, c'est-à-dire si $|a|\,|z|^p > R$. Le rayon de convergence de la série entière $\sum_{n\geqslant 0} (a_n a^p) z^{np}$ est donc $+\infty$ si $R := +\infty$ et $\big(R/|a|\big)^{1/p}$ si $R < +\infty$. La somme de cette série entière est clairement $z \mapsto f(az^p)$.

4. Considérons deux séries entières $F := \sum_{n\geqslant 0} a_n z^n$ et $G := \sum_{n\geqslant 0} b_n z^n$, de rayons de convergence respectifs R_F et R_G. Supposons que $a_n \sim b_n$ lorsque n tend vers $+\infty$. Pour tout réel $r > 0$, les séries $\sum_{n\geqslant 0} |a_n|\, r^n$ et $\sum_{n\geqslant 0} |b_n|\, r^n$ sont alors de même nature, d'où aussitôt $R_F = R_G$. Supposons de même que, pour tout n, a_n, b_n soient réels et $0 \leqslant a_n \leqslant b_n$. Pour tout réel $r > 0$, la convergence de $\sum_{n\geqslant 0} b_n r^n$ entraîne celle de $\sum_{n\geqslant 0} a_n r^n$, d'où l'inégalité $R_F \geqslant R_G$ (le vérifier). Observer que c'est à chaque fois le théorème 29 de la page précédente qui permet de conclure.

Le résultat suivant permet, dans certains cas, de calculer le rayon de convergence R d'une série entière $\sum_{n\geqslant 0} a_n z^n$:

Proposition 30. On suppose que $a_n \neq 0$ dès que n est assez grand et que $|a_{n+1}/a_n|$ a une limite ℓ (éventuellement $+\infty$) quand n tend vers $+\infty$. Alors $R = 1/\ell$, avec la convention que $R = +\infty$ si $\ell := 0$ et $R = 0$ si $\ell := +\infty$.

Démonstration. Soit $z \in \mathbb{C}^*$ fixé, posons $u_n := |a_n z^n|$. Dès que n est assez grand, $u_n \neq 0$ et $u_{n+1}/u_n = |z|\,|a_{n+1}/a_n|$. Ainsi $\lim\limits_{n\to+\infty} u_{n+1}/u_n = |z|\,\ell$. Appliquons la règle de d'Alembert (*cf.* le module IV.4 de [L1]). Si $|z|\,\ell < 1$, la série $\sum\limits_{n\geqslant 0} u_n$ converge, *i.e.* la série $\sum\limits_{n\geqslant 0} a_n z^n$ converge absolument. Si $|z|\,\ell > 1$, u_n ne tend pas vers 0 quand n tend vers $+\infty$, donc $\sum\limits_{n\geqslant 0} a_n z^n$ diverge. Nous en savons assez sur $\mathcal{D}$ pour déterminer R.

Si $\ell := 0$, $\sum\limits_{n\geqslant 0} a_n z^n$ converge pour tout $z \in \mathbb{C}$, d'où $\mathcal{D} := \mathbb{C}$ et $R = +\infty$. Si $\ell := +\infty$, $\sum\limits_{n\geqslant 0} a_n z^n$ diverge pour tout $z \in \mathbb{C}^*$, d'où $\mathcal{D} := \{0\}$ et $R = 0$. Si $0 < \ell < +\infty$, $\sum\limits_{n\geqslant 0} a_n z^n$ converge pour $|z| < 1/\ell$ et diverge pour $|z| > 1/\ell$, d'où $R = 1/\ell$ en vertu du théorème. ∎

Exemples.

1. Prenons $a_n := 1/n^2$ pour $n \geqslant 1$. Alors $a_{n+1}/a_n = \big[n/(n+1)\big)^2$ tend vers 1 lorsque n tend vers $+\infty$. Le rayon de convergence de la série entière $\sum\limits_{n\geqslant 1} z^n/n^2$ est donc 1. Dans cet exemple, $\sum\limits_{n\geqslant 1} z^n/n^2$ converge absolument lorsque $|z| = 1$.
2. Prenons $a_n := n!$ pour $n \geqslant 0$. Alors $a_{n+1}/a_n = n+1$ tend vers $+\infty$ avec n, donc $R := 0$. Le domaine de convergence de la série entière $\sum\limits_{n\geqslant 0} n!z^n$ est ainsi $\{0\}$.
3. Prenons $a_n := 1/n!$ pour $n \geqslant 0$. Alors $a_{n+1}/a_n = 1/(n+1)$ tend vers 0 lorsque n tend vers $+\infty$, donc $R := +\infty$. Le domaine de convergence de la série entière $\sum\limits_{n\geqslant 0} z^n/n!$ est ainsi $\mathbb{C}$. Nous reviendrons plus loin sur cette série entière.

La proposition précédente ne doit pas être appliquée systématiquement. Considérons par exemple la série entière $\sum\limits_{n\geqslant 0} z^{n^2}$. C'est une série *lacunaire*, en ce sens que beaucoup de ses coefficients sont nuls. On a ici $a_n := 1$ si n est un carré et $a_n := 0$ sinon. La proposition ne s'applique donc pas. Il est en revanche facile de voir directement que $R = 1$. En effet la série $\sum\limits_{n\geqslant 0} z^{n^2}$ diverge lorsque $|z| \geqslant 1$, car son terme général ne tend pas vers 0. On en déduit l'inégalité $R \leqslant 1$. Si $|z| < 1$, on a $\left|z^{n^2}\right| \leqslant |z^n|$ pour tout n ; la convergence de $\sum\limits_{n\geqslant 0} |z^n|$ implique donc, par comparaison, celle de $\sum\limits_{n\geqslant 0} \left|z^{n^2}\right|$. D'où $R \geqslant 1$, et donc $R = 1$.

Autre exemple : prenons $a_n := 2 + (-1)^n$, c'est-à-dire $a_n := 3$ si n est pair et $a_n := 1$ si n est impair. Alors a_{n+1}/a_n vaut $1/3$ si n est pair et 3 si n est impair, donc a_{n+1}/a_n n'a

pas de limite quand n tend vers $+\infty$. La proposition ne s'applique pas non plus. Le lecteur pourra vérifier directement que $R = 1$ dans ce cas.

Il existe une formule donnant le rayon de convergence d'une série entière dans tous les cas (contrairement à la proposition 30 de la page précédente). Cette formule fait appel à la notion de « limite supérieure » d'une suite réelle $(u_n)_n$. Voici de quoi il s'agit. Pour tout $p \in \mathbb{N}$, posons $t_p := \sup_{k \geqslant p} u_k \in \mathbb{R} \cup \{+\infty\}$. Noter que $t_p \in \mathbb{R}$ si, et seulement si, la suite $(u_n)_n$ est majorée. Supposons que ce soit le cas. La suite réelle $(t_p)_{p \geqslant 0}$ est alors décroissante, elle possède donc une limite, finie ou égale à $-\infty$. Cette limite est appelée *limite supérieure* de la suite $(u_n)_n$, on la note $\limsup_{n \to +\infty} u_n$. On a donc par définition :

$$\limsup_{n \to +\infty} u_n := \lim_{p \to +\infty} \big(\sup_{n \geqslant p} u_n\big). \tag{17}$$

Si la suite (u_n) n'est pas majorée, on pose $\limsup_{n \to +\infty} u_n := +\infty$. Dans $\mathbb{R} \cup \{\pm\infty\}$, la suite (u_n) possède donc *toujours* une limite supérieure L, alors qu'elle n'a pas forcément de limite. Nous laissons le lecteur vérifier les faits suivants (l'assertion 3 est un bon exercice) :

1. On a $L = +\infty$ si, et seulement si, (u_n) n'est pas majorée.
2. On a $L = -\infty$ si, et seulement si, (u_n) a pour limite $-\infty$.
3. Lorsque L est un nombre réel, c'est la plus grande valeur d'adhérence de (u_n).
4. Si (u_n) a une limite finie ℓ, on a $L = \ell$.

La notion de limite supérieure conduit d'abord à un critère de convergence d'une série à termes réels positifs, complétant la règle de d'Alembert vue dans [L1].

Proposition 31 (Règle de Cauchy). Soient (u_n) une suite de réels positifs et L la limite supérieure de la suite $\big(\sqrt[n]{u_n}\big)$.

1. Si $L < 1$, la série $\sum_{n \geqslant 0} u_n$ converge.
2. Si $L > 1$ (éventuellement $L = +\infty$), la série $\sum_{n \geqslant 0} u_n$ diverge.

Lorsque $L = 1$, on ne peut rien dire *a priori*.

Démonstration. Pour tout $p \in \mathbb{N}$, posons $t_p := \sup_{n \geqslant p} \sqrt[n]{u_n}$, d'où $L := \lim_{p \to +\infty} t_p$.

Supposons $L < 1$; soit $a \in]L, 1[$. Pour tout entier p assez grand, on a $t_p \leqslant a$, en particulier $\sqrt[p]{u_p} \leqslant a$, c'est-à-dire $u_p \leqslant a^p$. La série $\sum_{n \geqslant 0} u_n$ converge, par comparaison avec la série géométrique $\sum_{n \geqslant 0} a^n$.

Supposons que $L > 1$, et soit $a \in]1, L[$. Puisque L est valeur d'adhérence de la suite $\big(\sqrt[n]{u_n}\big)$, il existe des entiers n arbitrairement grands tels que $\sqrt[n]{u_n} \geqslant a$, d'où $u_n \geqslant a^n > 1$. La série $\sum_{n \geqslant 0} u_n$ diverge, car son terme général ne tend pas vers 0. ∎

La règle de Cauchy est « plus puissante » que la règle de d'Alembert. Plus précisément, dans chaque cas où l'on peut appliquer la règle de d'Alembert, la règle de Cauchy s'applique aussi, mais il y a des cas où la règle de Cauchy s'applique et celle de d'Alembert ne s'applique pas (*cf.* l'exercice II.2.25). Comme pour la règle de d'Alembert, il faut bien se rappeler que la règle de Cauchy ne concerne que les séries à termes réels positifs.

Proposition 32 (formule d'Hadamard). Le rayon de convergence R d'une série entière $\sum_{n\geqslant 0} a_n z^n$ est donné par la formule suivante :

$$\frac{1}{R} = \limsup_{n\to+\infty} |a_n|^{1/n}, \tag{18}$$

avec la convention que $R = 0$ si la limite supérieure vaut $+\infty$ et $R = +\infty$ si cette limite supérieure vaut 0.

Démonstration. Notons L la limite supérieure en question. Pour tout réel $r > 0$, la limite supérieure de la suite de terme général $|a_n r^n|^{1/n}$ est évidemment rL si $L \in \mathbb{R}$ et $+\infty$ si $L = +\infty$. Si $L = +\infty$, la règle de Cauchy montre que la série $\sum_{n\geqslant 0} |a_n|\, r^n$ diverge pour tout $r > 0$, donc $R = 0$. Supposons que $L \in [0, +\infty[$. La règle de Cauchy montre que $\sum_{n\geqslant 0} |a_n|\, r^n$ converge lorsque $rL < 1$ et diverge lorsque $rL > 1$.

On en déduit que $R = +\infty$ si $L = 0$ et $R = 1/L$ si $L > 0$. ∎

Prenons par exemple $a_n := 2 + (-1)^n$. Pour tout n, a_n vaut 1 ou 3 suivant que n soit pair ou impair, donc $1 \leqslant a_n^{1/n} \leqslant 3^{1/n}$. Comme $3^{1/n}$ tend vers 1 lorsque n tend vers $+\infty$, on a $\lim_{n\to+\infty} a_n^{1/n} = 1$. D'après la formule d'Hadamard, le rayon de convergence de la série entière $\sum_{n\geqslant 0} a_n z^n$ vaut 1.

Jusqu'à présent nous n'avons étudié que la convergence simple d'une série entière. Voici un théorème très important concernant la convergence normale :

Théorème 33. Soit $\sum_{n\geqslant 0} a_n z^n$ une série entière, de rayon de convergence $R > 0$.

1. Pour tout réel r fixé tel que $0 < r < R$, la série $\sum_{n\geqslant 0} a_n z^n$ converge *normalement*, en particulier uniformément, sur le disque fermé $\overline{D}(0, r)$.
2. La fonction $f : z \mapsto \sum_{n\geqslant 0} a_n z^n$ est continue sur le disque ouvert $D(0, R)$ si $R \in \mathbb{R}$, sur $\mathbb{C}$ si $R := +\infty$.

Démonstration. Pour l'assertion 1, le théorème 29 de la page 501 montre que la série $\sum_{n\geqslant 0} |a_n|\, r^n$ converge. Pour tous $z \in \overline{D}(0, r)$ et $n \in \mathbb{N}$, on a $|a_n z^n| \leqslant |a_n|\, r^n$, d'où la convergence normale de $\sum_{n\geqslant 0} a_n z^n$ sur $\overline{D}(0, r)$.

Chacune des fonctions $z \mapsto a_n z^n$ est continue sur $\mathbb{C}$. Vu l'assertion 1 et le théorème 18 de la page 489, f est donc continue sur $\overline{D}(0, r)$, cela pour tout $r \in]0, R[$. Il en résulte que f est continue en tout point a tel que $|a| < R$, d'où l'assertion 2. ∎

Soit R le rayon de convergence de $\sum_{n\geqslant 0} a_n z^n$. Lorsque le domaine de convergence $\mathcal{D}$ contient un point z_0 tel que $|z_0| := R > 0$, on peut se demander si la somme $f : \mathcal{D} \to \mathbb{C}$ de la série entière est continue au point z_0. On peut montrer que ce n'est pas vrai en général, mais le théorème ci-après, appelé aussi *Lemme d'Abel*, donne un résultat partiel important, qui sera illustré plus loin (*cf.* le corollaire 44 de la page 516). La preuve de ce théorème est basée sur la transformation d'Abel.

Théorème 34 (théorème radial d'Abel).
Considérons une série entière $\sum_{n\geqslant 0} a_n z^n$ de rayon de convergence R, $0 < R < +\infty$.
Soit $z_0 \in \mathbb{C}$. Supposons que $|z_0| := R$ et que la série $\sum_{n\geqslant 0} a_n z_0^n$ converge.
Alors $f(z_0) := \sum_{n=0}^{+\infty} a_n z_0^n$ est la limite, lorsque $r \in]-1,1[$ tend vers 1 par valeurs inférieures, de $f(rz_0) := \sum_{n=0}^{+\infty} a_n (rz_0)^n$.

Démonstration. Supposons d'abord que $z_0 := 1$ et donc $R := 1$. Par hypothèse, la série $\sum_{n\geqslant 0} a_n$ est alors convergente. Il s'agit de montrer que la fonction $r \mapsto f(r) := \sum_{n\geqslant 0} a_n r^n$ de $]-1,1]$ dans $\mathbb{C}$ est continue à gauche en 1. Il suffit pour cela de vérifier que la série de fonctions $\sum_{n\geqslant 0} a_n r^n$ converge *uniformément* sur $[0,1]$, ou encore que cette série de fonctions satisfait le critère de Cauchy uniforme.
Soit $\varepsilon > 0$. Puisque la série $\sum_{n\geqslant 0} a_n$ converge, elle vérifie le critère de Cauchy. Il existe donc $N \in \mathbb{N}$ tel que, pour tous entiers p, q vérifiant $N \leqslant p < q$, on ait $\left|\sum_{n=p}^{q} a_n\right| \leqslant \varepsilon$.
Considérons un réel $r \in [0,1]$ et deux entiers p, q tels que $N \leqslant p < q$. Appliquons l'égalité 11 de la page 491 en prenant $\lambda_n := r^n$ pour tout n. Nous obtenons :

$$\sum_{n=p}^{q} a_n r^n = A_q r^q + \sum_{n=p}^{q-1} A_n (r^n - r^{n+1}), \quad \text{où} \quad A_n := \sum_{k=p}^{n} a_k \quad \text{pour} \quad n = p, \dots, q.$$

Pour tout n, $r^n - r^{n+1}$ est réel positif et $|A_n| \leqslant \varepsilon$. L'inégalité triangulaire donne donc :

$$\left|\sum_{n=p}^{q} a_n r^n\right| \leqslant \varepsilon\big[r^q + \sum_{n=p}^{q-1} (r^n - r^{n+1})\big] \;=\; \varepsilon r^p \;\leqslant \varepsilon.$$

La série de fonctions $\sum_{n\geqslant 0} a_n r^n$ satisfait donc le critère de Cauchy uniforme sur $[0,1]$, d'où la conclusion dans le cas $z_0 := 1$.
Si $z_0 \neq 1$, appliquons ce qui précède à la série entière $\sum_{n\geqslant 0} (a_n z_0^n) z^n$, dont le rayon de convergence est 1, d'après l'exemple 3 de la page 502. Puisque la série $\sum_{n\geqslant 0} a_n z_0^n$ converge, sa somme est la limite de $\sum_{n=0}^{+\infty} (a_n z_0^n) r^n$ lorsque $r \in]-1,1[$ tend vers 1 par valeurs inférieures, d'où la conclusion dans le cas général. ∎

Considérons la série entière $F := \sum_{n\geqslant 0}(-1)^n z^n$. Son rayon de convergence est 1, et sa somme $f : D(0,1) \to \mathbb{C}$ est donnée par $f(z) := 1/(1+z)$. Posons $z_0 := 1$. Lorsque $r \in [0,1[$ tend vers 1_-, $f(rz_0) = 1/(1+r)$ tend vers $1/2$. Cependant la série $\sum_{n\geqslant 0}(-1)^n z_0^n = \sum_{n\geqslant 0}(-1)^n$ diverge. La réciproque « naturelle » du théorème d'Abel est donc fausse. L'exercice ci-après donne une réciproque partielle du théorème lorsque les a_n vérifient certaines conditions supplémentaires.

Exercice 8.

Soit $\sum_{n\geqslant 0} a_n z^n$ une série entière de rayon de convergence 1 et $f :]-1,1[\to \mathbb{C}$ sa somme. On suppose que $f(x)$ a une limite finie ℓ quand x tend vers 1^-.

1. Si les a_n sont réels positifs, montrer que la série $\sum_{n\geqslant 0} a_n$ converge et a pour somme ℓ.

2. Montrer qu'on a la même conclusion si $a_n = o(1/n)$ lorsque n tend vers $+\infty$ (théorème de Tauber). Pour cela, prouver l'inégalité suivante, pour tout $x \in [0,1[$ et tout $N \in \mathbb{N}^*$:

$$\left| f(x) - \sum_{n=0}^{N} a_n \right| \leqslant (1-x)\sum_{n=0}^{N} |na_n| + \frac{1}{N(1-x)} \sup_{n>N} |na_n| . \qquad (*)$$

Conclure en prenant $x := 1 - 1/N$ et en faisant tendre N vers $+\infty$.

Solution.

1. Supposons que les a_n soient réels positifs. La fonction f est alors évidemment croissante sur $[0,1[$, donc $f(x) \leqslant \ell$ pour tout $x \in [0,1[$. Soit $N \in \mathbb{N}$. Pour tout $x \in [0,1[$, $\sum_{n=0}^{N} a_n x^n \leqslant \sum_{n=0}^{+\infty} a_n x^n = f(x) \leqslant \ell$. En faisant tendre x vers 1_-, on a donc $\sum_{n=0}^{N} a_n \leqslant \ell$. Il en résulte que la série $\sum_{n\geqslant 0} a_n$ converge, car elle est à termes réels positifs et ses sommes partielles sont majorées. Le théorème d'Abel montre alors que $\sum_{n=0}^{+\infty} a_n = \ell$.

2. Supposons que a_n soit négligeable devant $1/n$ lorsque n tend vers $+\infty$. Autrement dit, $\lim_{n\to+\infty} na_n = 0$, donc $\beta_N := \sup_{n>N} |na_n|$ tend vers 0 lorsque N tend vers $+\infty$. Soient $x \in [0,1[$ et $N \in \mathbb{N}^*$. On a :

$$\sum_{n=0}^{N} a_n - f(x) = \sum_{n=0}^{N} a_n - \sum_{n=0}^{+\infty} a_n x^n = \sum_{n=0}^{N} a_n(1-x^n) - \sum_{n=N+1}^{+\infty} a_n x^n .$$

Soit $n \in \mathbb{N}$. Si $n \leqslant N$, majorons $1 - x^n = (1-x)(1 + x + \cdots + x^{n-1})$ par $n(1-x)$. Si $n > N$, majorons $|a_n x^n| = |(na_n)(x^n/n)|$ par $\beta_N x^n/n$, ou encore par $\beta_N x^n/N$. Nous obtenons :

$$\left| \sum_{n=0}^{N} a_n - f(x) \right| \leqslant (1-x)\sum_{n=0}^{N} |na_n| + \frac{\beta_N}{N} \sum_{n=N+1}^{+\infty} x^n = (1-x)\sum_{n=0}^{N} |na_n| + \frac{\beta_N x^{N+1}}{N(1-x)} \cdot$$

L'inégalité $(*)$ en résulte. Dans cette inégalité, prenons $x := 1 - 1/N$. Il vient :

$$\left|\sum_{n=0}^{N} a_n - f(1-1/N)\right| \leqslant \beta_N + \frac{1}{N}\sum_{n=0}^{N} |na_n|\,.$$

Faisons tendre N vers $+\infty$. D'abord, β_N tend vers 0 et $f(1-1/N)$ tend vers ℓ. Pour conclure, *i.e.* pour montrer que la série $\sum\limits_{n\geqslant 0} a_n$ converge et a pour somme ℓ, il suffit donc de prouver que $(1/N)\sum\limits_{n=0}^{N} |na_n|$ tend vers 0. C'est une conséquence facile du fait que la suite $(na_n)_n$ converge vers 0 (*cf.* l'exercice II.4.6 de la page 701, portant sur les sommes de Cesàro).

Dans les sections suivantes, nous nous intéresserons aux fonctions qui sont somme d'une série entière. Une caractérisation de ces fonctions sera donnée dans le module II.6.

Définition 5. Soient D une partie ouverte de $\mathbb{C}$ (resp. de $\mathbb{R}$) et $f : D \to \mathbb{C}$ une fonction.

1. On dit que f est *développable en série entière* (en abrégé DSE) sur D s'il existe une série entière $F := \sum\limits_{n\geqslant 0} a_n z^n$ telle que $D \subset D(0, R_F)$ et $f(z) = \sum\limits_{n=0}^{+\infty} a_n z^n$ pour tout $z \in D$. On dit alors aussi que F *représente* f sur D.
2. On dit que f est *développable en série entière au voisinage de* 0 (en abrégé DSE_0) si elle est développable en série entière sur un voisinage ouvert de 0 dans $\mathbb{C}$ (resp. $\mathbb{R}$) inclus dans D.

Supposons que D soit un ouvert de $\mathbb{C}$. La fonction f est donc DSE_0 si, et seulement s'il existe un réel r vérifiant les conditions suivantes : $0 < r \leqslant R_F$, $D(0,r) \subset D$ et $f(z) = \sum\limits_{n=0}^{+\infty} a_n z^n$ pour tout $z \in D(0,r)$.

Supposons que D soit un ouvert de $\mathbb{R}$. La fonction f est DSE_0 si, et seulement s'il existe un réel r vérifiant les conditions suivantes : $0 < r \leqslant R_F$, $]-r, r[\subset D$ et $f(x) = \sum\limits_{n=0}^{+\infty} a_n x^n$ pour tout $x \in]-r, r[$.

Trois questions naturelles se posent, concernant les fonctions DSE_0 :

1. Comment savoir si une fonction donnée est DSE_0 ? Voir des réponses dans la section 3.4 si la variable est réelle et dans le module II.6 lorsque la variable est complexe.
2. Si une fonction est DSE_0, comment trouver une série entière la représentant ?
3. Comment, à l'inverse, calculer la somme d'une série entière donnée ?

Les questions 2 et 3 seront abordées dans les sections suivantes.

3.2 Opérations algébriques sur les séries entières

Une série entière $F := \sum_{n\geqslant 0} a_n z^n$ correspond simplement à une suite complexe (a_n). À F sont associés d'une part son rayon de convergence, noté ici R_F, et d'autre part la fonction (somme) $f : z \mapsto \sum_{n=0}^{+\infty} a_n z^n$, définie au moins pour $|z| < R_F$. Si l'on effectue une opération sur une ou plusieurs séries entières, on désire connaître le rayon de convergence de la série entière obtenue, ainsi que la fonction correspondant à cette série.

La première opération est évidente, c'est la loi externe. Si $F := \sum_{n\geqslant 0} a_n z^n$ et $\lambda \in \mathbb{C}$, par définition $\lambda F := \sum_{n\geqslant 0} \lambda a_n z^n$. Il est clair que, si $\lambda \neq 0$, $R_{\lambda F} = R_F$. Si f est la fonction associée à F, la fonction associée à λF est λf.

Passons à l'addition. Si $F := \sum_{n\geqslant 0} a_n z^n$ et $G := \sum_{n\geqslant 0} b_n z^n$, la somme $F + G$ est par définition la série entière $\sum_{n\geqslant 0} (a_n + b_n) z^n$. Voici ce que l'on peut dire de $F + G$:

Théorème 35. On a $R_{F+G} \geqslant \min(R_F, R_G)$, et il y a égalité lorsque $R_F \neq R_G$. Pour tout $z \in \mathbb{C}$ tel que $|z| < \min(R_F, R_G)$, on a $\sum_{n=0}^{+\infty} (a_n + b_n) z^n = \sum_{n=0}^{+\infty} a_n z^n + \sum_{n=0}^{+\infty} b_n z^n$.

Démonstration.

Soit $z \in \mathbb{C}$ tel que $|z| < \min(R_F, R_G)$. Chacune des séries numériques $\sum_{n\geqslant 0} a_n z^n$ et $\sum_{n\geqslant 0} b_n z^n$ converge, donc la série $\sum_{n\geqslant 0} (a_n + b_n) z^n$ converge, et sa somme vaut $\sum_{n=0}^{+\infty} a_n z^n + \sum_{n=0}^{+\infty} b_n z^n$. Puisque c'est vrai dès que $|z| < \min(R_F, R_G)$, le théorème 29 de la page 501 montre que $R_{F+G} \geqslant \min(R_F, R_G)$.

Si $R_F < R_G$, on a $R_{F+G} \geqslant R_F$. De $F = (F+G)+(-G)$, on tire $R_F \geqslant \min(R_{F+G}, R_G)$. D'où $R_{F+G} < R_G$ et $R_F \geqslant R_{F+G}$, donc $R_{F+G} = \min(R_F, R_G)$. ∎

Noter que chacune des séries entières $F := \sum_{n\geqslant 0} z^n$ et $-F$ a pour rayon de convergence 1, mais $F + (-F) = 0$ a pour rayon de convergence $+\infty > 1$.

Exercice 9.

Déterminer le rayon de convergence et la somme de la série entière $H := \sum_{n\geqslant 0} (2^n + 3^n) z^n$.

Solution. On a $H = F + G$, où $F := \sum_{n\geqslant 0} 2^n z^n$ et $G := \sum_{n\geqslant 0} 3^n z^n$. Soit $z \in \mathbb{C}$. La série géométrique $\sum_{n\geqslant 0} (2z)^n$ converge si, et seulement si, $|2z| < 1$, de sorte que $R_F = 1/2$. Si $|z| < 1/2$, on a de plus $\sum_{n=0}^{+\infty} (2z)^n = (1 - 2z)^{-1}$. Pour les mêmes raisons, $R_G = 1/3$ et $\sum_{n=0}^{+\infty} (3z)^n = (1 - 3z)^{-1}$ dès que $|z| < 1/3$. Le théorème donne

$R_H = \min(1/2, 1/3) = 1/3$ et $\sum_{n=0}^{+\infty}(2^n + 3^n)z^n = (1-2z)^{-1} + (1-3z)^{-1}$ dès que $|z| < 1/3$. Pour calculer R_H, on peut évidemment aussi appliquer ici la proposition 30 de la page 503.

Définition 6. Soient $F := \sum_{n\geqslant 0} a_n z^n$ et $G := \sum_{n\geqslant 0} b_n z^n$ deux séries entières.

On appelle *produit* de F et G la série entière $FG := \sum_{n\geqslant 0} c_n z^n$, en ayant posé, pour tout $n \in \mathbb{N}$:

$$c_n := \sum_{k=0}^{n} a_k b_{n-k} = \sum_{p+q=n} a_p b_q. \tag{19}$$

Voici ce que l'on peut dire de FG :

Théorème 36. On a $R_{FG} \geqslant \min(R_F, R_G)$. De plus, pour tout $z \in \mathbb{C}$ tel que $|z| < \min(R_F, R_G)$, on a $\sum_{n=0}^{+\infty} c_n z^n = \left(\sum_{n=0}^{+\infty} a_n z^n\right)\left(\sum_{n=0}^{+\infty} b_n z^n\right)$.

Démonstration. Soit $z \in \mathbb{C}$ tel que $|z| < \min(R_F, R_G)$. Chacune des deux séries complexes $\sum_{n\geqslant 0} a_n z^n$ et $\sum_{n\geqslant 0} b_n z^n$ converge absolument. Soit (γ_n) le *produit de Cauchy* des suites $(a_n z^n)$ et $(b_n z^n)$. Pour tout n, on a donc $\gamma_n := \sum_{k=0}^{n}(a_k z^k)(b_{n-k} z^{n-k}) = c_n z^n$. Comme nous l'avons vu dans le module IV.4 de [L1], la série $\sum_{n=0}^{+\infty} c_n z^n$ converge absolument, et sa somme vaut $\left(\sum_{n=0}^{+\infty} a_n z^n\right)\left(\sum_{n=0}^{+\infty} b_n z^n\right)$. Le théorème s'en déduit. ∎

Si $F := 1 - z$ et $G := \sum_{n\geqslant 0} z^n$, on a $R_F = +\infty$ et $R_G = 1$; cependant $FG = 1$ a pour rayon de convergence $+\infty \neq \min(R_F, R_G)$. Il est clair par ailleurs que la multiplication des séries entières est associative et commutative. Les séries entières forment en fait une $\mathbb{C}$-algèbre commutative (les trois lois étant l'addition, la multiplication et la loi externe).
Voici une application du théorème précédent :

Proposition 37. Soit $k \in \mathbb{N}$. Pour tout nombre complexe z tel que $|z| < 1$, on a :

$$\frac{1}{(1-z)^{k+1}} = \sum_{n=0}^{+\infty} \binom{n+k}{k} z^n. \tag{20}$$

Le rayon de convergence de la série entière figurant dans le second membre est égal à 1.

Démonstration. Considérons la série entière $F := \sum_{n\geqslant 0} z^n$. Son rayon de convergence

est 1, et sa somme est la fonction $f : z \mapsto (1-z)^{-1}$, de $D(0,1)$ dans $\mathbb{C}$. Pour tout $k \in \mathbb{N}$, notons H_k la série entière figurant dans le second membre de l'égalité (20). Évidemment $R_{H_k} \leqslant 1$ car, si $z := 1$, le terme général de la série ne tend pas vers 0.

Montrons, par récurrence sur $k \geqslant 0$, que $H_k = F^{k+1}$. C'est vrai pour $k := 0$. Supposons que $H_k = F^{k+1}$ pour un certain k. Pour tout $n \in \mathbb{N}$, le coefficient de z^n dans $F^{k+2} = F^{k+1}F = H_k F$ est alors égal, par définition du produit, à la somme $\displaystyle\sum_{h=0}^{n} \binom{h+k}{k}$.

L'égalité $F^{k+2} = H_{k+1}$ est alors une conséquence des égalités suivantes, valables pour tous $n, k \in \mathbb{N}$: $\displaystyle\sum_{h=0}^{n} \binom{h+k}{k} = \binom{n+k+1}{k+1}$. La preuve en est facile (fixer k et raisonner par récurrence sur n), *cf.* un exercice du module II.1 de [L1].

Soit $k \in \mathbb{N}$. Puisque $H_k = F^{k+1}$, une application itérée du théorème précédent montre que $R_{H_k} \geqslant R_F = 1$, d'où $R_{H_k} = 1$. De plus, dès que $|z| < 1$, le second membre de l'égalité (20) vaut $f(z)^{k+1} = (1-z)^{-(k+1)}$, d'où le résultat. ∎

Cette proposition permet de développer les fractions rationnelles en série entière :

Proposition 38. Soit $\Phi \in \mathbb{C}(X)$ une fraction rationnelle, écrite sous forme irréductible : $\Phi := P/Q$, où $P, Q \in \mathbb{C}[X]$ sont deux polynômes premiers entre eux et $\deg Q \geqslant 1$. Soit $\mathcal{R}$ l'ensemble des *pôles* de Φ (racines de Q). *On suppose que* $0 \notin \mathcal{R}$. La fonction $\varphi : \mathbb{C} \setminus \mathcal{R} \to \mathbb{C}$ définie par Φ est DSE_0. Plus précisément, il existe une série entière $F := \sum_{n \geqslant 0} a_n z^n$ vérifiant les conditions suivantes :

1. Le rayon de convergence de F est $R := \min\{|a| \mid a \in \mathcal{R}\}$.
2. Pour tout $z \in D(0,R)$, on a $\varphi(z) = P(z)/Q(z) = \displaystyle\sum_{n=0}^{+\infty} a_n z^n$.

Démonstration. Puisque Q est non constant, $\mathcal{R}$ n'est pas vide (théorème de d'Alembert-Gauß). L'idée est de décomposer Φ en *éléments simples*. On sait (revoir à ce sujet [L1]) que Φ est somme d'un polynôme (sa partie entière) et d'une combinaison linéaire d'éléments simples de la forme $1/(X-a)^k$, où $a \in \mathcal{R}$ et $k \in \mathbb{N}^*$. Plus précisément, on peut écrire :

$$\Phi(X) \;=\; E(X) + \sum_{a \in \mathcal{R}} \left(\sum_{k=1}^{m_a} \frac{t_{a,k}}{(X-a)^k} \right). \qquad (*)$$

Dans cette écriture, $E \in \mathbb{C}[X]$ est la partie entière de Φ. Pour tout $a \in \mathcal{R}$, m_a est la *multiplicité* de a comme pôle de Φ, *i.e.* comme racine de Q, et les $t_{a,k}$ sont des constantes complexes, le coefficient t_{a,m_a} étant en outre *non nul*. Soit $a \in \mathcal{R}$. Par hypothèse, $a \neq 0$, donc la proposition précédente et l'exemple 3 de la page 502 donnent pour tout $k \in \mathbb{N}^*$ le

développement suivant :

$$\frac{1}{(z-a)^k} = (-a)^{-k}(1-za^{-1})^{-k} = (-a)^{-k}\sum_{n=0}^{+\infty}\binom{n+k-1}{k-1}a^{-n}z^n.$$

Le rayon de convergence de la série entière obtenue est $|a|$. Regroupons les développements des différents éléments simples. On obtient une série entière $F := \sum_{n\geqslant 0} a_n z_n$ dont les coefficients a_n sont donnés par la formule suivante, en désignant par b_n les coefficients (presque tous nuls) du polynôme E :

$$a_n \;=\; b_n + \sum_{a\in\mathcal{R}}\sum_{k=1}^{m_a}(-1)^k a^{-(k+n)}t_{a,k}\binom{n+k-1}{k-1}.$$

D'après le théorème 35 de la page 509, le rayon de convergence R_F de F est au moins égal à R et, si $f : D(0,R_F) \to \mathbb{C}$ est la somme de la série entière F, on a $\varphi(z) = f(z)$ pour tout $z \in D(0,R)$.

Pour conclure, il suffit de montrer que $R_F = R$. Raisonnons par l'absurde, en supposant que $R_F > R$. Soit $a \in \mathcal{R}$ tel que $|a| = R$. Puisque $a \in D(0,R_F)$, f est continue au point a. Lorsque $r \in [0,1[$ tend vers 1_-, $f(ra) = \varphi(ra)$ tend donc vers $f(a)$. D'un autre côté, en remplaçant X par ra dans la formule $(*)$ et en se rappelant que $t_{a,m_a} \neq 0$, on voit que $\varphi(ra)$ est équivalent à $t_{a,m_a}\big[(r-1)a\big]^{-m_a}$, en particulier $|\varphi(ra)|$ tend vers $+\infty$. C'est absurde, et cette contradiction achève la démonstration. ∎

Exercice 10.

Soit $a > 0$ un réel fixé. Développer la fonction $f : z \mapsto \dfrac{1}{z^2 - 2z\cosh a + 1}$ en série entière au voisinage de 0, et préciser le rayon de convergence de la série entière obtenue.

Solution. On a $X^2 - 2X\cosh a + 1 = (X - e^a)(X - e^{-a})$. La fraction rationnelle $\Phi := 1/(X^2 - 2X\cosh a + 1)$ a donc deux pôles simples, e^a et e^{-a}. La partie entière de Φ est nulle ($\deg \Phi = -2$). Ainsi $\Phi := \dfrac{\lambda}{X - e^a} + \dfrac{\mu}{X - e^{-a}}\cdot$, où $(\lambda,\mu) \in \mathbb{R}^2$ est unique. Le calcul donne : $\lambda = 1/(2\sinh a) = -\mu$. Ici $R := e^{-a}$. Pour $|z| < e^{-a}$, on obtient :

$$\frac{1}{z-e^a} = \frac{-e^{-a}}{1-ze^{-a}} = -\sum_{n=0}^{+\infty} e^{-(n+1)a}\,z^n,\ \frac{1}{z-e^{-a}} = \frac{-e^a}{1-ze^a} = -\sum_{n=0}^{+\infty} e^{(n+1)a}\,z^n \quad \text{et}$$

$$f(z) = \frac{1}{2\sinh a}\sum_{n=0}^{+\infty}\big(e^{(n+1)a} - e^{-(n+1)a}\big)z^n = \frac{1}{\sinh a}\sum_{n=0}^{+\infty}\sinh\big((n+1)a\big)z^n.$$

Soit $F := \sum_{n\geqslant 0} a_n z^n$ la série entière obtenue. Nous savons que $R_F \geqslant e^{-a}$. Soit $r > e^{-a}$. Si n tend vers $+\infty$, $a_n \sim Ce^{na}$, où $C > 0$ est une constante (le vérifier). Ainsi, $a_n r^n$ tend vers $+\infty$, d'où $r \geqslant R_F$, ceci pour tout $r > e^{-a}$. D'où $e^{-a} \geqslant R_F$, et par suite $R_F = e^{-a}$.

Composition des séries entières

C'est une opération plus délicate. Commençons par définir l'*ordre* (ou valuation) d'une série entière entière non nulle $F := \sum_{n\geqslant 0} u_n z^n$: c'est le plus petit entier $n \geqslant 0$ tel que $u_n \neq 0$. Notons $\omega(F)$ cet ordre. Si $G := \sum_{n\geqslant 0} v_n z^n$ est une autre série entière non nulle, on a $\omega(FG) = \omega(F) + \omega(G)$. En effet, posons $p := \omega(F)$ et $q := \omega(G)$. Ainsi $u_p \neq 0$ mais $u_h = 0$ pour tout $h \in \mathbb{N}$, $h < p$. De même $v_q \neq 0$, mais $v_k = 0$ pour tout $k \in \mathbb{N}$, $k < q$. Par définition du produit FG, on a $FG := \sum_{n\geqslant 0} w_n z^n$ où, pour tout $n \in \mathbb{N}$, $w_n := \sum_{h=0}^{n} u_h v_{n-h}$. Alors $w_{p+q} = u_p v_q \neq 0$, mais $w_n = 0$ pour tout $n \in \mathbb{N}$ tel que $n < p + q$. En effet, soit $h \in [\![0, n]\!]$. Si $h < p$, on a $u_h = 0$. Si $h > p$, on a $v_{n-h} = 0$, car $n - h < n - p < q$. Ainsi $u_h v_{n-h} = 0$ pour tout $h \in [\![0, n]\!]$, d'où $w_n = 0$.

Soient $F := \sum_{n\geqslant 0} a_n z^n$, $G := \sum_{n\geqslant 0} b_n z^n$ deux séries entières, de rayons de convergence respectifs $R, R' > 0$. Soient f, g les fonctions correspondantes : $f(z) := \sum_{n=0}^{+\infty} a_n z^n$ dès que $|z| < R$ et $g(z) := \sum_{n=0}^{+\infty} b_n z^n$ dès que $|z| < R'$. *Faisons l'hypothèse suivante* : $\omega(G) \geqslant 1$, *i.e.* $b_0 = g(0) = 0$. Cherchons une série entière $H := \sum_{n\geqslant 0} c_n z^n$, de rayon de convergence strictement positif, et telle que $\sum_{n=0}^{+\infty} c_n z^n = f\big(g(z)\big)$ pour tout z de module assez petit. Pour tout $n \in \mathbb{N}$, considérons la série entière G^n (produit de n copies de G), de rayon de convergence au moins égal à R' vu le théorème 36 de la page 510. Écrivons $G^n := \sum_{p\geqslant 0} v_{n,p} z^p$. D'après l'alinéa précédent, $\omega(G^n) \geqslant n$, *i.e.* $v_{n,p} = 0$ dès que $p < n$.

Proposition 39. Sous ces hypothèses, posons $c_p := \sum_{n=0}^{p} a_n v_{n,p}$ pour tout $n \in \mathbb{N}$. La série entière $H := \sum_{n\geqslant 0} c_p z^p$, notée $H := F \circ G$, a un rayon de convergence strictement positif. Pour tout $z \in \mathbb{C}$ de module assez petit, on a de plus : $\sum_{p=0}^{+\infty} c_p z^p = f\big(g(z)\big)$.

Démonstration. Montrons que, si $r := |z| \in]0, R'[$ est assez petit , on a :

$$\sum_{n=0}^{+\infty} \left(\sum_{p=0}^{+\infty} a_n v_{n,p} z^p \right) = \sum_{p=0}^{+\infty} \left(\sum_{n=0}^{+\infty} a_n v_{n,p} z^p \right). \qquad (*)$$

Pour justifier cette interversion de sommations, invoquons le théorème 15 de la page 485. Soit $n \in \mathbb{N}$ un entier fixé. Puisque le rayon de convergence de G^n est au moins égal à R' et puisque $|z| < R'$, la série $\sum_{p\geqslant 0} v_{n,p} z^p$ converge absolument, et sa somme est $g(z)^n$, en vertu du théorème 36 de la

page 510. Ainsi la série $\sum_{p\geqslant 0} a_n v_{n,p} z^p$ converge absolument, et sa somme est $a_n g(z)^n$. Supposons un instant d'une part que la série $\sum_{n=0}^{+\infty} \left(\sum_{p=0}^{+\infty} |a_n v_{n,p} z^p| \right)$ converge, et d'autre part que $|g(z)| < R$. Nous pouvons alors appliquer le théorème 15 de la page 485, qui affirme que l'égalité $(*)$ est vraie. Puisque $|g(z)| < R$, le premier membre de l'égalité $(*)$ vaut $\sum_{n\geqslant 0} a_n g(z)^n = f\big(g(z)\big)$. Si $p \in \mathbb{N}$ est fixé, la série $\sum_{n=0}^{+\infty} a_n v_{n,p} z^p$ se réduit à la somme finie $\sum_{n=0}^{p} a_n v_{n,p} z^p = c_p z^p$. Ainsi $\sum_{p\geqslant 0} c_p z^p$ converge, donc le rayon de convergence de H est au moins égal à r, et sa somme est $f\big(g(z)\big)$.

Il reste à choisir r de sorte que les deux propriétés admises soient vérifiées. Introduisons la série entière $\Gamma := \sum_{n\geqslant 0} |b_n|\, z^n$. Son rayon de convergence est évidemment R' (seuls les modules des coefficients interviennent dans le calcul d'un rayon de convergence). Si $t \in \mathbb{C}$ et $|t| < R'$, posons $\gamma(t) := \sum_{n=0}^{+\infty} |b_n|\, t^n$. La fonction γ est continue. Comme $\gamma(0) = b_0 = 0$, il existe $\alpha > 0$ tel que $\big(|t| \leqslant \alpha\big) \Longrightarrow \big(|\gamma(t)| < R\big)$. Choisissons r de sorte que $0 < r < \min(\alpha, R')$. Alors $|g(z)| < R$:

$$|g(z)| \;=\; \left| \sum_{n=0}^{+\infty} b_n z^n \right| \;\leqslant\; \sum_{n=0}^{+\infty} |b_n|\, r^n \;=\; \gamma(r) \;<\; R.$$

Pour tout $n \in \mathbb{N}$, notons $w_{n,p}$ les coefficients de la série entière Γ^n. Ce sont des réels positifs, et l'on voit aisément, par récurrence sur n, que $|v_{n,p}| \leqslant w_{n,p}$ pour tous $n, p \in \mathbb{N}$. Pour tout $n \in \mathbb{N}$, on a donc : $\sum_{p=0}^{+\infty} |a_n v_{n,p} z^p| = |a_n| \sum_{p=0}^{+\infty} |v_{n,p}|\, r^p \leqslant |a_n| \sum_{p=0}^{+\infty} w_{n,p} r^p = |a_n|\, \gamma(r)^n$. Comme $\gamma(r) < R$, la série $\sum_{n\geqslant 0} |a_n|\, \gamma(r)^n$ converge. Il en résulte, par comparaison, que la série $\sum_{n=0}^{+\infty} \left(\sum_{p=0}^{+\infty} |a_n v_{n,p} z^p| \right)$ converge, ce qui achève la démonstration. ∎

Remarque. Si $R := +\infty$, cette preuve montre que le rayon de convergence de H est au moins égal à R' et que l'égalité $\sum_{p=0}^{+\infty} c_p z^p = f\big(g(z)\big)$ est vraie dès que $|z| < R'$.

3.3 Dérivation terme à terme d'une série entière

Nous allons voir que, si une série entière a un rayon de convergence $R > 0$, sa somme, restreinte à l'*intervalle de convergence ouvert* $]-R, R[$, est une fonction de classe $\mathcal{C}^\infty$.

Soit d'abord $F := \sum_{n\geqslant 0} a_n z^n$ une série entière. Posons :

$$F' \;:=\; \sum_{n\geqslant 1} n a_n z^{n-1} \;=\; \sum_{n\geqslant 0} (n+1) a_{n+1} z^n \quad \text{et} \quad \Phi \;:=\; \sum_{n\geqslant 0} a_n \frac{z^{n+1}}{n+1} \;=\; \sum_{n\geqslant 1} a_{n-1} \frac{z^n}{n}\cdot$$

Nous dirons que la série entière F' s'obtient *en dérivant F terme à terme* et que la série entière Φ s'obtient *en primitivant F terme à terme*. En dérivant Φ terme à terme, on obtient évidemment F. En primitivant F' terme à terme, on obtient $F - a_0$. En dérivant (terme à

terme) successivement, on définit des séries entières $F^{(p)}$ par récurrence sur p : $F^{(0)} := F$, $F^{(1)} := F'$, puis $F^{(p+1)} := \left(F^{(p)}\right)'$ pour tout $p \in \mathbb{N}$.

Proposition 40. On ne change pas le rayon de convergence d'une série entière en la dérivant ou en la primitivant terme à terme.

Démonstration. Il suffit de montrer que F et F' ont même rayon de convergence, car alors il en sera de même pour Φ et $\Phi' = F$. Considérons un réel $r > 0$. Si la suite $(na_n r^{n-1})_{n\geqslant 1}$ est bornée, les suites $(a_n r^{n-1})_{n\geqslant 1}$ et $(a_n r^n)_{n\geqslant 0}$ sont bornées *a fortiori*. Vu la définition du rayon de convergence, on a donc $R_F \geqslant R_{F'}$. Si $R_F := 0$, on a aussi $R_{F'} := 0$.

Supposons $R_F > 0$, et soient $r, \rho \in \mathbb{R}$ tels que $0 \leqslant \rho < r < R_F$. Il existe $M \in \mathbb{R}$ tel que $\forall n \in \mathbb{N},\ |a_n|\, r^n \leqslant M$. Si $n \geqslant 1$, $\left|na_n\rho^{n-1}\right| = n\,|a_n|\, r^{n-1}\left(\dfrac{\rho}{r}\right)^{n-1} \leqslant \dfrac{nM}{r}\left(\dfrac{\rho}{r}\right)^{n-1}$.

Comme $0 \leqslant \rho/r < 1$, on a $\lim\limits_{n\to+\infty} n\left(\dfrac{\rho}{r}\right)^{n-1} = 0$, donc la suite $(na_n\rho^{n-1})$ converge vers 0, d'où $R_{F'} \geqslant \rho$. C'est vrai pour tout $\rho \in [0, r[$, d'où $R_{F'} \geqslant r$. Cette nouvelle inégalité est vraie pour tout réel r tel que $0 < r < R_F$, et par suite $R_{F'} \geqslant R_F$, d'où $R_{F'} = R_F$.

Le théorème ci-dessous est essentiel, il relie les sommes des séries entières F et F' :

Théorème 41. Soit $F := \sum\limits_{n\geqslant 0} a_n z^n$ une série entière de rayon de convergence $R > 0$.

La fonction $f :]-R, R[\to \mathbb{C}$ définie par $f(x) := \sum\limits_{n=0}^{+\infty} a_n x^n$ est de classe $\mathcal{C}^1$, et :

$$f'(x) \;=\; \sum_{n=1}^{+\infty} na_n x^{n-1} \;=\; \sum_{n=0}^{+\infty}(n+1)a_{n+1}x^n \quad \text{pour tout} \quad x \in]-R, R[\,. \qquad (21)$$

Autrement dit, *si la variable x ne prend que des valeurs réelles*, f' est la somme de la série entière F'. Le cas d'une variable complexe sera étudié dans le module II.6.

Démonstration. La série de fonctions $\sum\limits_{n\geqslant 0} a_n x^n$ converge simplement vers f. Chacune des fonctions $x \mapsto a_n x^n$ étant de classe $\mathcal{C}^1$, il suffit, en vertu du théorème 24 de la page 492, de montrer que la série des dérivées, c'est-à-dire $\sum\limits_{n\geqslant 1} na_n x^{n-1}$, converge uniformément sur tout compact $J \subset]-R, R[$. Pour un tel compact J, on a $\inf J > -R$ et $\sup J < R$, donc il existe un réel $r > 0$ tel que $r < R$ et $J \subset [-r, r]$. D'après la proposition précédente et le théorème 33 de la page 505, la série $\sum\limits_{n\geqslant 0} a_n x^n$ converge normalement, donc uniformément, sur le disque $\overline{D}(0, r)$, *a fortiori*, elle converge uniformément sur le segment $[-r, r]$, donc aussi sur J.

Voici le théorème « inverse » du précédent :

Théorème 42. Soit $F := \sum_{n\geqslant 0} a_n z^n$ une série entière de rayon de convergence $R > 0$. Soit $f :]-R, R[\to \mathbb{C}$ sa somme : $f(x) := \sum_{n=0}^{+\infty} a_n x^n$. Pour tout $x \in]-R, R[$, on a :

$$\int_0^x f(t)\,\mathrm{d}t \;=\; \sum_{n=0}^{+\infty} a_n \frac{x^{n+1}}{n+1}\cdot \tag{22}$$

Démonstration. Soit donc Φ la série entière déduite de F par primitivation terme à terme. Le rayon de convergence de Φ est R. Soit $\varphi :]-R, R[\to \mathbb{C}$ la somme de la série entière Φ. Puisque $\Phi' = F$, le théorème précédent montre que φ est de classe $\mathcal{C}^1$ et que $\varphi' = f$. D'où la conclusion, puisque $\varphi(0) = 0$ (le terme constant de Φ est nul).

Donnons une première application importante de ce théorème.

Proposition 43. Le rayon de convergence de chacune des séries entières $\sum_{n\geqslant 1} (-1)^{n-1} z^n/n$ et $\sum_{n\geqslant 0} (-1)^n z^{2n+1}/(2n+1)$ est 1. Pour tout $x \in]-1, 1[$, on a :

$$\ln(1+x) \;=\; \sum_{n=1}^{+\infty} (-1)^{n-1} \frac{x^n}{n} \quad \text{et} \quad \arctan(x) \;=\; \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n+1}}{2n+1}\cdot \tag{23}$$

Démonstration. Soit $\mathbb{D} = D(0,1)$. La série entière $F := \sum_{n\geqslant 0} (-1)^n z^n$ a pour rayon de convergence 1, et sa somme est la fonction $z \mapsto 1/(1+z)$ de $\mathbb{D}$ dans $\mathbb{C}$. Par ailleurs, la série entière déduite de F par primitivation terme à terme est $\Phi := \sum_{n\geqslant 1} (-1)^{n-1} z^n/n$. D'après le théorème précédent, la fonction $x \mapsto \sum_{n=1}^{+\infty} (-1)^{n-1} x^n/n$ de $]-1, 1[$ dans $\mathbb{R}$ est la primitive de $x \mapsto 1/(1+x)$ nulle en 0, ce qui établit la première égalité de l'énoncé.

Pour la deuxième égalité de l'énoncé, on remplace ci-dessus z par z^2, et l'on conclut de même, en notant que la fonction $\arctan : \mathbb{R} \to \mathbb{R}$ est la primitive de $x \mapsto 1/(1+x^2)$ nulle en 0. Les détails sont laissés au lecteur. On prendra garde au fait suivant : $\arctan x$ est défini pour tout réel x, mais la deuxième série en (23) diverge lorsque $|x| > 1$.

Corollaire 44. On peut exprimer $\ln 2$ et $\pi/4$ à l'aide des séries suivantes :

$$\sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n} \;=\; \ln 2 \quad \text{et} \quad \sum_{n=}^{+\infty} \frac{(-1)^n}{2n+1} = \frac{\pi}{4}\cdot \tag{24}$$

Démonstration. Chacune de ces deux séries converge, en vertu du critère spécial des séries alternées. D'après le théorème radial d'Abel et la proposition précédente, la somme de la première série de l'énoncé est la limite, lorsque $x \in]-1, 1[$ tend vers 1^-, de $\ln(1+x)$. Cette limite est $\ln 2$, puisque la fonction $x \mapsto \ln(1+x)$ est continue en 1.

De même, la somme de la deuxième série est la limite, lorsque $x \in]-1, 1[$ tend vers 1^-, de $\arctan x$. Cette limite est $\arctan 1 = \pi/4$, car la fonction $\arctan$ est continue en 1.

Voici maintenant une conséquence *fondamentale* du théorème 41 de la page 515 :

Théorème 45. Soient $F := \sum_{n\geqslant 0} a_n z^n$ une série entière de rayon de convergence $R > 0$. La fonction $f :]-R, R[\to \mathbb{C}$ définie par $f(x) := \sum_{n=0}^{+\infty} a_n x^n$ est de classe $\mathcal{C}^\infty$. Pour tout $p \in \mathbb{N}^*$, la dérivée $p^{ème}$ de f est donnée par la formule suivante :

$$\forall x \in]-R, R[, \quad f^{(p)}(x) = \sum_{n=0}^{+\infty} (n+p)(n+p-1)\cdots(n+2)(n+1)a_{n+p}x^n. \quad (25)$$

En particulier, $a_p = \dfrac{f^{(p)}(0)}{p!}$ pour tout $p \in \mathbb{N}$.

Démonstration. Montrons d'abord que $F^{(p)} = \sum_{n\geqslant 0} (n+p)\cdots(n+2)(n+1)a_{n+p}z^n$ pour tout entier $p \geqslant 1$. Si $p := 1$, c'est la définition de F'. Supposons la formule vraie pour un certain $p \geqslant 1$. Notons b_n les coefficients de la série entière $F^{(p)}$. Soit $n \in \mathbb{N}$. Par définition de $F^{(p+1)} := \left(F^{(p)}\right)'$, le coefficient de z^n dans $F^{(p+1)}$ est $(n+1)b_{n+1}$, c'est-à-dire :

$$(n+1)\big[(n+p+1)\cdots(n+3)(n+2)a_{n+p+1}\big] = (n+p+1)\cdots(n+2)(n+1)a_{n+p+1},$$

ce qui établit la formule annoncée à l'ordre $p+1$.
Par récurrence sur $p \geqslant 1$, il résulte de la proposition 40 de la page 515 et du théorème 41 de la page 515 que le rayon de convergence de la série entière $F^{p)}$ est R, que f est de classe $\mathcal{C}^p$, et que, sur $]-R, R[$, $f^{(p)}$ est la somme de la série entière $F^{(p)}$. Ainsi f est de classe $\mathcal{C}^\infty$, et la formule (25) est vraie pour tout $p \geqslant 1$. En particulier, si $p \in \mathbb{N}^*$, $f^{(p)}(0) = p!a_p$, d'où la formule donnant a_p, formule encore valable si $p := 0$. ∎

Remarque. Considérons deux séries entières $F := \sum_{n\geqslant 0} a_n z^n$ et $G := \sum_{n\geqslant 0} b_n z^n$. Supposons qu'il existe un réel t, $0 < t \leqslant \min(R_F, R_G)$, tel que $\sum_{n=0}^{+\infty} a_n x^n = \sum_{n=0}^{+\infty} b_n x^n$ pour tout $x \in]-t, t[$. Alors $a_n = b_n$ pour tout n. Il suffit d'appliquer la dernière assertion du théorème à la série entière $F - G$ (*cf.* aussi le théorème 47 de la page 522).

Exemple. Définissons $f :]-1, +\infty[\to \mathbb{R}$ par $f(x) := \ln(1+x)/x$ si $x \neq 0$ et $f(0) := 1$. Il est clair que f est de classe $\mathcal{C}^\infty$ en-dehors de 0. La dérivée de $x \mapsto \ln(1+x)$ en 0 est 1, *i.e.* $\lim_{x\to 0} \ln(1+x)/x = 1$, ce qui prouve la continuité de f en 0. Par contre, il n'est pas tout à fait évident de montrer directement que f est de classe $\mathcal{C}^\infty$ au voisinage de 0.
Procédons indirectement, en partant de la série entière $G := \sum_{n\geqslant 1} (-1)^{n-1} z^n/n$, de rayon de convergence 1. On peut écrire $G = zF$, où $F := \sum_{n\geqslant 0} (-1)^n z^n/(n+1)$. Il est clair que le rayon de convergence de F est le même que celui de G. Sur l'intervalle $]-1, 1[$, notons g la somme de G et φ la somme de F.

Pour tout $x \in]-1,1[$, la proposition 43 de la page 516 montre que $g(x) = \ln(1+x)$, et évidemment $g(x) = x\varphi(x)$, d'où $\varphi(x) = \ln(1+x)/x := f(x)$ lorsque $x \neq 0$.
Comme $\varphi(0) = 1 = f(0)$, on a $\varphi = f$. Le théorème précédent montre alors que f est de classe $\mathcal{C}^\infty$ sur $]-1,1[$. En prime, si $p \in \mathbb{N}$, $f^{(p)}(0)$ est le quotient du coefficient de z^p dans F par $p!$: $f^{(p)}(0) = (-1)^p/(p+1)!$.

Exponentielle complexe

Revenons rapidement sur la définition et quelques propriétés de la fonction $z \mapsto \exp z$ sur $\mathbb{C}$. Considérons la série entière $F := \sum_{n\geqslant 0} z^n/n!$. Nous avons vu que son rayon de convergence est $+\infty$. La somme de cette série entière est donc une fonction continue de $\mathbb{C}$ dans $\mathbb{C}$ (théorème 33 de la page 505), c'est par définition la fonction exponentielle $\exp$:

$$\exp z \ := \ \sum_{n=0}^{+\infty} \frac{z^n}{n!} \quad \text{pour tout} \quad z \in \mathbb{C}. \tag{26}$$

Ainsi $\exp 0 = 1$. La première propriété fondamentale de l'exponentielle est la suivante :

$$\exp(a+b) \ = \ (\exp a)(\exp b) \quad \text{pour tous} \quad a, b \in \mathbb{C}. \tag{27}$$

Chacune des séries $\sum_{p\geqslant 0} a^p/p!$ et $\sum_{q\geqslant 0} b^q/q!$ converge absolument. Par produit de Cauchy, on en déduit que $(\exp a)(\exp b) = \sum_{n=0}^{+\infty} c_n$, en posant, pour tout $n \in \mathbb{N}$:

$$c_n \ := \ \sum_{p=0}^{n} \frac{a^p}{p!}\,\frac{b^{n-p}}{(n-p)!} \ = \ \frac{1}{n!}\sum_{p=0}^{n} \binom{n}{p} a^p b^{n-p} \ = \ \frac{(a+b)^n}{n!}\cdot$$

La dernière égalité est la formule du binôme. Ainsi $\sum_{n=0}^{+\infty} c_n = \exp(a+b)$, comme désiré.
La formule (27) montre que $\forall a \in \mathbb{C},\ \exp(a)\exp(-a) = 1$, d'où l'on tire $\exp a \neq 0$ et $(\exp a)^{-1} = \exp(-a)$. Ainsi, $\exp : (\mathbb{C}, +) \to (\mathbb{C}^*, \times)$ est un morphisme de groupes.
Une autre propriété essentielle de F est que $F' = F$. Plus précisément, F est la seule série entière ayant cette propriété et dont le terme constant soit 1. Soit en effet $G := \sum_{n\geqslant 0} a_n z^n$ une série entière. Comme $G' := \sum_{n\geqslant 0} (n+1)a_{n+1} z^n$, l'égalité $G' = G$ est vraie si, et seulement si, $(n+1)a_{n+1} = a_n$ pour tout n, d'où aussitôt notre assertion.
Comme application, fixons $z \in \mathbb{C}$ et définissons $f : \mathbb{R} \to \mathbb{C}$ par $f(t) := \exp(tz)$. Pour tout t réel, $f(t) = \sum_{n=0}^{+\infty} a_n t^n$, avec $a_n := z^n/n!$ pour tout n. D'après le théorème 41 de la page 515, f est de classe $\mathcal{C}^1$, et $f'(t) = \sum_{n=0}^{+\infty} (n+1)a_{n+1} t^n = \sum_{n=0}^{+\infty} z^{n+1} t^n/n! = z\exp(tz)$ pour tout $t \in \mathbb{R}$. En fait, f est la seule fonction de classe $\mathcal{C}^1$ de $\mathbb{R}$ dans $\mathbb{C}$ telle que $f(0) = 1$ et $f'(t) = zf(t)$ pour tout $t \in \mathbb{R}$. Soit en effet g une fonction ayant les mêmes propriétés. Puisque f ne s'annule pas, $h := g/f : \mathbb{R} \to \mathbb{C}$ est de classe $\mathcal{C}^1$, et un calcul

immédiat montre que $h' = 0$. Ainsi h est constante. Comme $f(0) = 1 = g(0)$, on a $h = 1$, c'est-à-dire $g = f$. Retenons la formule suivante, dans laquelle $z \in \mathbb{C}$ est fixé :

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\exp(tz)\big) \;=\; z\exp(tz) \quad \text{pour tout} \quad t \in \mathbb{R}. \tag{28}$$

Une autre notation usuelle de $\exp z$ est e^z. On peut définir e^z comme limite, ainsi que l'exercice suivant le montre.

Exercice 11.

Démontrer pour tout nombre complexe z la formule suivante :

$$\sum_{n=0}^{+\infty} \frac{z^n}{n!} \;=\; \lim_{n\to+\infty} \left(1+\frac{z}{n}\right)^n. \tag{29}$$

Solution. Notons e^z le premier membre de cette égalité. Fixons un nombre complexe z et un réel $\varepsilon > 0$. La série $\sum\limits_{n\geqslant 0} z^n/n!$ convergeant absolument, il existe un entier $p \geqslant 2$ tel que $\sum\limits_{k>p} |z|^k/k! \leqslant \varepsilon$. Fixons un tel p. Pour tout entier $n > p$, la formule du binôme donne :

$$\left(1+\frac{z}{n}\right)^n - e^z = 1 + z + \sum_{k=2}^{n} \binom{n}{k}\frac{z^k}{n^k} - \sum_{k=0}^{+\infty}\frac{z^k}{k!} = \sum_{k=2}^{n}\binom{n}{k}\frac{z^k}{n^k} - \sum_{k=2}^{+\infty}\frac{z^k}{k!}\cdot$$

$$= \underbrace{\sum_{k=2}^{p}\left(\frac{n(n-1)\cdots(n-k+1)}{n^k} - 1\right)\frac{z^k}{k!}}_{S_1(n)} + \underbrace{\sum_{k=p+1}^{n}\binom{n}{k}\frac{z^k}{n^k}}_{S_2(n)} - \underbrace{\sum_{k=p+1}^{+\infty}\frac{z^k}{k!}}_{S_3}\cdot$$

Puisque z et p sont fixés, il est clair que $\lim\limits_{n\to+\infty} S_1(n) = 0$. Il existe donc $N > p$ tel que $|S_1(n)| \leqslant \varepsilon$ pour tout $n \geqslant N$. Par ailleurs, $|S_3| \leqslant \varepsilon$, vu le choix de p. De même $|S_2(n)| \leqslant \varepsilon$ dès que $n > p$ car, si $k \in [\![p+1, n]\!]$, on a :

$$0 \;\leqslant\; \binom{n}{k}\frac{1}{n^k} \;=\; \frac{n(n-1)\cdots(n-k+1)}{k!\,n^k} \;\leqslant\; \frac{1}{k!}\cdot$$

Ainsi $|(1+z/n)^n - e^z| \leqslant 3\varepsilon$ pour tout $n \geqslant N$, ce qui établit la formule (29).

Une fois l'exponentielle définie sur $\mathbb{C}$ comme somme d'une série entière, on définit sans difficulté, *sur* $\mathbb{C}$, les fonctions $\cosh, \sinh, \cos, \sin$ par les formules suivantes :

$$\cosh z := \frac{e^z + e^{-z}}{2},\ \sinh z := \frac{e^z - e^{-z}}{2},\ \cos z := \frac{e^{iz} + e^{-iz}}{2},\ \sin z := \frac{e^{iz} - e^{-iz}}{2i}\cdot$$

Chacune de ces fonctions est la somme d'une série entière de rayon de convergence $+\infty$. On obtient les égalités suivantes, valables pour tout $z \in \mathbb{C}$:

$$\cos z \;=\; \sum_{p=0}^{+\infty} \frac{(-1)^p}{(2p)!}\, z^{2p}, \qquad \sin z \;=\; \sum_{p=0}^{+\infty} \frac{(-1)^p}{(2p+1)!}\, z^{2p+1}, \tag{30}$$

$$\cosh z \;=\; \sum_{p=0}^{+\infty} \frac{1}{(2p)!}\, z^{2p}, \qquad \sinh z \;=\; \sum_{p=0}^{+\infty} \frac{1}{(2p+1)!}\, z^{2p+1}. \tag{31}$$

En effet, si $n \in \mathbb{N}$, $i^n - (-i)^n$ vaut 0 si n est pair et $2i(-1)^p$ si $n := 2p+1$ est impair. D'où :

$$\sin z \;=\; \frac{1}{2i} \sum_{n=0}^{+\infty} \frac{\big[i^n + (-i)^n\big] z^n}{n!} \;=\; \frac{1}{2i} \sum_{p=0}^{+\infty} \frac{2i(-1)^p z^{2p+1}}{(2p+1)!}\cdot$$

On en déduit la seconde formule (30), et les autres formules se démontrent de même. À partir de la formule (27), on établit facilement les formules usuelles de trigonométrie, comme $\cos(a+b) = \cos a \cos b - \sin a \sin b$ ou $\cosh(a+b) = \cosh a \cosh b + \sinh a \sinh b$. Nous laissons le lecteur retrouver la formule usuelle $\cos^2 z + \sin^2 z = 1$ lorsque $z \in \mathbb{C}$. D'après l'exercice précédent, la restriction à $\mathbb{R}$ de la fonction sinus, par exemple, n'est autre que la fonction sinus définie dans le module IV.3 de [L1] (le vérifier !).

Exercice 12.

Développement en série entière de la fonction $f : x \mapsto \cos x \cosh x$?

Solution. La solution naturelle est d'effectuer le produit. Le théorème 36 de la page 510 montre que f est somme de la série entière $\sum_{n\geqslant 0} c_n z^{2n}$, de rayon de convergence $+\infty$, où $c_n := \sum_{p+q=n} \dfrac{(-1)^p}{(2p)!\,(2q)!}$ pour tout $n \in \mathbb{N}$. Cette expression de c_n n'est pas très satisfaisante. Il y a en fait beaucoup plus simple. Soit $x \in \mathbb{C}$. On a :

$$\begin{aligned} 4\cos x \cosh x &:= (e^x + e^{-x})(e^{ix} + e^{-ix}) = e^{(1+i)x} + e^{(1-i)x} + e^{(-1+i)x} + e^{(-1-i)x} \\ &= \sum_{k=0}^{+\infty} \big[(1+i)^k + (1-i)^k + (-1+i)^k + (-1-i)^k\big] \frac{x^k}{k!} = \sum_{k=0}^{+\infty} \frac{a_k z^k}{k!}\cdot \end{aligned}$$

Si k est impair, a_k est évidemment nul. Si $k := 2n$ est pair, il vient :

$$a_k = 2\big[(1+i)^{2n} + (1-i)^{2n}\big] = 2\big[(2i)^n + (-2i)^n\big] = 2^{n+1}\big[i^n + (-i)^n\big].$$

Ainsi $a_{2n} = 0$ si n est impair et, si $n := 2m$ est pair, $a_{2n} = 2^{2m+2}(-1)^m$. On en déduit que, si $n \in \mathbb{N}$, $c_n = 0$ lorsque n n'est pas multiple de 4 et, lorsque $n := 4m$, $c_n = (-1)^m 2^{2m}/(4m)!$. En vertu de la remarque de la page 517, nous avons ainsi démontré, pour tout $n \in \mathbb{N}$, la formule suivante :

$$\sum_{p=0}^{n} \frac{(-1)^p}{(2p)!\,(2n-2p)!} \;=\; \begin{cases} 0 & \text{si } n \text{ n'est pas multiple de 4} \\ \frac{(-1)^m 2^{2m}}{(4m)!} & \text{si } n := 4m \text{ , d'où :} \end{cases}$$

$$\cos x \cosh x \;=\; \sum_{m=0}^{+\infty} \frac{(-1)^m 2^{2m}}{(4m)!}\, z^{4m}.$$

Exercice 13.

Développement en série entière de la fonction $f : x \mapsto \ln(1 + x + x^2)$?

Solution. Là encore, l'idée naturelle, consistant à écrire f comme composée de $x \mapsto x+x^2$ et de $y \mapsto \ln(1+y)$, n'est pas la meilleure. Observons d'abord que f est définie sur $\mathbb{R}$, car $1+x+x^2 > 0$ pour tout x (signe du trinôme). L'idée est que, si $|x| < 1$, l'égalité $1+x+x^2 = (1-x^3)/(1-x)$ donne $f(x) = \ln(1-x^3) - \ln(1-x)$, d'où :

$$f(x) = \sum_{n=1}^{+\infty} \frac{x^n}{n} - \sum_{n=1}^{+\infty} \frac{x^{3n}}{n} = \sum_{p=1}^{+\infty} a_p x^p,$$

en posant, pour tout $p \in \mathbb{N}^*$, $a_p = 1/p$ si p n'est pas multiple de 3 et $a_p = -2/p$ si p est multiple de 3. Ce développement de f en série entière est valable sur $]-1,1[$. Soit R le rayon de convergence de la série entière obtenue. On a donc $R \geqslant 1$. En fait $R := 1$ car, si $r > 1$, $|a_p|\, r^p \geqslant r^p/p$, donc $|a_p|\, r^p$ tend vers $+\infty$ lorsque p tend vers $+\infty$.

Exercice 14.

Calculer $f(x) := \displaystyle\sum_{n=0}^{+\infty} \frac{1}{(3n)!}\, x^{3n}$ pour $x \in \mathbb{R}$.

Solution. L'exponentielle doit sûrement intervenir ! L'idée est de partir de la racine cubique de l'unité $j := \exp(2i\pi/3) = (-1+i\sqrt{3})/2$. Nous savons que $1+j+j^2 = 0$ et $j^3 = 1$. On en déduit que, pour tout entier k, $1+j^k+j^{2k}$ vaut 0 si k n'est pas multiple de 3 et vaut 3 si k est multiple de 3 (remplacer k par son reste modulo 3). On a ensuite :

$$\exp(x) = \sum_{k=0}^{+\infty} \frac{x^k}{k!}, \qquad \exp(jx) = \sum_{k=0}^{+\infty} \frac{j^k x^k}{k!}, \qquad \exp(j^2 x) = \sum_{k=0}^{+\infty} \frac{j^{2k} x^k}{k!}.$$

Compte tenu du calcul précédent, il vient :

$$e^x + e^{jx} + e^{j^2 x} = \sum_{k=0}^{+\infty} \frac{1+j^k+j^{2k}}{k!}\, x^k = 3 \sum_{n=0}^{+\infty} \frac{x^{3n}}{(3n)!}.$$

Le premier membre de cette égalité n'est pas satisfaisant, car f est à valeurs réelles. Mais :

$$\exp(jx) = \exp\big((-1+i\sqrt{3})x/2\big) = e^{-x/2} e^{ix\sqrt{3}/2} = e^{-x/2}\big[\cos\big(x\sqrt{3}/2\big) + i \sin\big(x\sqrt{3}/2\big)\big],$$

et de même $\exp(j^2 x) = e^{-x/2}\big[\cos\big(x\sqrt{3}/2\big) - i \sin\big(x\sqrt{3}/2\big)\big]$. D'où l'expression réelle :

$$f(x) \;=\; \frac{e^x + 2e^{-x/2} \cos\big(x\sqrt{3}/2\big)}{3}.$$

3.4 Fonctions développables en série entière

Soient $R > 0$ et $I :=]-R, R[$. Une fonction $f : I \to \mathbb{C}$ qui est développable en série entière sur I a des propriétés remarquables, par exemple elle est de classe $\mathcal{C}^\infty$, comme le montre le théorème 45 de la page 517. Nous connaissons déjà des exemples de telles fonctions. Avec $R := +\infty$, il y a les fonctions $\exp, \cos, \sin$ et $\cosh, \sinh$. Avec $R := 1$, il y a les fonctions $x \mapsto \ln(1+x)$, $x \mapsto \arctan x$ et $x \mapsto (1-x)^{-(p+1)}$, où $p \in \mathbb{N}$.

La proposition ci-dessous est une conséquence des résultats des sections précédentes.

Proposition 46. Considérons deux fonctions f et g définies sur un ouvert D de $\mathbb{C}$ (resp. de $\mathbb{R}$).

1. Si f et g sont DSE sur D, alors f', $f+g$, fg et λf (où $\lambda \in \mathbb{C}$) sont DSE sur D.
2. Si f et g sont DSE_0 et si $g(0)=0$, la fonction $f \circ g$, définie dans un voisinage de 0 dans $\mathbb{C}$ (resp. $\mathbb{R}$), est également DSE_0.

Avant d'étudier les fonctions DSE de façon plus poussée, donnons une définition, dans laquelle J est un intervalle de $\mathbb{R}$:

Définition 7. Soit $f : J \to \mathbb{C}$ une fonction de classe $\mathcal{C}^\infty$ et a un point intérieur à J. La série $\sum\limits_{n\geqslant 0} f^{(n)}(a)(z-a)^n/n!$ est appelée *série de Taylor de f en a*.

Lorsque $a := 0$, cette série s'écrit : $\sum\limits_{n\geqslant 0} f^{(n)}(0)z^n/n!$, c'est une série entière.

Théorème 47. Soit f une fonction DSE sur I. Alors f est de classe $\mathcal{C}^\infty$ et, si une série entière F représente f sur I, F est nécessairement la série de Taylor de f en 0.

Démonstration. Soit F une série entière représentant f sur I (F existe, car f est DSE sur I), de sorte que $R_F \geqslant R$. Soit $g :]-R_F, R_F[\to \mathbb{C}$ la somme de la série entière F. Les restrictions de f et g à I sont donc égales. D'après le théorème 45 de la page 517, g est de classe $\mathcal{C}^\infty$, donc f est de classe $\mathcal{C}^\infty$. De plus, pour tout $n \in \mathbb{N}$, le coefficient de z^n dans F est $g^{(n)}(0)/n! = f^{(n)}(0)/n!$, donc F est la série de Taylor de f en 0. ∎

Pour une fonction f de classe $\mathcal{C}^\infty$ au voisinage de 0, il y a donc *au plus une* série entière représentant f au voisinage de 0, c'est la série de Taylor de f en 0. À ce stade, il est essentiel de noter les points suivants :

1. La série de Taylor de f en 0 peut très bien avoir un rayon de convergence nul. Dans ce cas, f n'est pas développable en série entière au voisinage de 0. Voici un exemple. Pour tout $n \in \mathbb{N}$, définissons une fonction $f_n : \mathbb{R} \to \mathbb{C}$ par $f_n(x) = 2^{-n}e^{in^2x}$. Cette fonction est de classe $\mathcal{C}^\infty$ et, pour tout $x \in \mathbb{R}$, $f_n^{(p)}(x) = 2^{-n}(in^2)^p e^{in^2x}$, d'où $\left|f_n^{(p)}(x)\right| = 2^{-n}n^{2p}$. La série $\sum\limits_{n\geqslant 0} n^{2p}\,2^{-n}$ étant convergente (la règle de d'Alembert s'applique aisément), la série de fonctions $\sum\limits_{n\geqslant 0} f_n^{(p)}$ converge donc normalement. En vertu du théorème 28 de la page 495, la fonction $f : x \mapsto \sum\limits_{n=0}^{+\infty} f_n(x)$ est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$, et l'on a en particulier, pour tout $p \in \mathbb{N}$: $f^{(p)}(0) = \sum\limits_{n=0}^{+\infty} 2^{-n}(in^2)^p$. Remplaçons p par $4p$: $f^{(4p)}(0) = \sum\limits_{n=0}^{+\infty} 2^{-n}n^{8p}$ pour tout $p \in \mathbb{N}$. On peut minorer $f^{(4p)}(0)$

par le terme d'indice $n := 2p$ de cette série : $f^{(4p)}(0) \geqslant 2^{-2p}(2p)^{8p} = 2^{6p}p^{8p}$.

Soit $F := \sum_{k\geqslant 0} a_k z^k$ la série de Taylor de f en 0. Pour tout k, $a_k := f^{(k)}(0)/k!$. Soit $p \in \mathbb{N}$. Alors a_{4p} est réel et minoré par $2^{6p}p^{8p}/(4p)!$. Comme $(4p)!$ est le produit de $4p$ entiers majorés par $4p$, on a $(4p)! \leqslant (4p)^{4p}$, d'où $a_{4p} \geqslant 2^{-2p}p^{4p} \geqslant p^{2p}$ dès que $p \geqslant 2$. Cela étant, soit $r > 0$. Pour tout $p \in \mathbb{N}$, $a_{4p}r^{4p} \geqslant (pr^2)^{2p}$, donc $a_{4p}r^{4p}$ tend vers $+\infty$ lorsque p tend vers $+\infty$. Le rayon de convergence de la série entière F est donc nul.

2. Même si la série de Taylor de f en 0 a un rayon de convergence strictement positif, sa somme peut ne pas coïncider avec f dans un voisinage de 0 dans $\mathbb{R}$. Dans ce cas, f n'est pas développable en série entière au voisinage de 0. Voici un exemple. Soit $f : \mathbb{R} \to \mathbb{R}$ la fonction définie par $f(0) := 0$ et $f(x) := \exp(-1/x^2)$ si $x \neq 0$. La fonction f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^*$. Lorsque $x \neq 0$ tend vers 0, $-1/x^2$ tend vers $-\infty$, et $\exp(-1/x^2)$ tend vers 0, donc f est continue en 0. Pour tout $k \in \mathbb{N}$, on a aussi $\lim\limits_{x\to 0} x^{-k}\exp(-1/x^2) = 0$ car l'exponentielle l'emporte sur toute puissance de x. Ainsi $\lim\limits_{x\to 0} \exp(-1/x^2)/x = 0$, donc f est dérivable en 0 et $f'(0) = 0$. On montre ensuite, par récurrence sur p, qu'il existe une fraction rationnelle g_p telle que $f^{(p)} = g_p f$ sur $\mathbb{R}^*$. Toujours par récurrence sur p, il en résulte que f est p fois dérivable et $f^{(p)}(0) = 0$. Finalement, f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$, et toutes ses dérivées en 0 sont nulles. La série de Taylor de f en 0 est donc (identiquement) nulle ; son rayon de convergence est $+\infty$, mais sa somme n'a rien à voir avec f (elle est nulle !).

3. Supposons que f soit développable en série entière au voisinage de 0. Pour expliciter *la* série entière représentant f au voisinage de 0, le meilleur moyen n'est pas toujours de calculer les coefficients de Taylor $f^{(n)}(0)/n!$. Il peut être plus judicieux d'utiliser des développements en série entière « usuels » et des opérations sur les séries entières.

Formule du binôme généralisée

Il s'agit une application du théorème précédent. Soit α un nombre complexe fixé. La fonction puissance $t \mapsto t^\alpha$ est définie sur $]0,+\infty[$ par $t^\alpha := \exp(\alpha \ln t)$. La fonction $x \mapsto (1+x)^\alpha$ est donc définie, sur $]-1,+\infty[$, par la formule $(1+x)^\alpha := \exp\big((\alpha \ln(1+x)\big)$. Si $\alpha \in \mathbb{N}$, la formule du binôme donne les coefficients du polynôme $(1+x)^\alpha$. Le théorème ci-dessous est une généralisation de cette formule : on remplace les sommes finies par des sommes de séries. Rappelons que, pour tout $p \in \mathbb{N}$ et tout nombre complexe t, on pose :

$$\binom{t}{p} := \frac{t(t-1)\cdots(t-p+1)}{p!} \quad \text{et en particulier} \quad \binom{t}{0} := 1.$$

Lorsque t est entier et $t \geqslant p$, on retrouve le coefficient du binôme habituel.

Théorème 48. Soit α un nombre complexe fixé. Pour tout $x \in]-1,1[$, on a :

$$(1+x)^\alpha = \sum_{n=0}^{+\infty} \binom{\alpha}{n} x^n \quad \text{(formule du binôme généralisée).} \tag{32}$$

Si $\alpha \notin \mathbb{N}$, le rayon de convergence de la série entière figurant dans le second membre est égal à 1.

Il est important de noter que, lorsque $k \in \mathbb{N}$ et $\alpha := -(k+1)$, cette formule ne généralise que partiellement la formule (20) de la page 510, dans laquelle z était *complexe* (et $|z|<1$). La raison est que $\ln(1+x)$ n'est *a priori* défini que si x est réel et $x > -1$. Voir cependant le module II.6 pour le cas où z est complexe.

Démonstration. Si $\alpha \in \mathbb{N}$, il s'agit de la formule du binôme ordinaire. Supposons donc que $\alpha \notin \mathbb{N}$. D'abord la fonction $\varphi : x \mapsto (1+x)^\alpha$ est de classe $\mathcal{C}^\infty$ sur $]-1,+\infty[$. En effet, à partir de la formule $\varphi(x) = \exp\bigl(\alpha \ln(1+x)\bigr)$, la règle de dérivation d'une fonction composée montre que φ est dérivable et $\varphi'(x) = \alpha\varphi(x)(1+x)^{-1} = \alpha(1+x)^{\alpha-1}$ pour tout $x > -1$. Un raisonnement par récurrence évident montre que, pour tout $n \in \mathbb{N}^*$, φ est n fois dérivable et $\varphi^{(n)}(x) = \alpha(\alpha-1)\cdots(\alpha-n+1)(1+x)^{\alpha-n}$ pour tout $x > -1$. En particulier, pour tout $n \in \mathbb{N}^*$, $\varphi^{(n)}(0) = \alpha(\alpha-1)\cdots(\alpha-n+1)$. On en déduit que la série entière figurant dans le second membre de la formule (32) n'est autre que la série de Taylor de φ en 0. Cela ne suffit pas, nous l'avons dit, à établir le théorème, car nous ne savons pas *a priori* que φ est DSE sur $I :=]-1,1[$. Procédons alors de manière indirecte.

Posons $F := \sum_{n\geqslant 0} z^n/n!$ et $G := \alpha \sum_{n\geqslant 1} (-1)^{n-1} z^n/n$. Le rayon de convergence de la série entière F est $+\infty$, et sa somme $f : \mathbb{C} \to \mathbb{C}$ est la fonction exponentielle. Le rayon de convergence de la série entière G est 1, et sa somme g est définie sur $D(0,1)$. Nous savons que $g(x) := \alpha \ln(1+x)$ pour tout $x \in]-1,1[$.

Puisque le terme constant de G est nul, nous pouvons considérer la série entière composée $H := F \circ G$, écrivons-la $H := \sum_{n\geqslant 0} c_n z^n$. La remarque de la page 514 montre que le rayon de convergence R de H est au moins égal à 1 et que, si $h : D(0,1) \to \mathbb{C}$ est la somme de cette série entière, on a $h(z) = f\bigl(g(z)\bigr) = \exp\bigl(g(z)\bigr)$ pour tout $z \in D(0,1)$. En particulier, pour tout $x \in]-1,1[$, $h(x) = \exp\bigl(\alpha \ln(1+x)\bigr) = (1+x)^\alpha$. Cela montre que H représente la fonction φ sur I. Cette fonction est ainsi DSE sur I.

Pour identifier H, il est inutile de composer vraiment F et G, car nous savons que H est la série de Taylor de φ en 0 (théorème 47 de la page 522). Ainsi H est la série entière apparaissant dans la formule (32). Il reste à montrer que $R = 1$. Si $n \in \mathbb{N}$, c_{n+1}/c_n est le quotient de $\binom{\alpha}{n+1}$ par $\binom{\alpha}{n}$, d'où $c_{n+1}/c_n = (\alpha-n)/(n+1)$. On en déduit que c_{n+1}/c_n tend vers -1 lorsque n tend vers $+\infty$, d'où $R = 1$ d'après la proposition 30 de la page 503. ∎

Le cas $\alpha := -1/2$ est intéressant, nous allons le voir. Pour tout $n \in \mathbb{N}^*$, on a :

$$\binom{-1/2}{n} = \frac{1}{n!} \times (-1/2) \times (-1/2-1) \times \cdots \times (-1/2-n+1) = \frac{(-1)^n}{2^n n!} \times 1 \times 3 \times \cdots \times (2n-1).$$

Multiplions le numérateur et le dénominateur par $2\times 4\times\cdots\times(2n) = 2^n n!$. Au numérateur apparaît le produit de tous les entiers (pairs ou impairs) de 1 à $2n$. D'où :

$$\binom{-1/2}{n} \;=\; \frac{(-1)^n(2n)!}{2^{2n}(n!)^2} \;=\; \frac{(-1)^n}{2^{2n}}\binom{2n}{n}.$$

La formule du binôme généralisé donne donc, en remplaçant x par $-x$:

$$\frac{1}{\sqrt{1-x}} \;=\; \sum_{n=0}^{+\infty}\frac{1}{2^{2n}}\binom{2n}{n}x^n \quad\text{puis}\quad \frac{1}{\sqrt{1-x^2}} \;=\; \sum_{n=0}^{+\infty}\frac{1}{2^{2n}}\binom{2n}{n}x^{2n}.$$

Ces développements sont valables pour tout $x\in\,]-1,1[$, le rayon de convergence de chacune des séries entières qui apparaissent valant 1. L'intérêt est que, sur l'intervalle $]-1,1[$, nous connaissons la primitive de $x\mapsto(1-x^2)^{-1/2}$ nulle en 0, c'est la fonction $\arcsin$. Nous pouvons donc appliquer le théorème 42 de la page 516. La fonction $\arcsin$ est développable en série entière sur $]-1,1[$, son développement étant donné par la formule :

$$\arcsin x \;=\; \sum_{n=0}^{+\infty}\frac{1}{(2n+1)2^{2n}}\binom{2n}{n}x^{2n+1} \quad\text{pour tout}\quad x\in\,]-1,1[\,. \tag{33}$$

Le rayon de convergence de la série entière figurant dans le second membre vaut 1. Dans les calculs ci-dessus, si l'on remplace x par x^2 et non par $-x^2$, on obtient ceci :

$$\frac{1}{\sqrt{1+x^2}} \;=\; \sum_{n=0}^{+\infty}\frac{(-1)^n}{2^{2n}}\binom{2n}{n}x^{2n}.$$

La primitive de $x\mapsto(1+x^2)^{-1/2}$ nulle en 0 est la fonction argsh. D'où :

$$\operatorname{argsh} x \;:=\; \ln\bigl(x+\sqrt{1+x^2}\bigr) = \sum_{n=0}^{+\infty}\frac{(-1)^n}{(2n+1)2^{2n}}\binom{2n}{n}x^{2n+1} \quad\text{pour tout}\quad x\in\,]-1,1[\,.$$

Bien que la fonction argsh soit définie sur $\mathbb{R}$, la formule obtenue n'est valable que sur $]-1,1[$, puisque la série qui apparaît a pour rayon de convergence 1.

Remarque. Soient $F := \sum_{n\geqslant 0} z^n/n!$ et $G := \sum_{n\geqslant 1}(-1)^{n-1}z^n/n$. Ces séries entières ont pour rayon de convergence $+\infty$ et 1 respectivement. Notons $f:\mathbb{C}\to\mathbb{C}$ et $g:D(0,1)\to\mathbb{C}$ leurs sommes respectives. Les termes constants de G et $F-1$ étant nuls, on peut considérer les composées $F\circ G$ et $G\circ(F-1)$. Montrons que :

$$F\circ G \;=\; 1+z \quad\text{et}\quad G\circ(F-1) \;=\; z. \tag{34}$$

Soient $H := F\circ G$ et h sa somme. La remarque de la page 514 montre que le rayon de convergence de H est au moins égal à 1, et que $h(z) = f\bigl(g(z)\bigr) = \exp\bigl(g(z)\bigr)$ pour tout $z\in D(0,1)$. En particulier, si $x\in\,]-1,1[$, on a $g(x)=\ln(1+x)$ (*cf.* la proposition 43 de la page 516), donc $h(x) = \exp\bigl(\ln(1+x)\bigr) = 1+x$. L'égalité $H = 1+z$ résulte alors du théorème 47 de la page 522.

Soient de même $\Phi := G \circ (F-1)$ et φ sa somme. Si $z \in \mathbb{C}$ et si $|z|$ est assez petit, on a $\varphi(z) = g\big(\exp(z)-1\big)$. Si donc $x \in \mathbb{R}$ est assez voisin de 0 :

$$\varphi(x) = \ln\big(1+(\exp x - 1)\big) = \ln(\exp x) = x.$$

D'où $\Phi = z$ comme précédemment.

Exercice 15.

Soit F une série entière de terme constant non nul. Montrer qu'il existe une unique série entière G telle que $FG := 1$. Si le rayon de convergence de F est strictement positif, montrer qu'il en est de même pour G (on pourra utiliser la proposition 39 de la page 513).

Solution. On peut supposer que $F := \sum_{n\geqslant 0} a_n z^n$, avec $a_0 := 1$ (diviser F par son terme constant). Soit $G := \sum_{n\geqslant 0} b_n z^n$ une série entière. On a $FG := \sum_{n\geqslant 0} c_n z^n$, les c_n étant donnés par les formules (19) de la page 510. L'égalité $FG = 1$ est vraie si, et seulement si, $c_0 := 1$ et $c_n := 0$ pour tout $n \geqslant 1$. Cela revient à dire que $b_0 := 1$ et que, pour tout $n \geqslant 1$, on a $a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0 = 0$, soit $b_n = -(a_1 b_{n-1} + a_2 b_{n-2} + \cdots + a_n b_0)$. Ces conditions déterminent une unique suite (b_n) (construite par récurrence), d'où l'existence et l'unicité d'une série entière G telle que $FG = 1$.

Soit R_F le rayon de convergence de F, supposons $R_F > 0$. Pour prouver l'inégalité $R_G > 0$, on peut majorer directement les $|b_n|$ (*cf.* l'exercice II.2.38). Procédons indirectement, en considérant les séries entières $U := 1 - F$ et $\Phi := \sum_{n\geqslant 0} z^n$. On a $R_U = R_F$ et $R_\Phi = 1$. Soient f, u, φ les sommes respectives de F, U et Φ. Évidemment $f + u := 1$ et par ailleurs $\varphi(z) := (1-z)^{-1}$ pour tout $z \in D(0,1)$. Le terme constant de U étant nul, on peut considérer la série entière $H := \Phi \circ U$ et sa somme h. D'après la proposition 39 de la page 513, $R_H > 0$, et il existe un réel $r \in]0, R_F[$ tel que, dès que $|z| < r$, on ait $|u(z)| < 1$ et $h(z) = \varphi\big(u(z)\big)$, c'est-à-dire $h(z) = \big(1-u(z)\big)^{-1}$. La dernière égalité équivaut à $h(z)\big(1-u(z)\big) = 1$, ou encore à $h(z)f(z) = 1$. Ainsi $h(z)f(z) = 1$ pour tout $z \in D(0,r)$ ce qui montre que $HF = 1$, en vertu du théorème 47 de la page 522. Il en résulte que $H := G$, d'où $R_G > 0$.

Séries entières et équations différentielles

Les séries entières, comme les séries de Fourier, interviennent fréquemment dans la résolution d'équations différentielles (à ce sujet, nous renvoyons le lecteur au module II.7). Contentons-nous de deux exemples, d'autres seront proposés en exercice.

Retrouvons d'abord la formule du binôme généralisé en utilisant une équation différentielle adéquate. Soient $\alpha \in \mathbb{C}\setminus\mathbb{N}$ fixé et $f : I :=]-1,1[\to \mathbb{C}$ la fonction $x \mapsto (1+x)^\alpha$. Elle est de classe $\mathcal{C}^\infty$, et $f'(x) = \alpha(1+x)^{\alpha-1}$ pour tout x. Ainsi $(1+x)f' - \alpha f = 0$. L'idée est de chercher une fonction g solution de l'équation différentielle $(1+x)y' - \alpha y = 0$ et qui soit DSE_0, puis de prouver que $f = g$ grâce à la théorie des équations différentielles linéaires. Procédons d'abord formellement (*i.e.* sans nous occuper de convergence), en cherchant les

séries entières $F := \sum_{n\geqslant 0} a_n z^n \neq 0$ telles que $H := (1+z)F' - \alpha F$ soit nulle.

$$\begin{aligned} H &= (1+z)\sum_{n\geqslant 0}(n+1)a_{n+1}z^n - \alpha\sum_{n\geqslant 0} a_n z^n \\ &= \sum_{n\geqslant 0}\big((n+1)a_{n+1} - \alpha a_n\big)z^n + \sum_{n\geqslant 0}(n+1)a_{n+1}z^{n+1} \\ &= \sum_{n\geqslant 0}\big((n+1)a_{n+1} - \alpha a_n\big)z^n + \sum_{n\geqslant 0} n a_n z^n \\ &= \sum_{n\geqslant 0}\big[(n+1)a_{n+1} - (\alpha-n)a_n\big]z^n. \end{aligned}$$

Ainsi $H = 0$ si, et seulement si, $(n+1)a_{n+1} := (\alpha-n)a_n$ pour tout $n \in \mathbb{N}$. Ces conditions déterminent tous les a_n en fonction de a_0. Un calcul immédiat montre que $H := 0$ si, et seulement si, $a_n := a_0\binom{\alpha}{n}$ pour tout n. Il en résulte que la série entière F intervenant dans la formule du binôme généralisé vérifie l'égalité $(1+z)F' - \alpha F = 0$. Nous savons que le rayon de convergence de cette série est 1. Soit alors $g : I \to \mathbb{C}$ la somme de la série entière F.

C'est une fonction de classe $\mathcal{C}^1$, vérifiant l'équation différentielle $(1+x)y' - \alpha y = 0$. De plus $f(0) = 1 = g(0)$, d'où $f = g$, par exemple parce que la dérivée de g/f est nulle et $(g/f)(0) = 1$.

Exercice 16.

Montrer que l'équation différentielle $(E) : 4xy'' + 2y' + y = 0$ possède, à un coefficient multiplicatif près, une unique solution DSE sur $\mathbb{R}$, et expliciter cette solution.

Solution. On raisonne comme ci-dessus. Soit $F := \sum_{n\geqslant 0} a_n z^n$ une série entière. Voyons à quelles conditions $H := 4zF'' + 2F' + F$ est nulle. On a :

$$\begin{aligned} H &= 4z\sum_{n\geqslant 0}(n+2)(n+1)a_{n+2}z^n + 2\sum_{n\geqslant 0}(n+1)a_{n+1}z^n + \sum_{n\geqslant 0} a_n z^n \\ &= 4\sum_{n\geqslant 0} n(n+1)a_{n+1}z^n + \sum_{n\geqslant 0}\big[2(n+1)a_{n+1} + a_n\big]z^n \\ &= \sum_{n\geqslant 0}\big[2(n+1)(2n+1)a_{n+1} + a_n\big]z^n. \end{aligned}$$

Ainsi $H := 0$ si, et seulement si, $a_{n+1} := -\big[(2n+1)(2n+2)\big]^{-1}a_n$ pour tout $n \in \mathbb{N}$. Cette relation de récurrence se résout aisément : $H := 0$ si, et seulement si, $a_n = (-1)^n a_0/(2n)!$ pour tout n. Posons donc $F := \sum_{n\geqslant 0}(-1)^n z^n/(2n)!$. À un coefficient multiplicatif près, F est la seule série entière telle que $4zF'' + 2F' + F = 0$. Le rayon de convergence de F est clairement $+\infty$. La somme $f : \mathbb{R} \to \mathbb{R}$ de cette série entière est alors une solution de

l'équation (E) (*cf.* le théorème 45 de la page 517), et ce qui précède montre que c'est, à un coefficient multiplicatif près, la seule solution de (E) développable en série entière sur $\mathbb{R}$. Il est facile d'expliciter f, car on reconnaît les développements des fonctions cos et cosh, *cf.* les formules (30) et (31) de la page 519. Pour tout $x \in \mathbb{R}$, on a

$$f(x) := \cos(\sqrt{x}) \quad \text{si} \quad x \geqslant 0 \qquad \text{et} \qquad f(x) := \cosh(\sqrt{-x}) \quad \text{si} \quad x \leqslant 0.$$

Caractérisation des fonctions DSE_0

Soit f une fonction de classe $\mathcal{C}^\infty$ sur un intervalle $I :=]-R, R[$ (avec $R > 0$, éventuellement $R := +\infty$) et $F := \sum_{n\geqslant 0} a_n x^n$ sa série de Taylor en 0. À quelle condition F représente-t-elle f au voisinage de 0 ? Pour tout $x \in I$ et tout $n \in \mathbb{N}$, écrivons :

$$f(x) \;=\; \sum_{k=0}^{n} a_k x^k + R_n(x) \;=\; \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}\, x^k + R_n(x), \qquad (*)$$

$R_n(x)$ étant le reste dans la formule de Taylor à l'ordre n, appliquée à f en 0. Dire que F représente f sur I signifie que la série de fonctions $\sum_{n\geqslant 0} a_n x^n$ converge simplement vers f sur I, ou encore que, pour tout $x \in I$, la suite $\big(R_n(x)\big)$ converge vers 0. On comprend alors qu'une majoration convenable des restes $R_n(x)$, à l'aide de l'une des formules de Taylor, puisse permettre de prouver que f est DSE_0.

Théorème 49. Soit $f : I \to \mathbb{C}$ une fonction de classe $\mathcal{C}^\infty$. Pour que f soit DSE_0, il faut, et il suffit, qu'il existe trois réels strictement positifs C, ω et $r \leqslant R$ tels que, pour tout $n \in \mathbb{N}$ et tout $x \in]-r, r[$, on ait : $\big|f^{(n)}(x)\big| \leqslant Cn!\,\omega^n$.

Démonstration. Notons toujours $F := \sum_{n\geqslant 0} a_n x^n$ la série de Taylor de f en 0. Compte tenu du théorème 47 de la page 522, le problème est de voir si F représente f au voisinage de 0. En séparant la partie réelle et la partie imaginaire de f, on peut supposer que f est à valeurs réelles.

Supposons qu'il existe des réels $\omega, C > 0$ et $r \in]0, R]$ tels que les inégalités de l'énoncé soient valables. Soient $x \in]-r, r[$ et $n \in \mathbb{N}$. Puisque $\big|f^{(n+1)}(c)\big| \leqslant C(n+1)!\,\omega^{n+1}$ pour tout point c entre 0 et x, l'inégalité de Taylor-Lagrange donne $|R_n(x)| \leqslant C\big(|x|\,\omega\big)^{n+1}$. Posons $u := \min(r, \omega^{-1})$. Pour tout $x \in]-u, u[$, la suite $\big(R_n(x)\big)$ converge donc vers 0. Ainsi, la série de Taylor de f en 0 représente f sur $]-u, u[$.

Supposons inversement que f soit DSE_0. Il existe donc un réel r, $0 < r < \min(R, R_F)$, tel que F représente f sur $]-r, r[$. Fixons un réel $\rho \in]r, R_F[$. Par définition du rayon de convergence, la suite $(a_n\rho^n)$ est bornée : il existe un réel $M > 0$ tel que $|a_k|\,\rho^k \leqslant M$, c'est-à-dire $|a_k| \leqslant M\rho^{-k}$, pour tout $k \in \mathbb{N}$. Soit $n \in \mathbb{N}$. D'après le théorème 45 de la page 517, $f^{(n)}$ est représentée par $F^{(n)}$ sur $]-r, r[$ et, pour tout $x \in]-r, r[$, on a :

$$f^{(n)}(x) \;=\; \sum_{k=0}^{+\infty} (k+n)\cdots(k+2)(k+1)a_{n+k}x^k \;=\; n!\sum_{k=0}^{+\infty} \binom{n+k}{n} a_{n+k}x^k.$$

Pour tout $k \in \mathbb{N}$, on a $\left|x^k\right| \leqslant r^k$ et $|a_{n+k}| \leqslant M\rho^{-(n+k)}$, d'où :

$$\left|f^{(n)}(x)\right| \leqslant Mn!\,\rho^{-n} \sum_{k=0}^{+\infty} \binom{n+k}{n} \left(\frac{r}{\rho}\right)^k = \frac{Mn!\,\rho^{-n}}{\left(1-\frac{r}{\rho}\right)^{n+1}} = \frac{Mn!\,\rho}{(\rho-r)^{n+1}}.$$

La première des deux égalités résulte de la proposition 37 de la page 510. Il suffit ainsi de poser $C := M\rho(\rho-r)^{-1}$ et $\omega := (\rho-r)^{-1}$ pour que les inégalités de l'énoncé soient vérifiées. ∎

Exercice 17.

Soit $f : I \to \mathbb{R}$ une fonction de classe $\mathcal{C}^\infty$ telle que f et toutes ses dérivées soient positives sur I (on dit alors que f est *absolument monotone*). Montrer, à l'aide de la formule de Taylor et du théorème précédent, que f est DSE_0 (théorème de Bernstein).

Solution. Fixons un réel r tel que $0 < r < R/2$; soit M la borne supérieure de $|f|$ sur $[-r, 2r]$. Considérons un entier positif n et un point $x \in]-r, r[$. La formule de Taylor avec reste intégral donne :

$$f(2r) = \sum_{k=0}^{n} \frac{f^{(k)}(x)(2r-x)^k}{k!} + \int_x^{2r} \frac{f^{(n+1)}(t)(2r-t)^n}{n!}\,\mathrm{d}t.$$

Vu l'hypothèse, l'intégrale est positive et $f^{(k)}(x) \geqslant 0$ pour $k = 0, \ldots, n$. On en déduit les inégalités $0 \leqslant f^{(n)}(x)(2r-x)^n/n! \leqslant f(2r) \leqslant M$, d'où $0 \leqslant (2r-x)^n f^{(n)}(x) \leqslant Mn!$. Il en résulte que $\left|f^{(n)}(x)\right| = f^{(n)}(x) \leqslant Mn!(2r-x)^{-n} \leqslant Mn!\,r^{-n}$. Le théorème précédent, appliqué en prenant $C := M$ et $\omega := 1/r$, montre que f est DSE_0.

En se référant à la preuve du théorème 49 de la page ci-contre, on a $u := r$, donc la série de Taylor de f en 0 représente f sur $]-r, r[$. Si par exemple $R := +\infty$, f est DSE sur $\mathbb{R}$ tout entier.

3.5 Exponentielle d'une matrice carrée

Soit $\mathcal{A}$ une algèbre de Banach sur $\mathbb{C}$: $\mathcal{A}$ est une $\mathbb{C}$-algèbre, munie d'une norme $\|\cdot\|$ telle que $\|ab\| \leqslant \|a\|\|b\|$ pour tous $a, b \in \mathcal{A}$, et en outre $\mathcal{A}$ est *complète*. Considérons une série entière $\sum_{n\geqslant 0} a_n z^n$, de rayon de convergence $R > 0$ et de somme f.

Proposition et définition 50. Soit A un élément de $\mathcal{A}$ tel que $\|A\| < R$. La série $\sum_{n\geqslant 0} a_n A^n$ de $\mathcal{A}$ est absolument convergente, et en particulier convergente. Sa somme sera notée $f(A)$, par abus de langage. Ainsi $f(A) := \sum_{n=0}^{+\infty} a_n A^n$.

L'application $f : B(0, R) \to \mathcal{A}$ est continue.

Démonstration. Pour tout $n \geqslant 1$, on a $\|A^n\| \leqslant \|A\|^n$, d'où $\|a_n A^n\| \leqslant |a_n|\,\|A\|^n$. Puisque $\|A\| < R$, la série $\sum_{n\geqslant 0} |a_n|\,\|A\|^n$ converge, donc $\sum_{n\geqslant 0} \|a_n A^n\|$ converge, par comparaison. Ainsi, la série $\sum_{n\geqslant 0} a_n A^n$ est absolument convergente. Cette série converge donc, puisque $\mathcal{A}$ est complète (théorème 41 de la page 431). Prouvons la continuité de f sur

$B(0,R)$; si $R := +\infty$, convenons que $B(0,R) := \mathcal{A}$. Soit $r \in]0,R[$. La série d'applications de $\mathcal{A}$ dans $\mathcal{A}$ définissant f converge normalement, et en particulier uniformément, sur la boule $B(0,r)$. Si $n \in \mathbb{N}$ est fixé, l'application $A \mapsto A^n$ de $\mathcal{A}$ dans elle-même est continue. La raison est que la multiplication de $\mathcal{A} \times \mathcal{A}$ dans $\mathcal{A}$ est continue. La continuité de f sur $B(0,r)$ résulte alors du théorème 8 de la page 480.

Pour la série entière $\sum_{n\geqslant 0} z^n/n!$, de rayon de convergence $+\infty$, la définition précédente permet donc de poser, pour tout $A \in \mathcal{A}$: $\exp A := \sum_{n\geqslant 0} \frac{1}{n!} A^n$ ($\exp A$ est aussi noté e^A).

L'application $\exp : \mathcal{A} \to \mathcal{A}$ est *continue*, d'après la proposition précédente.

Proposition 51. Si $A, B \in \mathcal{A}$ commutent ($AB = BA$), on a $e^{A+B} = e^A e^B$.

Démonstration. Il suffit de reprendre la démonstration faite dans le cas où $\mathcal{A} := \mathbb{C}$ (formule (27) de la page 518). Nous avions utilisé un résultat de [L1] (module IV.4) sur le produit de Cauchy. La preuve de ce résultat est encore valable lorsque $\mathbb{C}$ est remplacé par $\mathcal{A}$, la convergence absolue dans $\mathcal{A}$ remplaçant la convergence absolue dans $\mathbb{C}$. En effet, les trois ingrédients de ladite preuve étaient l'inégalité triangulaire, l'inégalité $|AB| \leqslant |A|\,|B|$, et le fait que la convergence absolue implique la convergence.

Intéressons-nous au cas où $\mathcal{A} := M_n(\mathbb{C})$, avec $n \geqslant 1$. Munissons par exemple $M_n(\mathbb{C})$ de la norme $|||\cdot|||$ subordonnée à une norme $\|\cdot\|$ sur $\mathbb{C}^n$ (*cf.* la section 4). Nous avons bien $|||AB||| \leqslant |||A|||\,|||B|||$ pour toutes matrices A, B, et $|||I_n||| = 1$. L'algèbre normée $\big(M_n(\mathbb{C}), |||\cdot|||\big)$ est complète, car de dimension finie. Ce qui précède peut donc être appliqué. Soit $A \in M_n(\mathbb{C})$. Par définition, l'exponentielle de A, notée $\exp A$ ou e^A, est donnée par la formule suivante (l'indice de sommation ne doit pas être n) :

$$\exp A := \sum_{k=0}^{+\infty} \frac{1}{k!} A^k = \lim_{m\to+\infty} \sum_{k=0}^{m} \frac{1}{k!} A^k. \tag{35}$$

Il faut bien comprendre d'abord que e^A est une *limite* (comme toute somme de série).

Exercice 18.

Soit $A \in M_n(\mathbb{C})$. Montrer que $\exp A$ est un polynôme en A, *i.e.* que $\exp A \in \mathbb{C}[A]$.

Solution. On a $\mathbb{C}[A] := \big\{P(A) \mid P \in \mathbb{C}[X]\big\}$. C'est un sous-espace vectoriel de $M_n(\mathbb{C})$. Comme nous sommes en dimension finie, $\mathbb{C}[A]$ est un e.v. n. complet, donc fermé dans $M_n(\mathbb{C})$. Or, les sommes partielles de la série définissant $\exp A$ appartiennent à $\mathbb{C}[A]$. La limite de ces sommes, c'est-à-dire e^A, appartient donc à $\mathbb{C}[A]$. Il existe donc $P \in \mathbb{C}[X]$ tel que $e^A = P(A)$. Mais attention : P n'est pas unique (on peut par exemple lui ajouter le polynôme minimal de A).

De plus, P dépend de A : il n'existe pas de polynôme $Q \in \mathbb{C}[X]$ tel que $e^M = Q(M)$ pour *toute* matrice M (penser au cas $n := 1$!).

Cet exercice montre en particulier que A *et* $\exp A$ *commutent*. La première propriété essentielle de l'exponentielle de matrices est donnée par la proposition ci-dessus : lorsque $A, B \in M_n(\mathbb{C})$ *commutent*, $e^{A+B} = e^A e^B$. Voici une conséquence :

Proposition 52. Si $A \in M_n(\mathbb{C})$, $\exp A$ est inversible et son inverse est $\exp(-A)$.

Démonstration. On a clairement $\exp(0) = I_n$. Comme A et $-A$ commutent, on en déduit que $e^A e^{-A} = e^{A+(-A)} = e^0 = I_n$, d'où la conclusion. ∎

Ainsi exp est une application *continue* de $M_n(\mathbb{C})$ dans $GL_n(\mathbb{C})$. Attention : si $n \geqslant 2$, exp n'est pas un morphisme du groupe additif $M_n(\mathbb{C})$ dans le groupe multiplicatif $GL_n(\mathbb{C})$ (*cf.* l'exercice II.2.42).
Le théorème ci-dessous est essentiel dans l'étude des systèmes différentiels linéaires à coefficients constants (*cf.* le module II.7, ainsi que le module IV.6 de [L1]).

Théorème 53. Soit $A \in M_n(\mathbb{C})$. La fonction $f : t \mapsto e^{tA}$ de $\mathbb{R}$ dans $M_n(\mathbb{C})$ est de classe $\mathcal{C}^1$, sa dérivée est $t \mapsto Ae^{tA}$. Si $X \in \mathbb{C}^n$ est un vecteur colonne fixé, la fonction $Y : t \mapsto e^{tA}X$ de $\mathbb{R}$ dans $\mathbb{C}^n$ est solution de l'équation différentielle $Y' = AY$.

Démonstration. Si $k \in \mathbb{N}$, notons f_k la fonction $t \mapsto (t^k/k!)A^k$. La série de fonctions $\sum_{k\geqslant 1} f_k$ converge simplement, et sa somme est f. Chaque f_k est de classe $\mathcal{C}^1$ et sa dérivée, nulle si $k := 0$, est $t \mapsto \big(t^{k-1}/(k-1)!\big)A^k$ si $k \geqslant 1$. La série de fonctions $\sum_{k\geqslant 1} f_k'$ s'écrit donc $\sum_{k\geqslant 0} f_k A$, c'est-à-dire $t \mapsto \sum_{k\geqslant 0} f_k(t)A$. Sur un segment $[-R, R]$ donné, la série $\sum_{k\geqslant 0} f_k A$ converge normalement, donc uniformément, car $|||f_k(t)A||| \leqslant (R^k/k!)|||A|||^{k+1}$.
D'après le théorème 24 de la page 492, f est de classe $\mathcal{C}^1$ sur $[-R, R]$ et, si $|t| \leqslant R$, $f'(t) = \sum_{k=0}^{+\infty} f_k(t)A$. L'application $M \mapsto MA$ de $M_n(\mathbb{C})$ dans $M_n(\mathbb{C})$ est linéaire, donc continue (on est en dimension finie). On en déduit l'égalité $\sum_{k=0}^{+\infty} f_k(t)A = \left(\sum_{k=0}^{+\infty} f_k(t)\right)A$ (passage à la limite), soit $f'(t) = e^{tA}A$. Enfin e^{tA} et A commutent, donc $f'(t) = Ae^{tA}$, et ceci pour tout $t \in \mathbb{R}$. D'où la première assertion du théorème.
Considérons l'application $Y : t \mapsto e^{tA}X = f(t)X$. L'application $M \mapsto MX$ de $M_n(\mathbb{C})$ dans $\mathbb{C}^n$ est linéaire, donc continue. On en déduit que Y est de classe $\mathcal{C}^1$ et que sa dérivée est $t \mapsto f'(t)X = Ae^{tA}X = AY$, d'où la deuxième assertion du théorème. ∎

Calcul de l'exponentielle d'une matrice

Commençons par le cas d'une matrice diagonale $D := \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$.
Pour tout $k \in \mathbb{N}$, $D^k = \mathrm{Diag}(\lambda_1^k, \ldots, \lambda_n^k)$. Ainsi e^D est limite, lorsque p tend vers $+\infty$,

de $\mathrm{Diag}\left(\sum_{k=0}^{p}\lambda_1^k/k!,\dots,\sum_{k=0}^{p}\lambda_n^k/k!\right)$, d'où la formule suivante :

$$\exp((\mathrm{Diag}(\lambda_1\dots,\lambda_n)) \;=\; \mathrm{Diag}(e^{\lambda_1},\dots,e^{\lambda_n}). \tag{36}$$

Le cas d'une matrice diagonalisable se déduit de la proposition suivante :

Proposition 54. Pour toutes matrices $A \in M_n(\mathbb{C})$ et $P \in GL_n(\mathbb{C})$, on a : $\exp\bigl(PAP^{-1}\bigr) = P\exp(A)P^{-1}$.

Démonstration. L'application $L : M \mapsto PMP^{-1}$ de $M_n(\mathbb{C})$ dans $M_n(\mathbb{C})$ est un automorphisme de l'algèbre $M_n(\mathbb{C})$, donc $PM^kP^{-1} = (PMP^{-1})^k$ pour tous $k \in \mathbb{N}$ et $M \in M_n(\mathbb{C})$. De plus L est linéaire, donc continue car $M_n(\mathbb{C})$ est de dimension finie. D'où :

$$\begin{aligned} P\exp(A)P^{-1} = L\left(\lim_{p\to+\infty}\sum_{k=0}^{p}\frac{1}{k!}A^k\right) &= \lim_{p\to+\infty}\sum_{k=0}^{p}\frac{1}{k!}L(A^k) \\ &= \lim_{p\to+\infty}\sum_{k=0}^{p}\frac{1}{k!}(PAP^{-1})^k = \exp\bigl(PAP^{-1}\bigr). \end{aligned}$$

Pour une matrice nilpotente $N \in M_n(\mathbb{C})$, $\exp(N)$ est une somme finie. En effet $N^n = 0$, donc $e^N = \sum_{k=0}^{n-1}\frac{1}{k!}N^k$. Dans le cas général, le calcul de $\exp A$ se ramène aux cas particuliers précédemment traités (matrices diagonalisables et matrices nilpotentes), vu le théorème suivant, qui fait intervenir la décomposition de Dunford (*cf.* le théorème 44 de la page 250) :

Théorème 55. Soit $A := D + N$ la décomposition de Dunford de la matrice $A \in M_n(\mathbb{C})$. Alors $e^A = e^De^N$.

Démonstration. Rappelons que D est diagonalisable, N est nilpotente, et $DN = ND$. Le fait que D et N commutent intervient ici de manière cruciale. C'est lui qui permet d'appliquer la proposition 51 de la page 530, donnant les égalités $e^A = e^{D+N} = e^De^N$.

Les valeurs propres de l'exponentielle d'une matrice A se déduisent de celles de A :

Proposition 56. Soit $A \in M_n(\mathbb{C})$. Si le polynôme caractéristique χ_A de A s'écrit $\prod_{j=1}^{n}(X-\alpha_j)$, le polynôme caractéristique de e^A est : $\prod_{j=1}^{n}(X-e^{\alpha_j})$. En particulier, le déterminant de e^A est $e^{\mathrm{Tr}\,A}$.

Démonstration. Supposons d'abord que $A := (a_{i,j})$ soit triangulaire supérieure, de sorte que χ_A s'écrit comme dans l'énoncé, avec $\alpha_j := a_{j,j}$ pour tout j. Pour tout $k \in \mathbb{N}$, $(1/k!)A^k$ est triangulaire supérieure, et ses termes diagonaux sont les $\alpha_j^k/k!$, $j = 1,\dots,n$.

Par sommation, on en déduit que e^A est triangulaire supérieure, ses termes diagonaux étant les e^{α_j}, $j = 1, \dots, n$. Le polynôme caractéristique de e^A est donc le produit des $X - \mathrm{e}^{\alpha_j}$. Dans le cas général, il existe $P \in GL_n(\mathbb{C})$ et une matrice triangulaire supérieure T telles que $A := PTP^{-1}$. Alors $\chi_A = \chi_T$, et de même e^A et e^T ont même polynôme caractéristique, puisque $e^A = Pe^TP^{-1}$. D'où la première assertion de l'énoncé. Le déterminant de e^A est le produit de $(-1)^n$ par le terme constant du polynôme caractéristique de e^A, c'est donc le produit des e^{α_j}. Par ailleurs la trace de A est la somme des α_j, et par suite $\det(e^A) = \prod_j e^{\alpha_j} = \exp\bigl(\sum_j \alpha_j\bigr) = \exp(\mathrm{Tr}\, A)$. ∎

On en déduit immédiatement le corollaire suivant :

Corollaire 57. Le spectre de e^A est formé des e^λ, λ décrivant le spectre de A. Plus précisément, si $\mu \in \mathbb{C}^*$ est valeur propre de e^A, sa multiplicité est la somme, pour tous les $\lambda \in \mathrm{sp}\, A$ tels que $e^\lambda = \mu$, des multiplicités des λ comme valeurs propres de A.

Exercice 19.

Calculer l'exponentielle de la matrice $A := \begin{pmatrix} -1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}$ (*cf.* l'exercice 13 de la page 250).

Solution. Il ne coûte pas plus cher de calculer e^{tA} pour tout $t \in \mathbb{R}$. La décomposition de Dunford $A := D + N$ de A a été donnée dans l'exercice précité, dont on reprend les notations et les calculs. Partons de l'égalité $D = I_3 - 2\Pi_1$, d'où $tD = tI_3 - 2t\Pi_1$ pour tout t. Puisque tI_3 et $-2t\Pi_1$ commutent, $\exp(tD) = \exp(tI_3)\exp(-2t\Pi_1) = e^t \exp(-2t\Pi_1)$. Rappelons que $\Pi_1^2 = \Pi_1$, d'où $\Pi_1^k = \Pi_1$ pour tout $k \geqslant 1$. Ainsi :

$$\exp(-2t\Pi_1) = I_3 + \sum_{k=1}^{+\infty} \frac{(-2t)^k}{k!}\,\Pi_1 = I_3 + (e^{-2t} - 1)\Pi_1, \quad \text{puis}$$

$$\exp(tD) = e^t\bigl[I_3 + (e^{-2t} - 1)\Pi_1\bigr] = e^t I_3 + (e^{-t} - e^t)\Pi_1 = \begin{pmatrix} e^{-t} & e^t - e^{-t} & e^{-t} - e^t \\ 0 & e^t & 0 \\ 0 & 0 & e^t \end{pmatrix}.$$

Passons à N. On a $N^2 = 0$, d'où : $\exp(tN) = I_3 + tN$. Enfin $tA = tD + tN$ et tD, tN commutent, d'où $\exp(tA) = \exp(tD)\exp(tN) = \exp(tD)(I_3 + tN)$, ce qui donne :

$$\exp(tA) \;=\; \begin{pmatrix} e^{-t} & e^t - e^{-t} & e^{-t} + (3t-1)e^t \\ 0 & e^t & 3te^t \\ 0 & 0 & e^t \end{pmatrix}.$$

Nous laissons le lecteur retrouver ce résultat en résolvant directement le système différentiel (triangulaire) correspondant à A, *cf.* le module II.7.

Pour terminer cette section, montrons que l'application exponentielle est *différentiable* :

Proposition 58. L'application $A \mapsto \exp A$ de $M_n(\mathbb{C})$ dans $M_n(\mathbb{C})$ est continûment différentiable. Sa différentielle en $0 \in M_n(\mathbb{C})$ est l'application identité de $M_n(\mathbb{C})$.

Démonstration. Posons $E := M_n(\mathbb{C})$. Soit f l'application $A \mapsto \exp A$ de E dans E. Pour tout $p \in \mathbb{N}$, notons f_p l'application $A \mapsto (1/p!)A^p$ de E dans E. Ainsi $f := \sum_{p=0}^{+\infty} f_p$, cette série de fonctions convergeant normalement sur tout compact de E. Soit $A \in E$ une matrice fixée. Il faut évaluer $f(A+H) - f(A) := e^{A+H} - e^A$ et voir ce qui se passe lorsque $H \in E$ tend vers 0. La difficulté est que, A et H ne commutant pas en général, e^{A+H} ne s'exprime pas simplement.

1. Soit $p \in \mathbb{N}^*$. L'application f_p est composée de l'application diagonale $M \mapsto (M, M, \dots, M)$ de E dans E^p, qui est linéaire, et de la multiplication $(M_1, \dots, M_p) \mapsto M_1 M_2 \cdots M_p$ de E^p dans E. Cette multiplication est p-linéaire, donc différentiable (*cf.* page 419). Il en résulte que f_p est différentiable au point A. En outre, vu la formule (20) de la page 419, on a :

$$(Df_p)_A(H) \;=\; \frac{1}{p!}\sum_{k=0}^{p-1} A^k H A^{p-k-1} \quad \text{pour toute matrice} \quad H \in E. \qquad (*)$$

   On en déduit la majoration $\|(Df_p)_A\| \leqslant \big(1/(p-1)!\big)\|A\|^{p-1}$. En développant $(A+H)^p$ comme somme de 2^p termes, on obtient aisément la majoration suivante, valable lorsque $\|H\| \leqslant 1$:

$$\|f_p(A+H) - f_p(A) - (Df_p)_A(H)\| \;\leqslant\; \frac{\big(1+\|A\|\big)^p \|H\|^2}{p!}. \qquad (**)$$

   L'application f_0 est constante, donc de différentielle nulle en tout point.

2. Compte tenu des majorations des $\|(Df_p)_A\|$, la série $\sum_{p\geqslant 1} (Df_p)_A$ de $\mathcal{L}(E)$ est absolument convergente, donc convergente. Soit $L \in \mathcal{L}(E)$ sa somme. Pour toute matrice $H \in E$ telle que $\|H\| \leqslant 1$, les inégalités $(**)$ donnent :

$$\big\|f(A+H) - f(A) - L(H)\big\| \;\leqslant\; \sum_{p=1}^{+\infty} \frac{\big(1+\|A\|\big)^p \|H\|^2}{p!} \;=\; \big[\exp\big(1+\|A\|\big) - 1\big]\|H\|^2.$$

   Il en résulte que f est différentiable au point A et que $Df_A = L$. Vu la formule $(*)$, on a de plus :

$$Df_A(H) \;=\; \sum_{p=1}^{+\infty} \frac{1}{p!}\left(\sum_{k=0}^{p-1} A^k H A^{p-k-1}\right) \quad \text{pour toute matrice} \quad H \in E. \qquad (37)$$

   Si $A := 0$, cette formule se réduit à $Df_0(H) = H$ (seul $p := 1$ intervient), donc Df_0 est l'application identité de E.

3. Il reste à montrer que l'application $A \mapsto Df_A$ de E dans $\mathcal{L}(E)$ est continue. Pour chaque p, l'application $A \mapsto (Df_p)_A$ de E dans $\mathcal{L}(E)$ est continue, comme composée de l'application diagonale de E dans E^{p-1} et d'une application $(p-1)$-linéaire de E^{p-1} dans $\mathcal{L}(E)$. D'où la conclusion, car $A \mapsto \sum_{p\geqslant 1} (Df_p)_A$ est une série d'applications de E dans $\mathcal{L}(E)$ convergeant normalement sur tout compact de E, toujours à cause des majorations des $\|(Df_p)_A\|$. ∎

Exercice type corrigé

II.2.0 1. Pour tout $n \in \mathbb{N}$, calculer l'intégrale de Wallis $I_{2n} = \displaystyle\int_0^{\pi/2} \sin^{2n} x \,\mathrm{d}x$, à l'aide de la formule $2i \sin x = e^{ix} - e^{-ix}$.

2. Sur l'intervalle $J :=]-1, 1[$, on considère l'équation différentielle (de Gauß) suivante :

$$(E) : \ (x^3 - x)y'' + (3x^2 - 1)y' + xy \ = \ 0.$$

Montrer que (E) possède une unique solution f, développable en série entière sur J et vérifiant la condition initiale $f(0) = 1$.

3. Pour tout $x \in J$, établir la formule suivante, à l'aide de séries entières appropriées :

$$f(x) \ = \ \frac{2}{\pi} \int_0^{\pi/2} \frac{\mathrm{d}t}{\sqrt{1 - x^2 \sin^2 t}}. \qquad (*)$$

Solution.

1. Soit $n \in \mathbb{N}$. La fonction $x \mapsto \sin^{2n} x$ est π-périodique et paire. Son intégrale sur $[0, \pi/2]$ est donc le quart de son intégrale sur $[0, 2\pi]$. Transformons $\sin^{2n} x$ à l'aide de la formule du binôme :

$$\begin{aligned} 4(2i)^{2n} I_{2n} &= \int_0^{2\pi} (e^{ix} - e^{-ix})^{2n} \,\mathrm{d}x \\ &= \int_0^{2\pi} \left(\sum_{k=0}^{2n} (-1)^k \binom{2n}{k} (e^{-ix})^k (e^{ix})^{2n-k} \right) \mathrm{d}x \\ &= \sum_{k=0}^{2n} (-1)^k \binom{2n}{k} \int_0^{2\pi} e^{(2n-2k)ix} \,\mathrm{d}x. \end{aligned}$$

Dans la dernière somme, seule l'intégrale correspondant à $k := n$ est non nulle. En effet, lorsque $p \in \mathbb{Z}$, l'intégrale sur $[0, 2\pi]$ de e^{ipx} vaut 2π si $p := 0$ et 0 sinon. D'où :

$$4(2i)^{2n} I_{2n} \ = \ 2\pi (-1)^n \binom{2n}{n} \quad \text{et donc} \quad I_{2n} \ = \ \frac{\pi \binom{2n}{n}}{2^{2n+1}}.$$

2. Cherchons une série entière $F := \displaystyle\sum_{n \geqslant 0} a_n z^n$ telle que $a_0 := 1$ et que la série entière $H := (z^3 - z)F'' + (3z^2 - 1)F' + zF$ soit nulle. Calculons les coefficients de H :

$$\begin{aligned} (z^3 - z)F'' &= \sum_{n \geqslant 0} (n+2)(n+1) a_{n+2} (z^{n+3} - z^{n+1}) \\ &= \sum_{n \geqslant 1} (n-1)(n-2) a_{n-1} z^n - \sum_{n \geqslant 0} n(n+1) a_{n+1} z^n \\ (3z^2 - 1)F' &= \sum_{n \geqslant 0} (n+1) a_{n+1} (3z^{n+2} - z^n) \\ &= 3 \sum_{n \geqslant 1} (n-1) a_{n-1} z^n - \sum_{n \geqslant 0} (n+1) a_{n+1} z^n \\ zF &= \sum_{n \geqslant 0} a_n z^{n+1} = \sum_{n \geqslant 1} a_{n-1} z^n. \end{aligned}$$

Le coefficient constant de H est $-a_1$. Si $n \geqslant 1$, le coefficient de z^n dans H vaut :

$$a_{n-1}\big[(n-1)(n-2)+3(n-1)+1\big] - (n+1)^2 a_{n+1} = n^2 a_{n-1} - (n+1)^2 a_{n+1}.$$

Ainsi $H = 0$ si, et seulement si, $a_1 = 0$ et $(n+1)^2 a_{n+1} = n^2 a_{n-1}$ pour tout $n \geqslant 1$. Ces relations imposent que a_k soit nul pour tout $k \in \mathbb{N}$ impair. Les a_{2p}, $p \in \mathbb{N}$, doivent satisfaire la relation de récurrence : $a_{2p+2} = \left(\dfrac{2p+1}{2p+2}\right)^2 a_{2p}$. On en déduit l'existence et l'unicité de F. Pour tout $p \in \mathbb{N}$, on obtient aisément les formules $a_{2p+1} = 0$ et

$$a_{2p} = \left(\frac{1 \times 3 \times \cdots \times (2p-1)}{2 \times 4 \times \cdots \times (2p)}\right)^2 = \left(\frac{(2p)!}{2^{2p}(p!)^2}\right)^2 = \frac{\binom{2p}{p}^2}{2^{4p}}.$$

Vu la relation de récurrence ci-dessus, a_{2p+2}/a_{2p} tend vers 1 lorsque p tend vers $+\infty$. Le rayon de convergence de F est donc 1 (règle de d'Alembert). Soit alors $f : J \to \mathbb{R}$ la somme de la série entière F. Le théorème 45 de la page 517 montre que f est solution de l'équation (E), et c'est la seule solution possédant les propriétés voulues, compte tenu de l'unicité de F. Pour tout $x \in J$, on a explicitement :

$$f(x) \;=\; \sum_{p=0}^{+\infty} \frac{\binom{2p}{p}^2}{2^{4p}}\, x^{2p}. \qquad (**)$$

3. Soit $g : J \to \mathbb{R}$ le second membre de l'égalité $(*)$. Il n'est pas tout à fait évident de montrer que g est une solution de (E). On va plutôt développer g en série entière, puis comparer avec l'égalité $(**)$. Partons de la formule suivante, prouvée dans la section 3.4 :

$$\frac{1}{\sqrt{1-u^2}} \;=\; \sum_{n=0}^{+\infty} \frac{1}{2^{2n}} \binom{2n}{n} u^{2n}.$$

Ce développement est valable dès que $u \in J$. Si $x \in J$ et $t \in [0, \pi/2]$ sont fixés, remplaçons u par $x \sin t$:

$$\frac{1}{\sqrt{1-x^2\sin^2 t}} \;=\; \sum_{n=0}^{+\infty} \frac{1}{2^{2n}} \binom{2n}{n} x^{2n} \sin^{2n} t.$$

Si $x \in J$ est fixé, la série de fonctions de t dans le membre de droite converge normalement sur $[0, \pi/2]$, ce qui permet d'intervertir les signes $\sum$ et $\int$. D'où, par définition de I_{2n} :

$$g(x) \;=\; \frac{2}{\pi} \sum_{n=0}^{+\infty} \frac{1}{2^{2n}} \binom{2n}{n} x^{2n} I_{2n} \;=\; \frac{2}{\pi} \sum_{n=0}^{+\infty} \frac{1}{2^{2n}} \binom{2n}{n} x^{2n} \left(\frac{\pi \binom{2n}{n}}{2^{2n+1}}\right) \;=\; f(x).$$

La première égalité vient de l'interversion des signes $\sum$ et $\int$, la deuxième de la valeur de I_{2n} obtenue dans la question 1, et la troisième de l'égalité $(**)$.

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.2.1 Pour tout $(n,x) \in \mathbb{N}^* \times \mathbb{R}$, on pose $f_n(x) := \arctan(x/n)$. Montrer que, sur $\mathbb{R}$, la suite de fonctions (f_n) converge vers 0 simplement mais pas uniformément, bien que la suite (f_n') converge uniformément vers 0. Commentaire ?

II.2.2 Pour tout $(n,x) \in \mathbb{N}^* \times \mathbb{R}_+$, on pose

$$f_n(x) := (1-x/n)^n \quad \text{si} \quad x \leqslant n \qquad \text{et} \qquad f_n(x) := 0 \quad \text{si} \quad x > n.$$

Montrer que la suite (f_n) converge uniformément sur $\mathbb{R}_+$, en précisant sa limite.

II.2.3 Soit a un réel fixé, $a > 0$. Si $(n,x) \in \mathbb{N} \times [0,1]$, on pose $f_n(x) := x^{n^a} - x^{(n+1)^a}$. Montrer que la suite (f_n) converge vers 0 uniformément sur $[0,1]$.

II.2.4 On pose $f_n(x) := \left(\displaystyle\int_0^x (1-t^2)^n \,\mathrm{d}t\right)\left(\displaystyle\int_0^1 (1-t^2)^n \,\mathrm{d}t\right)^{-1}$ pour tous $n \in \mathbb{N}$ et $x \in [0,1]$. Étudier la convergence simple et la convergence uniforme de la suite (f_n).

II.2.5 Soient k un réel positif fixé. Si une suite (f_n) de fonctions k-lipschitziennes de $[a,b]$ dans $\mathbb{R}$ $(a<b)$, converge simplement, montrer qu'elle converge uniformément.

II.2.6 Soit I un intervalle ouvert de $\mathbb{R}$. Si une suite $(f_n : I \to \mathbb{R})$ de fonctions *convexes* converge simplement, montrer, à l'aide de l'exercice précédent, qu'elle converge uniformément sur tout segment contenu dans I.

II.2.7 Soit $f : \mathbb{R}_+ \to \mathbb{R}$ une fonction continue, ayant en $+\infty$ une limite finie ℓ.
Montrer que f est limite uniforme (sur $\mathbb{R}_+$) d'une suite de combinaisons linéaires de fonctions du type $x \mapsto e^{-kx}$, $k \in \mathbb{N}$, en considérant la fonction $x \mapsto f(-\ln x)$ sur $]0,1]$.

II.2.8 $*$ Soient $a \in]0,1[$ et $I := [-a,a]$.

1. Soient $f : I \to \mathbb{R}$ une fonction continue telle que $f(0) := 0$ et $\varepsilon > 0$. Montrer qu'il existe un polynôme P à coefficients dans $\mathbb{Z}$ tel que $|f(x) - P(x)| \leqslant \varepsilon$ pour tout $x \in I$. On appliquera le théorème de Weierstraß à la fonction $t \mapsto f(\sqrt[k]{t})$ définie sur $[-a^k, a^k]$, k étant un entier impair assez grand.

2. Dans $\big(\mathcal{C}(I,\mathbb{R}), \|\cdot\|_\infty\big)$, quelle est l'adhérence de l'ensemble des fonctions polynômes à coefficients dans $\mathbb{Z}$?

II.2.9 $*$ Soit J un segment contenu dans $]0,1[$. Montrer que l'ensemble $\mathcal{A}$ des fonctions polynômes à coefficients dans $\mathbb{Z}$ est dense dans $\big(\mathcal{C}(J,\mathbb{R}), \|\cdot\|_\infty\big)$. Pour cela, étant donné un nombre premier p, on montrera que $Q_p := \big[1 - x^p - (1-x)^p\big]/p$ est à coefficients entiers, et on majorera $\left\|\dfrac{1}{p} - Q_p\right\|_\infty$, en notant (par abus) $Q_p : J \to \mathbb{R}$ la fonction induite par Q_p.

II.2.10 $*$ Soit E un sous-espace vectoriel de $\mathcal{C}\big([a,b],\mathbb{R}\big)$, de dimension finie n $(a<b)$.

1. Montrer qu'il existe des points distincts $x_1,\dots,x_n \in [a,b]$ tels que les formes linéaires $f\mapsto f(x_k)$ $(k=1,\dots,n)$ forment une base de E^*, dual de E.

2. En déduire que, si une suite (f_p) de E converge simplement vers une fonction f, la convergence est uniforme et $f\in E$.

II.2.11 $**$ Montrer qu'une fonction $f:[a,b]\to\mathbb{R}$ est limite uniforme de fonctions en escalier si, et seulement si, elle possède en tout point une limite à gauche et une limite à droite finies. Pour la définition d'une fonction en escalier, *cf.* page 550.

Exercices sur la section 2

II.2.12 Montrer que la formule $f(x):=\displaystyle\sum_{n=1}^{+\infty}(-1)^n\ln\left(1+\frac{x^2}{n(1+x^2)}\right)$ définit sur $\mathbb{R}$ une fonction de classe $\mathcal{C}^1$.

II.2.13 Montrer que la fonction f définie dans l'exercice 6 de la page 498 est de classe $\mathcal{C}^\infty$ sur $]0,+\infty[$.

II.2.14 Par interversion des signes $\sum$ et $\displaystyle\int$, prouver la formule suivante :

$$\sum_{n=1}^{+\infty}\frac{1}{n^n}\;=\;\int_0^1\frac{\mathrm{d}x}{x^x}.$$

II.2.15 Soit $n\in\mathbb{Z}$. Calculer $\displaystyle\int_0^{2\pi}e^{\cos t}\cos(\sin t-nt)\,\mathrm{d}t$ par la même méthode.

II.2.16 Montrer que la série $\displaystyle\sum_{k\geqslant 2}\big(\zeta(k)-1\big)$ converge et calculer sa somme. Montrer que la série $\displaystyle\sum_{k\geqslant 2}\big(\zeta(k)-1\big)/k$ converge et que sa somme est $1-C$, où C est la constante d'Euler. On rappelle que C est la limite, quand n tend vers $+\infty$, de $1+1/2+\cdots+1/n-\ln n$.

II.2.17 Montrer que $\Gamma'(1)=-C$, en désignant par C la constante d'Euler.

II.2.18 Montrer que la série de fonctions $x\mapsto\displaystyle\sum_{n\geqslant 0}\frac{(-1)^n}{n+x}$ définit une fonction de classe $\mathcal{C}^\infty$ sur $\mathbb{R}\setminus\{0,-1,-2,\dots\}$.

II.2.19 $*$ En quels points la fonction $x\mapsto\displaystyle\sum_{n=1}^{+\infty}2^{-n}E(nx)$ est-elle continue ? On note ici $E(t)$ la partie entière d'un réel t.

II.2.20 $*$ Soit $t \in \mathbb{R}^*$. Montrer que la série $\sum_{n\geqslant 1} n^{-(1+it)}$ diverge.

On comparera $\sum_{n\geqslant 1} \cos(t\ln n)/n$ et $\displaystyle\int_1^{+\infty} \cos(t\ln x)\,\mathrm{d}x/x$.

II.2.21 $*$ Soit $(a_n)_{n\geqslant 1}$ une suite de $\mathbb{C}$.
On considère sur $\mathbb{C}$ la série de fonctions $z \mapsto \sum_{n\geqslant 1} a_n n^{-z}$ (une telle série est appelée *série de Dirichlet*). On suppose que la série converge en un point $z_0 \in \mathbb{C}$. Montrer que, pour tout $\alpha \in]0,\pi/2[$ fixé, cette série converge uniformément sur $S := \{z_0 + re^{i\theta} \mid r > 0,\ -\alpha \leqslant \theta \leqslant \alpha\}$. On se ramènera au cas où $z_0 := 0$ et l'on utilisera une transformation d'Abel appropriée.
Application : soient $u \in]0,1[$, $v \in \mathbb{R}$ et $s := u + iv$. Montrer que la série $\sum_{n\geqslant 1} n^{-s}$ diverge.

II.2.22 $*$

1. Montrer que la formule :

$$f(x) \;:=\; \sum_{n=0}^{+\infty} \frac{(-1)^n}{n!(x+n)}\cdot$$

définit une fonction f de classe $\mathcal{C}^\infty$ sur $\mathcal{D} := \mathbb{R} \setminus \{0,-1,-2,\dots,\}$.

2. Soit $x \in \mathcal{D}$. À l'aide d'une décomposition en éléments simples, établir la formule :

$$ef(x) \;=\; \sum_{n=0}^{+\infty} \frac{1}{x(x+1)\cdots(x+n)}\cdot$$

Calculer $xf(x) - f(x+1)$.

3. Pour tout $x > 0$, montrer que $f(x) = \displaystyle\int_0^1 t^{x-1}e^{-t}\,\mathrm{d}t$.

II.2.23 $*$

1. Montrer que la formule ci-dessous définit une fonction 1-périodique continue sur $\mathbb{R} \setminus \mathbb{Z}$:

$$f(x) \;:=\; \sum_{n\in\mathbb{Z}} \frac{1}{(n-x)^2} \;:=\; \frac{1}{x^2} + \sum_{n=1}^{+\infty} \left[\frac{1}{(n-x)^2} + \frac{1}{(n+x)^2}\right].$$

2. Pour tout $x \in \mathbb{R} \setminus \mathbb{Z}$, on pose $g(x) := f(x) - \pi^2/\sin^2 \pi x$. Montrer que g se prolonge en une fonction continue, encore notée g, de $\mathbb{R}$ dans $\mathbb{R}$.

3. Pour tout $x \in \mathbb{R}$, montrer que $g(x/2) + g\big((x+1)/2\big) = 4g(x)$. En déduire les formules

$$\frac{\pi^2}{\sin^2 \pi x} \;=\; \sum_{n\in\mathbb{Z}} \frac{1}{(n-x)^2} \quad \text{pour tout} \quad x \in \mathbb{R} \setminus \mathbb{Z} \quad \text{et} \quad \sum_{n=1}^{+\infty} \frac{1}{n^2} \;=\; \frac{\pi^2}{6}\cdot$$

Exercices sur la section 3

II.2.24 Soit a un réel strictement positif. Déterminer la nature de la série suivante :

$$\sum_{n\geqslant 1} \frac{\big[(n+1)a\big]^n}{n^{n+1}},$$

en appliquant la règle de Cauchy, puis celle de d'Alembert.

II.2.25 Soit $\sum\limits_{n\geqslant 0} a_n$ une série à termes réels strictement positifs.

1. Supposons que la suite $(a_{n+1}/a_n)_n$ ait une limite ℓ, finie ou égale à $+\infty$. Montrer que ℓ est aussi limite de la suite $(\sqrt[n]{a_n})_n$. En déduire que, si la règle de d'Alembert permet de déterminer la nature de $\sum\limits_{n\geqslant 0} a_n$, c'est aussi le cas de la règle de Cauchy.

2. Examiner la réciproque, en prenant $a_n := 2^{E(n/2)} \cdot 3^{-n}$, en désignant par $E(t)$ la partie entière d'un réel t.

II.2.26 Convergence et calcul de $\displaystyle\sum_{n\geqslant 0}^{+\infty} \frac{x^n \cos na}{n!}$, lorsque $a, x \in \mathbb{C}$ sont donnés.

II.2.27 Développer la fonction $x \mapsto \sqrt{\dfrac{1+x}{1-x}}$ en série entière sur $]-1,1[$.

II.2.28 Soit $j := \exp(2i\pi/3)$. Montrer que la série de fonctions $x \mapsto \sum\limits_{n\geqslant 1} j^n x^n/n$ converge uniformément sur $[0,1]$. Soit $f : [0,1] \to \mathbb{C}$ sa somme. Calculer $f(x)$ lorsque $x \in [0,1[$. Montrer que la série $\sum\limits_{n\geqslant 1} j^n/n$ converge et calculer sa somme.

II.2.29 Développer la fonction $x \mapsto \arctan(x+1)$ en série entière au voisinage de 0 dans $\mathbb{R}$, en précisant le domaine de validité du développement obtenu.

II.2.30

1. Pour tous $n, p \in \mathbb{N}$, montrer que $\displaystyle\int_0^1 t^n (\ln t)^p \,\mathrm{d}t = \frac{(-1)^p\, p!}{(n+1)^{p+1}}$.

2. Développer $x \mapsto \displaystyle\int_0^1 t^{xt} \,\mathrm{d}t$ en série entière sur $\mathbb{R}$.

II.2.31 Soit $a \in \mathbb{C}$ tel que $|a| \neq 1$. À l'aide d'un développement en série entière, déterminer la série de Fourier de la fonction $x \mapsto 1/(e^{ix} - a)$.

II.2.32 Soit $\sum_{n\geqslant 0} a_n z^n$ une série entière de rayon de convergence $+\infty$ dont la somme $f : \mathbb{C} \to \mathbb{C}$ est « à croissance polynomiale » : il existe $K, a > 0$ tels que $|f(z)| \leqslant K\,|z|^a$ pour tout $z \in \mathbb{C}$.
Soit $r > 0$ fixé. À l'aide de la fonction $\theta \mapsto f(re^{i\theta})$ de $\mathbb{R}$ dans $\mathbb{C}$, montrer que $|a_p| \leqslant Kr^{a-p}$ pour tout $p \in \mathbb{N}$. En déduire que f est un polynôme.

II.2.33 Soient $n \in \mathbb{N}^*$ et a l'entier le plus proche de $n\sqrt{3}$.
Établir la minoration $\left|a - n\sqrt{3}\right| \geqslant 1/(2n\sqrt{3}+1)$.
En déduire le rayon de convergence de la série entière $\sum_{n\geqslant 1} z^n/\sin(\pi n\sqrt{3})$.

II.2.34 Soit $(p_n)_{n\geqslant 0}$ la suite croissante des nombres premiers. Quel est le rayon de convergence de la série entière $\sum_{n\geqslant 0} z^{p_n}/p_n$?

II.2.35 $*$ Pour quels $z \in \mathbb{C}$ l'expression $f(z) := \sum_{n\geqslant 1} (-1)^n/(n+z)$ a-t-elle un sens ?
Développer f en série entière au voisinage de 0, en précisant le rayon de convergence.

II.2.36 Soit $a \in \mathbb{C}^*$. À l'aide d'une équation différentielle linéaire d'ordre deux, développer la fonction $x \mapsto \sin(a \arcsin x)$ en série entière sur $]-1,1[$. Procéder de même avec la fonction $x \mapsto (\arcsin x)^2$.

II.2.37 Pour tout $n \in \mathbb{N}$, on pose $a_n := \displaystyle\int_0^1 \left(\frac{1+t^2}{2}\right)^n \mathrm{d}\,t$.
Montrer que la série entière $\sum_{n\geqslant 0} a_n z^n$ a pour rayon de convergence 1, et calculer sa somme sur $]-1,1[$.

II.2.38 $*$ Démontrer la deuxième assertion de l'exercice 15 de la page 526 directement, par une majoration convenable des modules des coefficients de G.

II.2.39 $*$ Pour tout $n \in \mathbb{N}$, soit a_n le nombre de façons de payer n centimes avec des pièces de $1, 2$ et 5 centimes.
Montrer que la série entière $\sum_{n\geqslant 0} a_n z^n$ est le produit des séries entières $\sum_{n\geqslant 0} z^n$, $\sum_{n\geqslant 0} z^{2n}$ et $\sum_{n\geqslant 0} z^{5n}$.
En déduire une formule donnant a_n en fonction de n, et ne comportant plus de somme.

II.2.40 $*$ Démontrer la formule suivante :

$$\pi \;=\; \sum_{n=0}^{+\infty} \frac{1}{16^n}\left[\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6}\right].$$

II.2.41 $*$

1. Pour tout $n \in \mathbb{N}$, démontrer la formule suivante :

$$\int_0^{\pi/2} \cos^{2n+1} x \, \mathrm{d}x \;=\; \frac{2^{2n}}{(2n+1)\binom{2n}{n}}.$$

2. En déduire (cela exige pas mal de calculs) l'égalité suivante :

$$\sum_{n=1}^{+\infty} \frac{n2^n}{\binom{2n}{n}} \;=\; \pi + 3.$$

II.2.42 Soient $A := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ et $B := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Montrer que $e^{A+B} \neq e^A e^B$.

II.2.43 Soit $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{R})$. Calculer explicitement $\exp(A)$ en fonction de a, b, c, d, d'abord lorsque $\mathrm{Tr}(A) = 0$, puis dans le cas général.

Intégration II.3

Si l'on « surfe naïvement sur le net » à partir des mots « intégration » ou « intégrale », on trouve (en se limitant au domaine mathématique…) une foule de notions : intégrale de Cauchy, intégrale de Newton, intégrale de Riemann, intégrale de Riemann généralisée, intégrale de Riemann complète, intégrale de Darboux, intégrale de Stieltjes, intégrale de Lebesgue, intégrale de Henstock-Kurzweil, intégrale de Denjoy, intégrale de Perron, intégrale de Daniell, intégrale de Radon, intégrale abstraite, intégrale d'Ito, intégrale de Feynmann (ou intégrale de chemin) et d'autres…

Pourquoi tant de notions ? Quelle(s) théorie(s) doit-on apprendre durant les premières années de l'enseignement universitaire ?

On trouvera ci-dessous dans l'encadré *Un peu d'Histoire* quelques éléments de réponse à la première question.

Pour essayer de répondre au mieux à la deuxième question, nous avons, par comparaison avec la première édition de nos livres, adopté une présentation *beaucoup plus progressive* de la théorie de l'intégration étalée sur les trois ouvrages de cette nouvelle édition. Nous avons ainsi tenu compte, d'une part, de la mise en place des nouveaux programmes de l'enseignement secondaire et des modifications corrélatives des enseignements universitaires et, d'autre part, des remarques de nos lecteurs et de nos collègues enseignants. Nous précisons nos choix à la fin de cette présentation.

Dans ce module nous présentons l'intégrale de Riemann pour les fonctions *d'une variable*.

Nous commençons, dans la section 1, par le cas des fonctions à valeurs réelles définies sur un segment. Nous définissons les fonctions intégrables par encadrement par des fonctions en escalier. Nous étudions ensuite les sommes de Darboux et les sommes de Riemann.

Nous poursuivons par une introduction à l'intégration des fonctions à valeurs vectorielles, d'abord pour le cas de la dimension finie, puis pour celui d'un espace de Banach (sections 3 et 4).

Nous verrons que la théorie de l'intégration au sens de Riemann permet d'intégrer de larges classes de fonctions : continues par morceaux, monotones, à variation bornée, réglées. L'étude des applications réglées fait l'objet de la section 5.

Un peu d'Histoire

L'histoire des théories de l'intégration est longue et compliquée. Nous donnerons seulement quelques jalons importants. Nous compléterons la partie consacrée aux développements récents dans le module *Intégration* de [L2-L3]. Cet historique fait inévitablement appel à quelques termes techniques, nous suggérons donc au lecteur d'y revenir au fur et à mesure de l'étude du module.

Antiquité et mathématiques arabes : la préhistoire du calcul intégral

Les premiers calculs d'aires et de volumes sont très anciens. Au IV^{e} siècle avant Jésus-Christ, le mathématicien grec Eudoxe de Cnide a inventé une méthode lui permettant de calculer le volume de pyramides, cônes et cylindres. Cette méthode dite *apagogique* (appelée aussi plus tard *méthode des anciens* ou *méthode d'exhaustion*) est assez proche de la notion moderne de limite. Elle repose sur un double raisonnement par l'absurde : pour montrer que l'aire A d'un domaine plan est égale à B, on montre que les hypothèses $A < B$ et $A > B$ conduisent à des contradictions (*cf.* la remarque de la page 556). Cette méthode a été reprise et développée par Archimède (-287, -212), par exemple pour le calcul des aires de « segments de parabole » ou celui du volume d'une sphère (ou, plus généralement, d'un ellipsoïde de révolution). Mais Archimède a confié à Erathostène de Cyrène (dans une lettre « testament », *Sur la Méthode*) qu'avant de *prouver* un résultat par la méthode d'exhaustion il le *devinait* (il écrit : *« … donne une idée que la conclusion est vraie »*) en utilisant une méthode heuristique, inspirée de l'atomisme, consistant, par exemple, à identifier une surface à une « somme » de segments parallèles la composant. Au cours des temps, le retentissement de cette idée heuristique va être considérable. Reprise quelques siècles plus tard par Cavalieri, avec ses « indivisibles », puis par Pascal, elle va conduire au calcul intégral de Leibniz et au développement des méthodes axées sur les *sommes de Riemann* (selon la terminologie moderne) qui se prolongent jusqu'à nos jours.

Les écrits d'Archimède ont été conservés dans le monde arabe et, à partir du milieu du IX^{e} siècle, divers mathématiciens (en particulier la famille Ibn Qurra, sur trois générations, et Ibn al Haytham, dit Alhazen) ont repris et prolongé ses idées pour calculer des aires et des volumes. Dans ces travaux, des *sommes de Riemann* sont utilisées de façon significative.

Cavalieri et Pascal : les indivisibles

En Europe, on trouve peu de contributions significatives avant les débuts du XVII^{e} siècle. Après divers travaux de quadrature du Toulousain Pierre de Fermat et de l'Anversois Grégoire de Saint-Vincent, l'Italien Bonaventura Cavalieri (1598-1647) introduit en 1635 la méthode des indivisibles (*Geometria indivilibus continuorum…*). Il part de l'idée qu'une surface est « constituée » de lignes « parallèles », les indivisibles, et « montre » l'égalité de deux aires en prouvant que les surfaces correspondantes sont constituées d'indivisibles deux à deux de même longueur. (Il a un principe analogue pour les volumes.) Il reprend ainsi avec beaucoup de hardiesse (peut-être due aux changements de perspective sur *l'infini* offerts par la pensée médiévale européenne et les idées de la Renaissance) les arguments d'Archimède dans sa lettre à Erathostène.

Un peu d'Histoire

La méthode de Cavalieri est très efficace mais on dirait aujourd'hui qu'elle est peu rigoureuse. Elle a été vigoureusement contestée à l'époque (l'Église l'a condamnée en 1649 !). Il semble que le savant français Gilles Personne de Roberval (inventeur d'une célèbre balance) ait découvert la méthode des indivisibles indépendamment de Cavalieri, mais il l'avait d'abord tenue secrète...

Vers le milieu du XVII^{e} siècle, les idées vont évoluer significativement grâce au philosophe et savant français Blaise Pascal (1623-1662). Il surmonte hardiment les aspects « paradoxaux » de la méthode des indivisibles en la traduisant « en termes apagogiques », réconciliant ainsi les deux approches utilisées par Archimède. Pour lui, on peut « remplacer » les indivisibles par des « sommes de Riemann ». Voici ce qu'il écrivait à propos de la quadrature du demi-disque :

« On n'entend autre chose par somme des ordonnées d'un cercle sinon la somme d'un nombre indéfini de rectangles faits de chaque ordonnée avec chacune des petites portions égales du diamètre dont la somme ne diffère de l'espace d'un demi-cercle que d'une quantité moindre qu'aucune donnée »

Nous sommes très près des définitions modernes données plus loin !

Intégration et calcul infinitésimal : Leibniz, Newton

L'invention du *calcul infinitésimal* au XVIII^{e} siècle, essentiellement due à Isaac Newton (1643-1727) et Gottfried Wilhelm Leibniz (1646-1716) change complètement les choses avec une *révolution conceptuelle* et de considérables retombées applicatives. Il faut noter qu'aussi bien Newton que Leibniz se sont inspirés des travaux de Pascal.

Newton introduit la notion de *fonction dérivée* d'une fonction (passage de la « fluente »à la « fluxion ») et, pour lui, l'opération d'intégration n'est pas autre chose que *l'inverse de celle de dérivation* (on dit aujourd'hui *primitive*).

L'inspiration de Newton est physique et cinématique, celle de Leibniz est plus philosophique et abstraite, avec une réflexion sur les relations entre discret et continu dans le prolongement des idées de Cavalieri et Pascal. Pour lui, comme pour Newton : *« L'intégration en tant que processus de sommation est l'inverse de la différentiation »*. Aujourd'hui, ces idées sont présentées sous le nom de *Théorème fondamental* dont nous donnerons plus loin diverses versions.

En théorie de l'intégration, les notations sont importantes. Le signe d'intégration a été introduit par Leibniz ainsi que le $\mathrm{d}x$ (écho de l'atomisme grec et des indivisibles). Le nom *intégrale* est dû à Jean Bernoulli (1695). La notation $\int_a^b$ est plus tardive. Elle est due à Joseph Fourier.

Un peu d'Histoire

Cauchy, Riemann, Darboux : les sommes de Riemann

Vers 1830, l'approche de l'intégration va évoluer pour diverses raisons. Parmi celles-ci, on peut noter d'une part l'émergence d'un souci de rigueur en analyse et d'autre part l'étude des séries de Fourier, introduites par celui-ci en 1822 (les coefficients de Fourier sont calculés par intégration).

En 1823, Cauchy définit l'intégrale d'une fonction *continue* sur un segment à l'aide de sommes de Riemann particulières, dites sommes de Cauchy (*cf.* la page 580). C'est un progrès important vers la rigueur, mais il confond la notion de fonction continue et celle de fonction *uniformément* continue...

En 1854, motivé par l'étude des séries de Fourier, le mathématicien allemand Bernhard Riemann (1826-1866) se pose le problème de définir l'intégrale de certaines fonctions discontinues, éventuellement très irrégulières. Il fonde pour cela la théorie de l'intégration qui porte son nom à partir des « sommes de Riemann » (dont nous avons signalé des avatars chez les mathématiciens arabes et plus tard chez Pascal puis ci-dessus chez Cauchy).

Vers 1875, le mathématicien français Gaston Darboux complète le travail de Riemann. C'est essentiellement sa théorie de l'intégration que nous présentons au début de ce module.

Lebesgue : intégration et mesure

Vers 1900, apparaissent des approches complètement nouvelles avec les travaux d'Émile Borel sur la mesure (1898) et d'Henri Lebesgue sur l'intégration. L'article fondateur des travaux de Lebesgue est une note aux *Comptes rendus de l'Académie des Sciences* (29 avril 1901), de moins de trois pages, dans laquelle il formule plusieurs idées « révolutionnaires ».

La nouvelle définition de l'intégrale proposée par Lebesgue repose sur des notions *d'ensembles mesurables* et de *mesure des ensembles*, qu'il esquisse aussi dans sa note (en quelques lignes !). Ces notions de mesure étant acquises, sa définition de l'intégrale est *très simple*. Pour définir l'intégrale d'une fonction f *bornée*, au lieu, comme ses prédécesseurs, de découper l'intervalle $[a,b]$ de l'axe des *abscisses*, il découpe un intervalle $[m,M]$ (contenant l'image de f) de l'axe des *ordonnées*. Il en déduit par image réciproque ($I \mapsto f^{-1}(I)$) une *partition finie* de $[a,b]$ en sous-ensembles sur lesquels f est « presque constante ». Il peut alors définir $\int_a^b f$ (par passage à la limite) pourvu que la mesure de ces sous-ensembles soit connue. Tout va bien si f vérifie une hypothèse ad-hoc, *la mesurabilité* : *« ... les fonctions telles que les ensembles qui figurent dans la définition de l'intégrale soient mesurables »*. Ainsi Lebesgue peut intégrer les fonctions qu'il appelle *sommables* (c'est-à-dire *bornées* et *mesurables*). Il montre qu'il obtient ainsi une généralisation *stricte* de l'intégrale de Riemann.

La motivation initiale de Lebesgue était la comparaison entre *primitive* (Newton et Leibniz) et *intégrale* (Cauchy et Riemann). Constatant que : *« ... toutes les dérivées ne sont pas intégrables au sens de Riemann »*, il se propose d'y remédier. Et il y remédie effectivement pourvu que la dérivée soit *bornée*. Pour lever cette dernière restriction il faudra attendre les intégrales de Denjoy, Perron, Henstock, Kurzweil...

Un peu d'Histoire

Au-delà de la motivation initiale, le véritable intérêt de l'intégrale de Lebesgue n'est apparu que plus tard avec, par exemple, les applications à l'analyse fonctionnelle (l'espace des « fonctions sommables » est *complet* : Riesz et Nagy, 1907). La notion de mesure a connu après Lebesgue un développement considérable (mesure de Radon en 1913, mesure abstraite de Carathéodory en 1918, théorie des probabilités après l'axiomatisation due à Kolmogoroff en 1933, mesure de Wiener...).

Il faut noter toutefois qu'une fonction *semi-intégrable* (au sens de Riemann généralisé) n'est pas intégrable au sens de Lebesgue. C'est une limitation trop oubliée de l'intégration au sens de Lebesgue. Les fonctions semi-intégrables, souvent présentées comme une « fantaisie de mathématicien » ou un prétexte à exercices, sont en fait essentielles dans diverses théories physiques.

Développements récents

Après Lebesgue, divers mathématiciens sont revenus sur le problème de la comparaison entre primitive et intégrale, qu'il n'avait que partiellement résolu. Ceci a motivé des travaux d'Arnaud Denjoy (1912) et Oskar Perron (1914) qui ont introduit par des approches très différentes des notions équivalentes d'intégrabilité permettant enfin de résoudre le problème de comparaison (le théorème fondamental de l'analyse). Mais les théories de l'intégration de Denjoy et Perron sont compliquées... Plus tard, indépendamment, Kurzweil (1957) et Henstock (1961) ont redécouvert l'intégrale de Denjoy-Perron. Leur approche est différente de celles de leurs deux prédécesseurs, elle est basée sur la notion de *jauge*. L'idée de départ est ancienne. Thabit ibn Qurra (pour la quadrature de la parabole) et plus tard Euler et Fermat avaient remarqué qu'il est utile, pour calculer des valeurs approchées de certaines intégrales, d'utiliser des subdivisions *en parties inégales*. Ainsi Euler écrivait : *« ... là où des changements peu importants de x conduisent à des variations violentes de $f(x)$, on devra prendre l'intervalle très petit »*. L'idée de Henstock et Kurzweil est de formaliser ces remarques en modifiant la définition de Riemann : au lieu de contrôler le pas des sommes de Riemann par un *nombre* $\delta > 0$, on le contrôle par une *fonction* strictement positive δ, la jauge, qui peut s'adapter aux « variations violentes » de la fonction à intégrer. L'intégrale de Henstock-Kurzweil est aussi appelée *intégrale de jauge* ou intégrale de Riemann complète. Elle généralise à la fois l'intégrale de Lebesgue et l'intégrale de Riemann sous sa forme généralisée.

Dans la section 6 nous étudierons l'intégrale fonction de sa borne supérieure et le *théorème fondamental de l'analyse* dans le cadre de l'intégrale de Riemann. Nous reviendrons sur le sujet dans le module *Integration* de [L2-L3] dans le cadre de l'intégrale de Henstock-Kurzweil où les difficultés disparaissent : la dérivée d'une fonction dérivable est toujours intégrable au sens de Henstock-Kurzweil.
La section 7 est consacrée à l'extension de la notion d'intégrale de Riemann à un intervalle réel quelconque. On parle alors d'intégrale impropre ou généralisée. C'est important pour un certain nombre d'applications, en particulier en physique. Il faut noter que, bien que la théorie de Lebesgue soit « plus forte » que celle de Riemann dans le cas de l'intégration sur un segment, il n'en est plus de même dans le cas de l'intégrale généralisée : par exemple la fonction $f : x \in \mathbb{R} \mapsto \frac{\sin x}{x} \in \mathbb{R}$ est intégrable au sens de Riemann généralisé (et au sens de Henstock-Kurzweil) mais pas au sens de Lebesgue.
Dans la dernière partie (section 8) nous étudions l'intégration des *limites simples* de fonctions intégrables au sens de Riemann (avec des applications à l'intégrabilité et la dérivabilité des fonctions définies par une intégrale). C'est une question difficile mais importante pour les applications. On reproche traditionnellement (à juste titre…) à l'intégrale de Riemann de « mal passer à la limite simple » pour justifier le recours à d'autres théories de l'intégration. Nous proposons une solution intermédiaire, plus accessible au débutant et bien suffisante pour résoudre un très grand nombre de problèmes avec, outre les résultats classiques dans le cas de la convergence uniforme, des versions « adaptées » des théorèmes de convergence monotone, encadrée et dominée des théories de Lebesgue et Henstock-Kurzweil. Ces dernières, dont la preuve est délicate, seront *provisoirement admises* ici et prouvées dans le module *intégration* de [L2-L3] à l'aide de la théorie de Henstock-Kurzweil.

Voici, pour terminer cette présentation, quelques précisions sur nos choix.
Durant les dernières décennies l'intégrale de Riemann a souvent été considérée comme « passée de mode » car « dépassée » par celle de Lebesgue. Nous tenons ce point de vue pour erroné, l'intégrale de Riemann a énormément d'applications (dont certaines « de niveau élevé » bien au-delà de celui du présent ouvrage…) tout en restant *d'exposition très simple*. Son apprentissage est donc tout à fait abordable et fort utile pour ceux qui en resteront là. C'est par ailleurs une bonne façon de se préparer à aborder les « grandes théories de l'intégration et de la mesure » pour ceux qui voudront aller plus loin. L'intégrale de Riemann est aussi, pour ceux qui poursuivront dans d'autres directions, une bonne introduction à divers sujets comme l'intégrale de Stieltjes, les intégrales stochastiques et l'intégrale de Feymann de la théorie des champs.
Cela étant, il n'est pas inutile de prendre conscience de certains « défauts » de l'intégrale de Riemann.

1. Certaines fonctions « utiles » ne sont pas Riemann-intégrables. Nous en donnerons quelques exemples *simples*.
2. Comme nous l'avons déjà signalé, l'intégrale de Riemann « passe mal » à la limite simple des suites de fonctions. Nous expliquons comment y remédier dans la section 8.
3. L'intégrale de Riemann s'étend à plusieurs variables, nous en ferons une présentation

dans le module *Intégration* de [L2-L3]. Mais cette extension n'est plus vraiment élémentaire et il n'y a pas d'énoncé très satisfaisant du théorème de Fubini permettant les calculs par intégrales itérées. Cela étant, quel que soit le cadre théorique adopté, une justification rigoureuse des calculs les plus simples est vite laborieuse, ce qui conduit en pratique plusieurs auteurs et utilisateurs à se passer de justification...

Pour aller plus loin et pallier ces défauts (et d'autres...), nous présenterons dans le module *Intégration* de [L2-L3] la théorie de l'intégrale de Henstock-Kurzweil. Celle-ci a de très nombreux avantages. C'est une *modification simple* de celle de Riemann vue du point de vue des *sommes de Riemann* mais elle englobe celle de Riemann, y compris sa version généralisée à un intervalle quelconque, et celle de Lebesgue, avec des théorèmes « plus performants » assez faciles à prouver. Nous verrons également comment déduire très simplement la théorie de Lebesgue de celle de Henstock-Kurzweil.

Notations

Dans tout ce module on désigne par $\mathbb{K}$ le corps $\mathbb{R}$ ou le corps $\mathbb{C}$ et par E un $\mathbb{K}$-espace vectoriel. On désigne par a et b deux réels tels que $a < b$ (sauf mention explicite du contraire, *cf.* ci-dessous les *Conventions* de la page 561).
On considérera des applications $f : [a,b] \to E$. Si $E := \mathbb{R}$ on parlera de *fonction* plutôt que d'application (cette convention n'étant valable que pour ce module).

1 Intégrale de Riemann d'une fonction réelle sur un segment

1.1 Subdivisions d'un segment

Définition 1.

1. On appelle *subdivision* d'un segment $[a,b]$ toute *famille* $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ de points de ce segment telle que $n \in \mathbb{N}^*$ et :
$$x_0 := a < x_1 < \cdots < x_n := b$$
2. On appelle *pas* ou *module* de la subdivison $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ le réel strictement positif :
$$\delta(\mathbf{u}) := \max_{i\in[\![0,n-1]\!]} (x_{i+1} - x_i).$$
3. On appelle *subdivision pointée* d'un segment $[a,b]$ la donnée d'une subdivision $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ et, pour tout $i \in [\![0, n-1]\!]$, d'un point $\xi_i \in [x_i, x_{i+1}]$. On note $\mathcal{D} := \Big\{\big([x_i, x_{i+1}], \xi_i\big)\Big\}_{i\in[\![0,n-1]\!]}$ une telle subdivision.

Si un segment pointé (J,ξ) est l'un des segments pointés $\big([x_i, x_{i+1}], \xi_i\big)$, on notera, par abus de langage $(J,\xi) \in \mathcal{D} = \Big\{\big([x_i, x_{i+1}], \xi_i\big)\Big\}_{i\in[\![0,n-1]\!]}$.
On appelle pas de la subdivision pointée $\mathcal{D}$ celui de la subdivision « sous-jacente » $\mathbf{u}$ et l'on note $\delta(\mathcal{D}) := \delta(\mathbf{u})$.

On appelle *image* de la subdivision $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ *l'ensemble* :

$$\operatorname{Im}\mathbf{u} := \{x_i\}_{i\in[\![0,n]\!]} \subset [a,b],$$

(une subdivision est par définition une application $\mathbf{u} : [\![0,n]\!] \to [a,b]$ et $\operatorname{Im}\mathbf{u}$ est l'image de cette application.)

Si $\mathbf{u}$ et $\mathbf{v}$ sont deux subdivisions de $[a,b]$ on dit que $\mathbf{u}$ est *plus fine* que $\mathbf{v}$ si $\operatorname{Im}\mathbf{v} \subset \operatorname{Im}\mathbf{u}$. On obtient ainsi une relation d'ordre sur l'ensemble des subdivisions.

Si $\mathbf{u}$ et $\mathbf{v}$ sont deux subdivisions de $[a,b]$, il existe une subdivision $\mathbf{w}$ plus fine que $\mathbf{u}$ et $\mathbf{v}$. Par exemple $\mathbf{w} := \mathbf{u}\cup\mathbf{v}$, où l'on a noté, par abus de langage, $\mathbf{u}\cup\mathbf{v}$ la subdivision obtenue en réunissant les éléments de $\mathbf{u}$ et $\mathbf{v}$, en les réordonnant et en supprimant les « doublés » ($\operatorname{Im}(\mathbf{u}\cup\mathbf{v}) = \operatorname{Im}\mathbf{u}\cup\operatorname{Im}\mathbf{v}$).

1.2 Applications en escalier à valeurs dans un espace vectoriel

Pour préparer les sections 3 et 4, nous étendons la définition des fonctions en escalier donnée dans le livre [L1] au cas des applications à valeurs dans un $\mathbb{K}$-espace vectoriel E. En première lecture on pourra se limiter au cas où $E = \mathbb{K}$.

Définition 2. Une application $f : [a,b] \to E$ est *en escalier* s'il existe une subdivision $(x_i)_{i\in[\![0,n]\!]}$ de $[a,b]$ telle que f soit constante sur chacun des intervalles ouverts $]x_i, x_{i+1}[$, $i \in [\![0,n-1]\!]$.
Une telle subdivision est dite *adaptée* à l'application en escalier f.

Si une subdivision est plus fine qu'une subdivision *adaptée* à f elle est aussi *adaptée* à f.

Si f est une fonction en escalier et si $\mathbf{u}$ est une subdivision adaptée à f, l'ensemble des *points de discontinuité* de f est contenu dans $\operatorname{Im}\mathbf{u}$, il est donc fini et il définit une subdivision $\mathbf{u}_f$ *moins fine* que $\mathbf{u}$. Cette subdivision est adaptée à f, c'est *le plus petit élément* de l'ensemble des subdivisions adaptées à f (muni de la relation d'ordre « plus fine »), on l'appelle la *subdivision minimale* adaptée à f.

On note $\mathcal{E}\big([a,b],E\big)$ l'ensemble des applications en escalier sur $[a,b]$ à valeurs dans E.

Si $E = \mathbb{R}$, on parle de *fonctions en escalier* et l'on note $\mathcal{E}\big([a,b],\mathbb{R}\big) = \mathcal{E}\big([a,b]\big)$.

Une fonction en escalier ne prend qu'un nombre fini de valeurs, elle est donc *bornée*.

La fonction caractéristique 1_A d'un sous-intervalle A de $[a,b]$ est en escalier.

Proposition 1.
Une application en escalier $f \in \mathcal{E}\big([a,b],E\big)$ peut s'écrire $f = \sum_{j\in[\![1,m]\!]} 1_{A_j}\,\lambda_j$, où les A_j sont des intervalles *disjoints* non vides de $[a,b]$ et $\lambda_j \in E$.

Démonstration. Soit $\mathbf{u}_f := (x_i)_{i\in[\![0,n]\!]}$ la subdivision minimale adaptée à f. Pour tout $i \in [\![0,n-1]\!]$ la fonction f est constante sur $]x_i, x_{i+1}[$, notons c_i sa valeur sur cet intervalle. On a :

$$f = \sum_{i\in[\![0,n]\!]} f(x_i)1_{\{x_i\}} + \sum_{i\in[\![0,n-1]\!]} c_i\,1_{]x_i,x_{i+1}[}.$$

Le résultat s'en déduit en ré-indexant la sommation. ∎

Remarque. L'écriture $f = \sum_{j\in[\![1,m]\!]} 1_{A_j}\,\lambda_h$, les A_j étant des intervalles disjoints, n'est pas en général unique ($1_{[0,1]} = 1_{\{0\}} + 1_{]0,1]}$), mais on peut vérifier qu'elle le devient si l'on impose que le nombre des A_j soit *minimal*.

Proposition 2.
(i) L'ensemble $\mathcal{E}\big([a,b],E\big)$ est un sous-espace vectoriel de $\mathcal{F}\big([a,b],E\big)$.
(ii) L'espace vectoriel $\mathcal{E}\big([a,b],\mathbb{K}\big)$ est engendré par les fonctions caractéristiques des sous-intervalles de $[a,b]$.

Démonstration.
(i) On reprend la preuve donnée dans le module *Introduction à l'intégration* de [L1] pour le cas $E := \mathbb{R}$.
(ii) D'après la proposition 1 ci-dessus, toute fonction en escalier est combinaison linéaire de fonctions caractéristiques d'intervalles. Inversement toute fonction caractéristique d'intervalle est en escalier. ∎

Proposition et définition 3. Soient $f \in \mathcal{E}\big([a,b],E\big)$ et $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ une subdivision adaptée à f. Si, pour tout $i \in [\![0,n-1]\!]$, on note $c_i \in E$ la valeur de f sur $]x_i, x_{i+1}[$, alors la quantité :

$$\sum_{i\in[\![0,n-1]\!]} (x_{i+1} - x_i)\,c_i$$

ne dépend pas de la subdivision adaptée $\mathbf{u}$ choisie.

On l'appelle *intégrale* de f sur $[a,b]$, on la note $\displaystyle\int_a^b f$. L'application $f \mapsto \displaystyle\int_a^b f$ est linéaire de $\mathcal{E}\big([a,b],E\big)$ dans E.

Démonstration. On reprend la preuve donnée dans le module *Introduction à l'intégration* de [L1] pour le cas $E := \mathbb{R}$. ∎

Exemple. Si $a \leqslant c \leqslant d \leqslant b$, $\displaystyle\int_a^b 1_{[c,d]} = \int_a^b 1_{]c,d[} = \int_a^b 1_{[c,d[} = \int_a^b 1_{]c,d]} = d - c.$

D'après la proposition 2 (ii), la forme linéaire $f \mapsto \displaystyle\int_a^b f$ est déterminée par sa valeur sur les fonctions caractéristiques d'intervalles que l'on calcule par la formule ci-dessus.
L'ensemble $\mathcal{F}\big([a,b]\big)$ est muni de la *relation d'ordre* : $f \leqslant g$ si, et seulement si, pour tout $x \in [a,b]$, $f(x) \leqslant g(x)$. On munit $\mathcal{E}\big([a,b]\big)$ de la relation d'ordre induite.
On vérifie facilement le résultat suivant.

Proposition 4. Soient $f, g \in \mathcal{E}\big([a,b]\big)$.

(i) Si $f \geqslant 0$, alors $\displaystyle\int_a^b f \geqslant 0$. La forme linéaire $f \mapsto \displaystyle\int_a^b f$ sur $\mathcal{E}\big([a,b]\big)$ est *positive*.

(ii) Si $f \leqslant g$, alors $\displaystyle\int_a^b f \leqslant \int_a^b g$.

Proposition 5. Pour tous $f, g \in \mathcal{E}\big([a,b]\big)$ les fonctions :

$$\sup(f,g), \quad \inf(f,g) \quad \text{et} \quad fg$$

sont des fonctions en escalier sur $[a,b]$.

Démonstration. Il existe une subdivision $(x_i)_{i\in[\![0,n]\!]}$ de $[a,b]$ adaptée à f et g. Sur tout intervalle ouvert $]x_i, x_{i+1}[$ $(i \in [\![0, n-1]\!])$ les fonctions $\sup(f,g)$, $\inf(f,g)$ et fg sont constantes. ∎

Si $f \in \mathcal{E}\big([a,b]\big)$ les fonctions :

$$f^+ := \sup(f,0), \quad f^- := \sup(-f,0) \quad \text{et} \quad |f| = \sup(f,-f) = f^+ + f^-$$

sont des fonctions en escalier sur $[a,b]$.

On a une *relation de Chasles* pour les fonctions en escalier (dont nous laissons la vérification au lecteur).

Proposition 6. Soit $f \in \mathcal{E}\big([a,b], E\big)$. Pour tout $c \in]a,b[$ les restrictions de f à $[a,c]$ et $[c,b]$ sont des fonctions en escalier (toujours notées f) et l'on a :

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

Les notions de subdivision, de subdivision pointée et de fonction en escalier s'étendent *mutatis mutandis* si l'on remplace le segment $[a,b]$ par un intervalle *borné* dont les bornes inférieure et supérieure sont respectivement a et b.

1.3 Fonctions intégrables, définition de l'intégrale et propriétés

Dans le module *Introduction à l'intégration* de [L1], nous avons défini la notion d'intégrale d'une fonction en escalier et nous en avons déduit la notion d'intégrale d'une fonction continue par morceaux. Nous allons reprendre la méthode utilisée en l'étendant à une classe *plus large* de fonctions.

Partant de l'ensemble des fonctions en escalier, l'idée est d'étendre cet ensemble de fonctions intégrables en appelant « intégrable » toute fonction qui peut être encadrée par des fonctions en escalier « aussi proches que l'on veut ». Deux fonctions en escalier f et g seront considérées comme « proches » si $\displaystyle\int_a^b |f-g|$ est « petit ». On prendra garde au fait que

l'application $(f,g) \mapsto \displaystyle\int_a^b |f-g|$ *n'est pas une distance* sur $\mathcal{E}\big([a,b]\big)$ car l'on peut avoir $\displaystyle\int_a^b |f-g| = 0$ et $f \neq g$. Une variante serait de considérer f et g comme « proches » si $\|g-f\|_\infty := \sup_{x\in[a,b]} \big|g(x)-f(x)\big|$ est « petit ». L'on obtiendrait alors un ensemble *strictement plus petit* de fonctions intégrables, l'ensemble des fonctions dites *réglées*. Nous reviendrons sur cette question dans la section 5.

1.3.1 Fonctions intégrables au sens de Riemann

Définition 3. On dit que $f : [a,b] \to \mathbb{R}$ est *intégrable au sens de Riemann* sur $[a,b]$ si, pour tout $\varepsilon > 0$, il existe des fonctions en escalier φ_ε et ψ_ε sur $[a,b]$ telles que :

$$\psi_\varepsilon \leqslant f \leqslant \varphi_\varepsilon \qquad \text{et} \qquad \int_a^b \varphi_\varepsilon - \int_a^b \psi_\varepsilon = \int_a^b (\varphi_\varepsilon - \psi_\varepsilon) \leqslant \varepsilon.$$

On écrira aussi que f est $\mathscr{R}$-intégrable. On notera $\mathscr{R}\big([a,b]\big)$ l'ensemble des fonctions intégrables au sens de Riemann sur $[a,b]$.
Il faut retenir qu'une fonction est $\mathscr{R}$-intégrable s'il est possible de l'encadrer par deux fonctions en escalier d'intégrales arbitrairement voisines.
Toute fonction en escalier étant bornée, il résulte immédiatement de la définition que toute fonction intégrable au sens de Riemann est *bornée*.

Proposition 7. Soit $f : [a,b] \to \mathbb{R}$ *intégrable au sens de Riemann* sur $[a,b]$. Alors pour tous $c,d \in [a,b]$, $c < d$, la restriction de f à $[c,d]$ est *intégrable au sens de Riemann* sur $[c,d]$.

Démonstration. Soient φ_ε et ψ_ε des fonctions en escalier sur $[a,b]$ telles que :

$$\psi_\varepsilon \leqslant f \leqslant \varphi_\varepsilon \qquad \text{et} \qquad \int_a^b \varphi_\varepsilon - \int_a^b \psi_\varepsilon = \int_a^b (\varphi_\varepsilon - \psi_\varepsilon) \leqslant \varepsilon.$$

Les restrictions de φ_ε et ψ_ε à $[c,d]$ sont des fonctions en escalier et, puisque $\varphi_\varepsilon - \psi_\varepsilon \geqslant 0$, on a, en utilisant la relation de Chasles pour les fonctions en escalier (*cf.* la proposition 6 de la page ci-contre) :

$$\int_c^d \varphi_\varepsilon - \int_c^d \psi_\varepsilon = \int_c^d (\varphi_\varepsilon - \psi_\varepsilon) \leqslant \int_a^b (\varphi_\varepsilon - \psi_\varepsilon) \leqslant \varepsilon. \qquad \blacksquare$$

Exercice 1.

Soit $f : [0,1] \to \mathbb{R}$ définie par $\begin{cases} f(x) = 1 & \text{si } x \in \mathbb{Q}, \\ f(x) = -1 & \text{si } x \notin \mathbb{Q}. \end{cases}$

Montrer que f n'est pas $\mathscr{R}$-integrable sur $[0,1]$.

Solution. Soient $\varphi \in \mathcal{E}_f^+$ et $\psi \in \mathcal{E}_f^-$. On a $\psi \leqslant f \leqslant \varphi$. Tout intervalle ouvert sur lequel φ est constante contient un point rationnel donc $1 \leqslant \varphi$ sur cet intervalle. Tout intervalle ouvert sur lequel ψ est constante contient un point irrationnel donc $\psi \leqslant -1$ sur cet intervalle. Il existe un segment $[a,b] \subset [0,1]$ de longueur non nulle sur lequel φ et ψ sont constantes, on a alors $2(b-a) \leqslant \displaystyle\int_a^b (\varphi - \psi) \leqslant \int_0^1 (\varphi - \psi)$, ce qui montre que f n'est pas $\mathscr{R}$-integrable sur $[0,1]$.

1.3.2 Intégrales supérieure et inférieure

Nous allons voir que l'on peut définir les notions d'intégrale supérieure et d'intégrale inférieure (au sens de Darboux) de *toute fonction bornée* sur $[a,b]$, ce qui nous permettra de reformuler la notion d'intégrabilité au sens de Riemann.

Soit $f : [a,b] \to \mathbb{R}$ une fonction *bornée*. Il existe $M, m \in \mathbb{R}$ tels que $m \leqslant f \leqslant M$. L'ensemble $\mathcal{E}_f^+ := \{\varphi \in \mathcal{E}([a,b] \mid f \leqslant \varphi\}$ n'est pas vide (il contient la fonction constante M) et l'ensemble $\mathcal{E}_f^- := \{\psi \in \mathcal{E}([a,b] \mid \psi \leqslant f\}$ n'est pas vide (il contient la fonction constante m). Si $\varphi \in \mathcal{E}_f^+$, on a $m \leqslant \varphi$, donc l'ensemble $\{\int_a^b \varphi \mid \varphi \in \mathcal{E}_f^+\} \subset \mathbb{R}$ est *minoré* par $\int_a^b m = m(b-a)$. De même si $\psi \in \mathcal{E}_f^-$, alors $\psi \leqslant M$ et l'ensemble $\{\int_a^b \psi \mid \psi \in \mathcal{E}_f^-\} \subset \mathbb{R}$ est *majoré* par $\int_a^b M = M(b-a)$. Cela justifie la définition suivante.

Proposition et définition 8. Soit $f : [a,b] \to \mathbb{R}$ une fonction *bornée*.

(i) L'ensemble $\{\int_a^b \varphi \mid \varphi \in \mathcal{E}_f^+\} \subset \mathbb{R}$ est borné inférieurement. Sa *borne inférieure* est appelée *intégrale supérieure* de f sur $[a,b]$ et notée $\overset{*}{\displaystyle\int_a^b} f$.

(ii) L'ensemble $\{\int_a^b \varphi \mid \psi \in \mathcal{E}_f^-\} \subset \mathbb{R}$ est borné supérieurement. Sa *borne supérieure* est appelée *intégrale inférieure* de f sur $[a,b]$ et notée $\displaystyle\int_{*a}^b f$.

Si f est une fonction en escalier, l'ensemble $\mathcal{E}_f^- := \{\psi \in \mathcal{E}([a,b] \mid \psi \leqslant f\}$ (resp. l'ensemble $\mathcal{E}_f^+ := \{\varphi \in \mathcal{E}([a,b] \mid f \leqslant \varphi\}$) admet f comme plus grand (resp. plus petit) élément. L'intégrale d'une fonction en escalier est donc égale à son intégrale supérieure et à son intégrale inférieure.

> ***Remarque***. Les notions d'intégrale supérieure et inférieure s'étendent au cas des fonctions *bornées* sur un intervalle *borné*.

Exemple. Soit $f : [0,1] \to \mathbb{R}$ définie par $\begin{cases} f(x) = 1 & \text{si } x \in \mathbb{Q}, \\ f(x) = 0 & \text{si } x \notin \mathbb{Q}. \end{cases}$

La fonction f est bornée. Calculons ses intégrales inférieure et supérieure.

Soient φ, ψ des fonctions en escalier sur $[0,1]$ telles que $\psi \leqslant f \leqslant \varphi$. Les ensembles $\mathbb{Q}\cap[0,1]$ et $[0,1]\setminus\mathbb{Q}$ étant denses dans $[0,1]$, on a $\psi \leqslant 0$ et $1 \leqslant \varphi$, sauf peut-être en un nombre fini de points, donc $\displaystyle\int_0^1 \psi \leqslant 0$ et $\displaystyle 1 \leqslant \int_0^1 \varphi$, d'où $\displaystyle 0 \leqslant \underset{*}{\int_0^1} f$ et $\displaystyle 1 \leqslant \overset{*}{\int_0^1} f$. Par ailleurs $0 \leqslant f \leqslant 1$ et 0 et 1 sont des fonctions en escalier, donc $\displaystyle\underset{*}{\int_0^1} f = 0$ et $\displaystyle\overset{*}{\int_0^1} f = 1$.

La proposition suivante donne les propriétés des intégrales supérieure et inférieure vis à vis de l'addition et de la multiplication par les scalaires.

Proposition 9. Soient $f, g : [a,b] \to \mathbb{R}$ deux fonctions *bornées*.
(i) Pour tout $\lambda \geqslant 0$ on a :

$$\overset{*}{\int_a^b} \lambda f = \lambda \overset{*}{\int_a^b} f \quad \text{et} \quad \underset{*}{\int_a^b} \lambda f = \lambda \underset{*}{\int_a^b} f$$

(ii) Pour tout $\lambda < 0$ on a :

$$\overset{*}{\int_a^b} \lambda f = \lambda \underset{*}{\int_a^b} f \quad \text{et} \quad \underset{*}{\int_a^b} \lambda f = \lambda \overset{*}{\int_a^b} f$$

(iii) On a :

$$\overset{*}{\int_a^b} (f+g) \leqslant \overset{*}{\int_a^b} f + \overset{*}{\int_a^b} g \quad \text{et} \quad \underset{*}{\int_a^b} f + \underset{*}{\int_a^b} g \leqslant \underset{*}{\int_a^b} (f+g). \tag{1}$$

Démonstration. Soient φ, ψ, η, θ des fonctions en escalier telles que :

$$\psi \leqslant f \leqslant \varphi \quad \text{et} \quad \theta \leqslant g \leqslant \eta.$$

(i) Pour tout $\lambda \geqslant 0$ on a $\lambda\psi \leqslant \lambda f \leqslant \lambda\varphi$ et $\lambda\varphi$, $\lambda\psi$ sont des fonctions en escalier. Le résultat s'en déduit.
(ii) Pour tout $\lambda < 0$ on a $\lambda\varphi \leqslant \lambda f \leqslant \lambda\psi$ et $\lambda\varphi$, $\lambda\psi$ sont des fonctions en escalier. Le résultat s'en déduit.
(iii) On a $\psi + \theta \leqslant f + g \leqslant \varphi + \eta$ et $\psi + \theta$, $\varphi + \eta$ sont des fonctions en escalier. On en déduit

$$\overset{*}{\int_a^b} (f+g) \leqslant \int_a^b \varphi + \int_a^b \eta \quad \text{et} \quad \int_a^b \psi + \int_a^b \theta \leqslant \underset{*}{\int_a^b} (f+g),$$

puis, en faisant varier φ, ψ, η, θ on en déduit par passage aux $\inf$ (resp. aux $\sup$) les inégalités (1). ∎

Il n'y a pas égalité en général dans les inégalités (1) comme le montre l'exemple suivant.

Exemple. Soient $f := 1_{\mathbb{Q}}$ et $g := -f$ sur $[0,1]$. On a $\displaystyle\overset{*}{\int_0^1} (f+g) = 0$, $\displaystyle\overset{*}{\int_0^1} f = 1$ et $\displaystyle\overset{*}{\int_0^1} g = 0$. La première inégalité de (1) se lit donc $0 \leqslant 1 + 0$. Elle est stricte.

1.3.3 Intégrale d'une fonction intégrable au sens de Riemann

On vérifie facilement le résultat suivant.

Proposition 10.
Soit $f : [a, b] \to \mathbb{R}$. Les deux conditions suivantes sont équivalentes :

(i) f est intégrable au sens de Riemann sur $[a, b]$;

(ii) f est bornée et ses intégrales supérieure et inférieure sont égales.

Cette proposition permet la définition suivante.

Définition 4. Soit $f : [a, b] \to \mathbb{R}$ une fonction intégrable au sens de Riemann sur $[a, b]$. On *définit l'intégrale de f sur $[a, b]$* par :

$$\int_a^b f = \overset{*}{\int_a^b} f = \underset{*}{\int_a^b} f.$$

Si f est une fonction en escalier cette définition coïncide évidemment avec celle donnée précédemment.

Remarque. Si l'on dispose d'une notion d'aire de *certaines* figures planes satisfaisant quelques conditions intuitives (d'une part, si $A \subset B$, aire$(A) \leqslant$ aire(A) et, d'autre part, l'aire du domaine compris entre l'axe des ordonnées et le graphe d'une fonction en escalier positive est l'aire usuelle), on peut montrer que l'aire A du domaine compris entre l'axe des ordonnées et le graphe d'une fonction f positive et intégrable au sens de Riemann (c'est-à-dire $\{(x, y) \in [a, b] \times \mathbb{R} \mid y \leqslant f(x)\}$) est nécessairement égale à $I := \displaystyle\int_a^b f$ en utilisant, dans le style d'Archimède, la méthode *apagogique* d'Eudoxe. Si l'on avait $A < I$, on aurait $I = \displaystyle\underset{*}{\int_a^b} \leqslant A < I$ et une contradiction. Si l'on avait $A > I$, on aurait $I = \displaystyle\overset{*}{\int_a^b} \geqslant A > I$ et également une contradiction. Il ne reste que la possibilité $A = I$.

Exercice 2.
Montrer que la fonction f définie dans l'exemple de la page 554 n'est pas $\mathscr{R}$-intégrable.

Solution. On a $\displaystyle\underset{*}{\int_0^1} f = 0$ et $\displaystyle\overset{*}{\int_0^1} f = 1$, donc $\displaystyle\underset{*}{\int_0^1} f \neq \overset{*}{\int_0^1} f$.
C'est une variante de l'exemple de la page 553.

Dans le module *Introduction à l'intégration* de [L1] nous avons montré la $\mathscr{R}$-intégrabilité des fonctions *continues* sur un segment. Nous allons donner ici une preuve un peu différente en utilisant la continuité uniforme.

Proposition 11. Toute fonction *continue* sur un segment est intégrable au sens de Riemann sur ce segment.

Démonstration. Soit $f : [a,b] \to \mathbb{R}$ une fonction continue.

Nous allons montrer que pour tout $\varepsilon > 0$ il existe des fonctions en escalier $\varphi \in \mathcal{E}_f^+$ et $\psi \in \mathcal{E}_f^-$ telles que $\displaystyle\int_a^b \varphi - \int_a^b \psi \leqslant \varepsilon$.

La fonction f étant continue sur le segment $[a,b]$, elle y est *uniformément continue* d'après le théorème de Heine. Donc pour tout $\varepsilon > 0$ il existe $n \in \mathbb{N}^*$ tel que si $x, y \in [a,b]$ sont tels que $|y-x| \leqslant \frac{b-a}{n}$, l'on ait $\big|f(y) - f(x)\big| \leqslant \frac{\varepsilon}{b-a}$. On définit une subdivision $\mathbf{u}_n$ par $x_i := a + i\frac{b-a}{n}$. On note pour tout $i \in [\![0, n-1]\!]$:

$$M_i := \sup_{x\in[x_i,x_{i+1}]} f(x) \quad \text{et} \quad m_i := \inf_{x\in[x_i,x_{i+1}]} f(x).$$

Il existe $\xi_i, \eta_i \in [x_i, x_{i+1}]$ tels que $f(\xi_i) = M_i$ et $f(\eta_i) = m_i$, donc $0 \leqslant M_i - m_i \leqslant \frac{\varepsilon}{b-a}$.

La fonction $\psi : [a,b] \to \mathbb{R}$ définie par :

$$\begin{cases} \forall i \in [\![0, n-1]\!],\ \forall x \in]x_i, x_{i+1}[, \quad \psi(x) := m_i \\ \forall i \in [\![0, n]\!], \psi(x_i) := f(x_i) \end{cases}$$

est un élément de $\mathcal{E}_f^-$ et la fonction $\varphi : [a,b] \to \mathbb{R}$ définie par :

$$\begin{cases} \forall i \in [\![0, n-1]\!],\ \forall x \in]x_i, x_{i+1}[, \quad \varphi(x) := M_i \\ \forall i \in [\![0, n]\!], \varphi(x_i) := f(x_i) \end{cases}$$

est un élément de $\mathcal{E}_f^+$.

On a :

$$\int_a^b \varphi - \int_a^b \psi = \sum_{i=0}^{n-1} (M_i - m_i)(x_{i+1} - x_i) \leqslant (b-a)\frac{\varepsilon}{b-a} = \varepsilon.$$

Ainsi f est $\mathscr{R}$-intégrable. ∎

1.3.4 Propriétés des fonctions intégrables et de l'intégrale

Proposition 12. L'ensemble $\mathscr{R}\big([a,b]\big)$ est un *sous-espace vectoriel* de $\mathcal{F}\big([a,b]\big)$ et l'application $f \in \mathscr{R}\big([a,b]\big) \mapsto \displaystyle\int_a^b f \in \mathbb{R}$ est une *forme linéaire* sur $\mathscr{R}\big([a,b]\big)$.

Démonstration. Soient $f, g \in \mathscr{R}\big([a,b]\big)$ et $\lambda \in \mathbb{R}$.

En utilisant les assertions (i) et (ii) de la proposition 9 et $\displaystyle\int_a^b f = \overline{\int_a^b} f = \underline{\int_a^b} f$ on obtient $\displaystyle\int_a^b \lambda f = \lambda \int_a^b f$.

On a $\int_a^b f = {}^*\!\int_a^b f = {}_*\!\int_a^b f$, et $\int_a^b g = {}^*\!\int_a^b g = {}_*\!\int_a^b g$. En utilisant (1) on obtient :

$$ {}^*\!\int_a^b (f+g) \leqslant \int_a^b f + \int_a^b g \leqslant {}_*\!\int_a^b (f+g). $$

On a aussi ${}_*\!\int_a^b (f+g) \leqslant {}^*\!\int_a^b (f+g)$, donc ${}_*\!\int_a^b (f+g) = {}^*\!\int_a^b (f+g)$. Ainsi $f+g$ est intégrable et $\int_a^b (f+g) = \int_a^b f + \int_a^b g$. ∎

Exercice 3.
Soient $f, g : [a,b] \to \mathbb{R}$ égales sauf en un *nombre fini* de points. Montrer que g est $\mathscr{R}$-intégrable si, et seulement si, f l'est et qu'alors $\int_a^b f = \int_a^b g$.

Solution. On considère $h := g - f$. Cette fonction est nulle sauf en un nombre fini de points, c'est donc une fonction en escalier et $\int_a^b h = 0$. On a $g = f + h$ et, d'après la proposition ci-dessus g est intégrable si, et seulement si, f l'est et $\int_a^b g = \int_a^b f + \int_a^b h = \int_a^b f$.

Proposition 13. Soient $f, g : [a,b] \to \mathbb{R}$ intégrables au sens de Riemann sur $[a,b]$.
(i) Si $f \geqslant 0$, alors $\int_a^b f \geqslant 0$. La forme linéaire $f \mapsto \int_a^b f$ sur $\mathscr{R}\big([a,b]\big)$ est *positive*.
(ii) Si $f \geqslant g$, alors $\int_a^b f \geqslant \int_a^b g$.

Démonstration. Si $f \geqslant 0$, la fonction nulle appartient à $\mathcal{E}_f^-$, donc $\int_a^b f \geqslant 0$.
Si $g \geqslant f$, $g - f \geqslant 0$ et il suffit d'appliquer le résultat précédent. ∎

Proposition 14. Soit $f : [a,b] \to \mathbb{R}$ intégrable au sens de Riemann sur $[a,b]$. Alors f^+, f^- et $|f|$ sont intégrables au sens de Riemann sur $[a,b]$. On a les inégalités :

$$ \Big|\int_a^b f\Big| \leqslant \int_a^b |f| \leqslant (b-a) \sup_{x\in[a,b]} \big|f(x)\big|. \tag{2} $$

Démonstration. On rappelle que l'on définit les parties positive et négative d'une fonction f par $f^+ := \sup(f, 0)$ et $f^- := \sup(-f, 0)$. On a $f = f^+ - f^-$ et $|f| = f^+ + f^-$. Montrons que si $f : [a,b] \to \mathbb{R}$ est intégrable au sens de Riemann sur $[a,b]$, f^+ l'est aussi.

Soient $\varepsilon > 0$, par définition, il existe des fonctions en escalier φ_ε et ψ_ε sur $[a,b]$ telles que $\psi_\varepsilon \leqslant f \leqslant \varphi_\varepsilon$ et $\displaystyle\int_a^b \varphi_\varepsilon - \int_a^b \psi_\varepsilon \leqslant \varepsilon$.

On vérifie que $\psi_\varepsilon^+ \leqslant f^+ \leqslant \varphi_\varepsilon^+$. Par suite, en changeant f en $-f$, $\varphi_\varepsilon^- \leqslant f^- \leqslant \psi_\varepsilon^-$ et $\psi_\varepsilon^- - \varphi_\varepsilon^- \geqslant 0$. On a $\varphi_\varepsilon - \psi_\varepsilon = \varphi_\varepsilon^+ - \psi_\varepsilon^+ + \psi_\varepsilon^- - \varphi_\varepsilon^-$, donc, $0 \leqslant \varphi_\varepsilon^+ - \psi_\varepsilon^+ \leqslant \varphi_\varepsilon - \psi_\varepsilon$. D'après la proposition 5 de la page 552, les fonctions φ_ε^+ et ψ_ε^+ sont en escalier. Ainsi $\displaystyle\int_a^b \varphi_\varepsilon^+ - \int_a^b \psi_\varepsilon^+ \leqslant \int_a^b \varphi_\varepsilon - \int_a^b \psi_\varepsilon \leqslant \varepsilon$. On en déduit l'intégrabilité de f^+. Celle de f^- est prouvée de façon similaire. On en déduit l'intégrabilité de $|f|$ en utilisant $|f| = f^+ + f^-$ et la proposition 12 ci-dessus.

Il reste à prouver l'inégalité (2). On a $f = f^+ - f^-$ et $|f| = f^+ + f^-$, donc :

$$\Big|\int_a^b f\Big| = \Big|\int_a^b (f^+ - f^-)\Big| \leqslant \Big|\int_a^b f^+\Big| + \Big|\int_a^b f^-\Big| = \int_a^b (f^+ + f^-) = \int_a^b |f|.$$

Soit $M := \sup\limits_{x\in[a,b]} \big|f(x)\big|$.

De $|f| \leqslant M$ l'on déduit, d'après (i), $\displaystyle\int_a^b |f| \leqslant \int_a^b M = (b-a)M$. ∎

> ***Remarque***. Si f est $\mathscr{R}$-intégrable, on peut aussi prouver la $\mathscr{R}$-intégrabilité de $|f|$ et l'inégalité (2) de la proposition ci-dessus en utilisant les propositions 28 et 27 ci-dessous. Nous le ferons plus loin dans le cas plus général des applications à valeurs dans un espace de Banach (*cf.* la proposition 13 de la page ci-contre).

On prendra garde au fait que $|f|$ peut être intégrable au sens de Riemann sur $[a,b]$ sans que f le soit. L'exercice suivant en donne un exemple.

Exercice 4.

Donner un exemple de fonction $f : [a,b] \to \mathbb{R}$ *non-intégrable* au sens de Riemann sur $[a,b]$ telle que $|f|$ soit *intégrable* au sens de Riemann sur $[a,b]$.

Solution. On reprend l'exemple de la fonction f de l'exercice 1 de la page 553. Cette fonction n'est pas $\mathscr{R}$-integrable tandis que sa valeur absolue, qui est la fonction constante 1, l'est.

Proposition 15. Soient $f, g : [a,b] \to \mathbb{R}$ intégrables au sens de Riemann sur $[a,b]$.
(i) Les fonctions $\sup(f,g)$ et $\inf(f,g)$ sont intégrables au sens de Riemann sur $[a,b]$.
(ii) La fonction fg est intégrable au sens de Riemann sur $[a,b]$.

Démonstration.
(i) On a $\sup(f,g) = \frac{1}{2}\big(f+g+|f-g|\big)$ et $\inf(f,g) = \frac{1}{2}\big(f+g-|f-g|\big)$. On déduit donc le résultat de la proposition 14 ci-dessus et de la proposition 12.

(ii) En utilisant $fg = (f^+ - f^-)(g^+ - g^-) = f^+g^+ - f^+g^- - f^-g^+ + f^-g^-$ et la proposition 12, on constate qu'il suffit de prouver le résultat pour f et g *positives*.
Soient $f, g \geqslant 0$ intégrables au sens de Riemann sur $[a,b]$. Elles sont bornées, il existe donc $M > 0$ tel que $\sup(f,g) \leqslant M$. Soit $\varepsilon > 0$, par définition il existe des fonctions en escalier $\varphi, \psi, \eta, \theta$ telles que :

$$\psi \leqslant f \leqslant \varphi, \quad \theta \leqslant g \leqslant \eta, \quad \int_a^b \varphi - \int_a^b \psi \leqslant \varepsilon/2M \quad \text{et} \quad \int_a^b \eta - \int_a^b \theta \leqslant \varepsilon/2M.$$

Quitte à remplacer ψ par ψ^+ et θ par θ^+, on peut supposer que $\psi, \theta \geqslant 0$. Quitte à remplacer φ par $\inf(\varphi, M)$ et η par $\inf(\eta, M)$, on peut aussi supposer que $\varphi \leqslant M$ et $\eta \leqslant M$. On a alors :

$$\varphi\eta - \psi\theta = \eta(\varphi - \psi) + \psi(\eta - \theta) \leqslant M(\varphi - \psi) + M(\eta - \theta).$$

Donc $\displaystyle\int_a^b \varphi\eta - \int_a^b \psi\theta \leqslant \varepsilon$. Par ailleurs $\psi\theta \leqslant fg \leqslant \varphi\eta$ et les fonctions $\varphi\eta$ et $\psi\theta$ sont en escalier (*cf.* la proposition 5). L'intégrabilité de fg s'en déduit. ∎

1.3.5 Relation de Chasles

Proposition 16. Soient $f : [a,b] \to \mathbb{R}$ et $c \in]a,b[$. Pour que f soit $\mathscr{R}$-intégrable sur $[a,b]$ il faut et il suffit que ses restrictions à $[a,c]$ et $[c,b]$ soient $\mathscr{R}$-intégrables respectivement sur $[a,c]$ et $[c,b]$. Alors, en notant f ces restrictions on a :

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

Démonstration. Supposons f $\mathscr{R}$-intégrable sur $[a,b]$. Soient $\varepsilon > 0$ et des fonctions en escalier $\varphi_\varepsilon,\ \psi_\varepsilon \in \mathcal{E}\big([a,b]\big)$ vérifiant les conditions de la définition 3. En appliquant à φ_ε et ψ_ε la proposition 6, on obtient :

$$\int_a^b \varphi_\varepsilon = \int_a^c \varphi_\varepsilon + \int_c^b \varphi_\varepsilon \quad \text{et} \quad \int_a^b \psi_\varepsilon = \int_a^c \psi_\varepsilon + \int_c^b \psi_\varepsilon.$$

Compte tenu de la *positivité* de $\varphi_\varepsilon - \psi_\varepsilon$ sur $[a,b]$, on en déduit :

$$\int_a^c \varphi_\varepsilon - \int_a^c \psi_\varepsilon \leqslant \varepsilon \quad \text{et} \quad \int_c^b \varphi_\varepsilon - \int_c^b \psi_\varepsilon \leqslant \varepsilon.$$

Ainsi les restrictions de f à $[a,c]$ et $[c,b]$ sont $\mathscr{R}$-intégrables respectivement sur $[a,c]$ et $[c,b]$.

Inversement, supposons que les restrictions de f à $[a,c]$ et $[c,b]$ soient $\mathscr{R}$-intégrables. Par définition, pour tout $\varepsilon > 0$ il existe des fonctions en escalier $\varphi_{1,\varepsilon},\ \psi_{1\varepsilon} \in \mathcal{E}\big([a,c]\big)$, $\varphi_{2,\varepsilon},\ \psi_{2,\varepsilon} \in \mathcal{E}\big([c,b]\big)$ vérifiant :

$$\forall x \in [a,c],\ \ \psi_{1,\varepsilon}(x) \leqslant f(x) \leqslant \varphi_{1,\varepsilon}(x), \quad \forall x \in [c,b],\ \ \psi_{2,\varepsilon}(x) \leqslant f(x) \leqslant \varphi_{2,\varepsilon}(x)$$

et

$$\int_a^c \varphi_{1,\varepsilon} - \int_a^c \psi_{1,\varepsilon} \leqslant \varepsilon/2, \quad \int_c^b \varphi_{2,\varepsilon} - \int_c^b \psi_{2,\varepsilon} \leqslant \varepsilon/2. \tag{3}$$

On définit des fonctions en escalier $\varphi_\varepsilon,\ \psi_\varepsilon \in \mathcal{E}\big([a,b]\big)$ par :

$$\begin{cases} \varphi_\varepsilon(x) := \varphi_{1,\varepsilon}(x) & \text{si} \quad x \in [a,c], \\ \varphi_\varepsilon(x) := \varphi_{2,\varepsilon}(x) & \text{si} \quad x \in \,]c,b] \end{cases} \quad \text{et} \quad \begin{cases} \psi_\varepsilon(x) := \psi_{1,\varepsilon}(x) & \text{si} \quad x \in [a,c], \\ \psi_\varepsilon(x) := \psi_{2,\varepsilon}(x) & \text{si} \quad x \in \,]c,b]. \end{cases}$$

On a :

$$\int_a^b \varphi_\varepsilon = \int_a^c \varphi_{1,\varepsilon} + \int_c^b \varphi_{2,\varepsilon} \quad \text{et} \quad \int_a^b \psi_\varepsilon = \int_a^c \psi_{1,\varepsilon} + \int_c^b \psi_{2,\varepsilon},$$

d'où, en utilisant (3) :

$$\int_a^b \varphi_\varepsilon - \int_a^b \psi_\varepsilon = \int_a^c \varphi_{1,\varepsilon} - \int_a^c \psi_{1,\varepsilon} + \int_c^b \varphi_{2,\varepsilon} - \int_c^b \psi_{2,\varepsilon} \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon. \qquad \blacksquare$$

En appliquant la proposition ci-dessus on retrouve un résultat donné dans le module *Intégration* de [L1].

Corollaire 17. Toute fonction continue par morceaux sur un segment est intégrable au sens de Riemann sur ce segment.

La relation de Chasles reste vraie pour les intégrales inférieures et supérieures d'une fonction bornée (la preuve de ce résultat est l'objet de l'exercice II.3.3).

Proposition 18. Soient $f : [a,b] \to \mathbb{R}$ une fonction *bornée* sur $[a,b]$ et $c \in \,]a,b[$. On a :

$$\overset{*}{\int_a^b} f = \overset{*}{\int_a^c} f + \overset{*}{\int_c^b} f \quad \text{et} \quad \underset{*}{\int_a^b} f = \underset{*}{\int_a^c} f + \underset{*}{\int_c^b} f.$$

Conventions

On peut étendre la définition de l'intégrabilité et de l'intégrale à des intervalles compacts *arbitraires* en faisant les conventions suivantes :

1. Soit $a \in \mathbb{R}$. Si $f : \{a\} \to \mathbb{R}$, on dit que f est $\mathscr{R}$-intégrable en a et l'on définit $\displaystyle\int_a^a f = 0$.
2. Soient $a, b \in \mathbb{R}$, avec $b < a$ et $f : [b,a] \to \mathbb{R}$. On dit que f est intégrable sur $[a,b]$ si, et seulement si, elle l'est sur $[b,a]$ et l'on définit son intégrale par $\displaystyle\int_a^b f := -\int_b^a f$.

En utilisant les conventions ci-dessus et la proposition 16, on vérifie facilement le résultat suivant.

Proposition 19. Soient $a, b, c \in \mathbb{R}$ et $f \in \mathscr{R}\big([\min(a,b,c), \max(a,b,c)]\big)$. Alors f est $\mathscr{R}$-intégrable sur chacun des intervalles compacts déterminés par deux quelconques des points a, b, c, et l'on a :

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

Proposition 20. Soit $f : [a,b] \to \mathbb{R}$ *bornée* et $\mathscr{R}$-intégrable sur tout sous-intervalle compact de $]a,b[$. Alors f est $\mathscr{R}$-intégrable sur $[a,b]$.

Démonstration. Soit M un majorant de $|f|$ sur $[a,b]$.

Soit $\varepsilon > 0$. On note $\varepsilon_1 := \inf\left(\frac{b-a}{3}, \frac{\varepsilon}{8M}\right)$. La restriction de f à $[a+\varepsilon_1, b-\varepsilon_1]$ est intégrable, il existe donc des fonctions en escalier η et θ sur $[a+\varepsilon_1, b-\varepsilon_1]$ telles que :

$$\theta \leqslant f \leqslant \eta \quad \text{sur} \quad [a+\varepsilon_1, b-\varepsilon_1] \quad \text{et} \quad \int_{a+\varepsilon_1}^{b-\varepsilon_1} \eta - \int_{a+\varepsilon_1}^{b-\varepsilon_1} \theta \leqslant \varepsilon/2.$$

On prolonge η et θ en des fonctions en escalier φ et ψ sur $[a,b]$ en posant $\varphi := M$ et $\psi := -M$ sur $[a, a+\varepsilon_1] \cup [b-\varepsilon_1, b]$. On a :

$$\psi \leqslant f \leqslant \varphi \quad \text{sur} \quad [a,b] \quad \text{et} \quad \int_a^b \varphi - \int_a^b \psi = \int_{a+\varepsilon_1}^{b-\varepsilon_1} (\eta - \theta) + 4M\varepsilon_1 \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Ainsi f est $\mathscr{R}$-intégrable sur $[a,b]$. ∎

On déduit facilement de la proposition ci-dessus le résultat suivant.

Corollaire 21. Soit $f : [a,b] \to \mathbb{R}$ une fonction *bornée* n'admettant qu'un nombre fini (éventuellement nul) de points de discontinuité. Alors f est $\mathscr{R}$-intégrable.

Exercice 5.

Montrer que la fonction $f : [-1,1] \to \mathbb{R}$ définie par :

$$f(x) := \begin{cases} \sin(1/x) & \text{si} \quad x \neq 0 \\ 0 & \text{si} \quad x = 0. \end{cases}$$

est intégrable au sens de Riemann.

Solution. Cette fonction est continue sur $[-1,1]$ sauf en 0. Elle est donc $\mathscr{R}$-intégrable d'après le corollaire ci-dessus.

1.3.6 Sous-ensembles élémentaires et longueurs

Comme nous l'avons vu dans « Un peu d'Histoire » page 544, les théories de l'intégration ont toujours oscillé entre deux démarches : commencer par *mesurer des ensembles* ou commencer par *intégrer des fonctions*. Nous avons choisi la seconde, nous allons expliquer comment on peut à partir de là « mesurer » *certains* ensembles.

Une théorie de l'intégration sur $\mathbb{R}$ permet de définir la « longueur » ou « mesure » de *certains* sous-ensembles. Par exemple si $A \subset [a,b]$ est tel que sa fonction caractéristique 1_A est $\mathscr{R}$-intégrable, on peut *définir* sa longueur par $\displaystyle\int_a^b 1_A$. On peut aussi essayer de définir dans certains cas la longueur par d'autres méthodes, plus « naïves ». Nous comparons ci-dessous ces deux approches sur quelques exemples simples et signalons une difficulté (sur laquelle nous reviendrons plus loin) qui montre une limitation de la théorie de Riemann et motive le recours à des théories de l'intégration plus puissantes, comme celles de Lebesgue ou de Henstock-Kurzweil.
Cette partie est également utile pour préparer l'étude des intégrales multiples (*cf.* le module intégration de [L2-L3]). *En attendant elle peut être omise en première lecture.*
Nous aurons besoin d'un résultat sur la structure des ouverts de $\mathbb{R}$ (*cf.* aussi l'exercice II.1.36 de la page 472 pour une autre preuve).

Proposition 22. Tout ouvert non vide de $\mathbb{R}$ est une réunion au plus dénombrable d'intervalles ouverts disjoints. La famille de ces intervalles est déterminée de manière unique.

Démonstration. Soit O un ouvert non vide de $\mathbb{R}$. Soit $a \in O$, posons :

$$E_a = \{x \in O \mid [a,x[\subset O\}.$$

Puisque O est ouvert E_a n'est pas vide, ou bien il n'est pas majoré et $[a,+\infty[\subset O$, ou bien E_a admet une borne supérieure μ_a et l'on a $[a,\mu_a[\subset O$, avec $[a,\mu_a] \not\subset O$.

De même posons :

$$F_a = \{y \in O \mid]y,a] \subset O\}.$$

Ou bien F_a n'est pas minoré et $]-\infty,a] \subset O$, ou bien il possède une borne inférieure λ_a et $]\lambda_a,a] \subset O$ avec $[\lambda_a,a] \not\subset O$.

Notons $I_a =]\lambda_a,\mu_a[$ avec éventuellement $\lambda_a = -\infty$ et $\mu_a = +\infty$ (I_a est le plus grand intervalle ouvert de $\mathbb{R}$ au sens large contenant a et inclus dans O).

On a $O = \bigcup_{a\in O} I_a$. Si $I_a \cap I_b \neq \varnothing$, alors $I_a \cup I_b$ est un intervalle inclus dans O, on en déduit que $I_a = I_b$. Les intervalles I_a forment donc une partition de O, $O = \bigcup_{a\in A} I_a$ où A est un sous ensemble convenable de O. On peut choisir $A \subset O \cap \mathbb{Q}$, ce qui prouve que A est au plus dénombrable.

Montrons l'unicité. Soit $O = \bigcup_{k\in K} J_k$ une réunion disjointe d'intervalles ouverts. Si $b \in J_k$, il existe $a \in A$ tel que $b \in I_a$ et d'après la définition de I_a on a $J_k \subset I_a$. Ainsi chaque J_k est contenu dans un I_a *unique*. Puisque $\bigcup_{k\in K} J_k = \bigcup_{a\in A} I_a$, on a nécessairement l'égalité $J_k = I_a$. ∎

On déduit facilement de la proposition ci-dessus que tout ouvert X de $[a,b]$ est réunion disjointe d'une famille *au plus dénombrable* d'intervalles ouverts de $[a,b]$. (On prendra garde au fait que, si $a < c \leqslant b$, $[a,c[$ est un intervalle ouvert de $[a,b]$ mais pas de $\mathbb{R}$.)
On vérifie facilement le résultat suivant.

Proposition et définition 23. Pour un sous-ensemble $X \subset [a,b]$ les conditions suivantes sont équivalentes :
(i) X est réunion *disjointe finie* de sous-intervalles de $[a,b]$;
(ii) X est réunion *finie* de sous-intervalles de $[a,b]$;
(iii) la fonction caractéristique 1_X restreinte à $[a,b]$ est en escalier.
Un sous-ensemble vérifiant ces conditions sera dit *élémentaire*.
On définit sa *longueur* par $\operatorname{long} X := \displaystyle\int_a^b 1_X$.

On notera que dans la proposition-définition ci-dessus on considère *tous* les intervalles : ouverts, fermés, semi-ouverts, réduits à un point ou vide.
On vérifie qu'une fonction en escalier est la fonction caractéristique d'un sous-ensemble élémentaire si, et seulement si, elle ne prend que les valeurs 0 ou 1.

Proposition 24. Si $X, Y \subset [a,b]$ sont des sous-ensembles élémentaires, $X \cup Y$, $X \cap Y$ et $1_{[a,b]\setminus X}$ sont des sous-ensembles élémentaires et l'on a :

$$\operatorname{long} X \cup Y + \operatorname{long} X \cap Y = \operatorname{long} X + \operatorname{long} Y.$$

$$\operatorname{long}\big([a,b] \setminus X\big) = b - a - \operatorname{long}(X)$$

Démonstration.
On a $1_{X\cup Y} = \sup\{1_X, 1_Y\}$, $1_{X\cap Y} = \inf\{1_X, 1_Y\}$, et $1_{X\cup Y} + 1_{X\cap Y} = 1_X + 1_Y$. ∎

On en déduit que la famille de parties de $[a,b]$ formée des sous-ensembles élémentaires est stable par réunion finie, intersection finie et passage au complémentaire (on dit que c'est une *algèbre de Boole*).
Si X est un intervalle d'origine c et d'extrémité d, sa longueur définie ci-dessus coïncide avec sa longueur « ordinaire », c'est-à-dire $d - c$. On a $\operatorname{long} \varnothing = 0$.
Si $X = \displaystyle\bigcup_{k\in[\![1,m]\!]} I_k$ (où, pour tout $k \in [\![1,m]\!]$, I_k est un intervalle de $[a,b]$), la réunion étant *disjointe*, on vérifie que $\operatorname{long} X = \displaystyle\sum_{k\in[\![1,m]\!]} \operatorname{long} I_k$. On notera que la décomposition en réunion finie disjointe d'intervalles *n'est pas unique* mais que, par définition, la longueur ne dépend pas du choix de la décomposition.
Si, plus généralement, on a une réunion disjointe $X = \displaystyle\bigcup_{k\in\mathbb{N}^*} I_k$ (où, pour tout $k \in \mathbb{N}^*$, I_k est un intervalle *ouvert* de $[a,b]$), les sommes partielles de la série à termes positifs $\displaystyle\sum_{k\in\mathbb{N}^*} \operatorname{long} I_k$ sont majorées par $b-a$, donc la série est convergente (et sa somme ne dépend pas de la numérotation des I_k).
On peut donc *étendre* « naïvement » à ce cas la définition de la longueur par $\operatorname{long} X := \displaystyle\sum_{k\in\mathbb{N}^*} \operatorname{long} I_k$.
(Les intervalles étant ouverts, la décomposition est unique.)
Nous savons que tout ouvert X de $[a,b]$ est réunion disjointe d'une famille *au plus dénombrable* d'intervalles ouverts (*cf.* la proposition 22). On peut donc, pour tout ouvert X de $[a,b]$ *définir* $\operatorname{long} X$ à partir du procédé décrit ci-dessus. Cette définition est (sous-une forme un peu plus générale) due à Borel et elle est le point de départ de sa théorie de la mesure.
Si $K \subset [a,b]$ est *compact*, $X := [a,b] \setminus K$ est ouvert, ce qui justifie la définition de la longueur de K par $\operatorname{long} K := b - a - \operatorname{long} X$.

On prendra garde au fait que, si $X \subset [a,b]$ est un ouvert, la fonction caractéristique 1_X *n'est pas nécessairement $\mathscr{R}$-intégrable*. On verra un exemple simple dans l'exercice II.3.5. On peut aussi considérer le cas où X est le complémentaire dans $[0,1]$ d'un ensemble de Cantor « épais » comme le compact de Smith-Volterra-Cantor (*cf.* plus loin la remarque de la page 598 et l'exercice II.3.6).

Ainsi *on ne peut pas toujours définir la longueur d'un ouvert ou d'un compact* en utilisant la théorie de l'intégrale de Riemann. En un certain sens les défauts de la théorie de Riemann « découlent de celui-ci ».
L'exercice ci-dessous montre que si X est ouvert et si 1_X est $\mathscr{R}$-intégrable, la définition de la longueur de X par l'intégrale de Riemann et la définition naïve coïncident. Ainsi la seconde est une *généralisation* (*stricte*) de la première et nous pourrons simplement parler de longueur sans préciser.

Exercice 6.
Soit une réunion disjointe $X = \bigcup_{k\in\mathbb{N}^*} I_k$ (où, pour tout $k \in \mathbb{N}^*$, I_k est un intervalle ouvert de $[a,b]$).

On *suppose* que 1_X est $\mathscr{R}$-intégrable. Montrer que $\operatorname{long} X := \displaystyle\int_a^b 1_X$.

Solution. Notons, pour tout $m \in \mathbb{N}^*$, $X_m = \bigcup_{k\in[\![1,m]\!]} I_k$. C'est un ensemble élémentaire et l'on a $1_{X_m} \leqslant 1_X$. Les fonctions 1_{X_m} étant en escalier on a

$$\operatorname{long} X = \sup_{m\in\mathbb{N}^*} \int_a^b 1_{X_m} \leqslant {}_*\!\int_a^b 1_X = \int_a^b 1_X.$$

Par ailleurs, soit ψ une fonction en escalier sur $[a,b]$, telle que $\psi \leqslant 1_X$. Quitte à la remplacer par $\sup(\psi, 0)$ ($0 \leqslant 1_X$), on peut supposer que $0 \leqslant \psi$. La fonction ψ est alors identiquement nulle sur le complémentaire de $\bigcup_{k\in\mathbb{N}^*} I_k$, donc un intervalle ouvert I sur lequel ψ est constante et non nulle est contenu dans un et un seul des I_k. La fonction ψ est nulle en dehors de la réunion d'un nombre *fini* d'intervalles ouverts disjoints et d'un nombre *fini* de points, il en résulte qu'il existe $N \in \mathbb{N}^*$ tel que ψ soit nulle sur le complémentaire de $X_N = \bigcup_{k\in[\![1,N]\!]} I_k$. On a alors (en utilisant $\psi \leqslant 1$) :

$$\int_a^b \psi \leqslant \int_a^b 1_{X_N} = \operatorname{long} X_N \leqslant \operatorname{long} X.$$

Donc $\displaystyle\int_a^b 1_X = {}_*\!\int_a^b 1_X \leqslant \operatorname{long} X$ et $\displaystyle\int_a^b 1_X = \operatorname{long} X$.

Soit $X \subset [a,b]$ non vide. On désigne par $\lambda^+(X)$ la *borne inférieure* des longueurs des sous-ensembles élémentaires *contenant* X et par $\lambda^-(X)$ la *borne supérieure* des longueurs des sous-ensembles élémentaires *contenus* dans X.

Proposition 25. Soit $X \subset [a,b]$ non vide., on a

$$\lambda^+(X) = {}^*\!\int_a^b 1_X \quad \text{et} \quad \lambda^-(X) = {}_*\!\int_a^b 1_X.$$

Démonstration.
On a $A \subset X$ si, et seulement si, $1_A \leqslant 1_X$ et $X \subset B$ si, et seulement si, $1_X \leqslant 1_B$. Si A et B sont des sous ensembles élémentaires, les fonctions 1_A et 1_B sont en escalier.

De $1_A \leqslant 1_X \leqslant 1_B$ on déduit alors que $\lambda^-(X) \leqslant \underset{*}{\int_a^b} 1_X \leqslant \overset{*}{\int_a^b} 1_X \leqslant \lambda^+(X)$.

Soit ψ une fonction en escalier telle que $\psi \leqslant 1_X$. On note $A := \{x \in [a,b] \mid \psi(x) > 0\}$, on vérifie que c'est un sous-ensemble élémentaire. Si $x \in A$, $1_X(x) > 0$, donc $1_X(x) = 1$ et $x \in X$, donc $A \subset X$. Ainsi $\psi \leqslant 1_A \leqslant 1_X$. On en déduit $\underset{*}{\int_a^b} 1_X \leqslant \lambda^-(X)$, puis $\lambda^-(X) = \underset{*}{\int_a^b} 1_X$. On prouverait de même $\lambda^+(X) = \overset{*}{\int_a^b} 1_X$ ∎

Corollaire 26. Soit $X \subset [a,b]$ non vide. Les deux conditions suivantes sont équivalentes :
(i) 1_X est $\mathscr{R}$-intégrable ;
(ii) $\lambda^+(X) = \lambda^-(X)$.
Si elles sont satisfaites, on a $\int_a^b 1_X = \lambda^+(X) = \lambda^-(X)$.

Si $X \subset [a,b]$ vérifie les conditions ci-dessus on dit que $\int_a^b 1_X$ est sa *mesure de Jordan*.

Dans le module *Intégration* de [L2-L3] nous définirons en dimension finie quelconque des sous-ensembles dits *cubables* (ou aussi *quarrables* en dimension 2) vérifiant des conditions analogues à celles du corollaire ci-dessus. Cette terminologie vient du calcul des aires et des volumes.

Exercice 7.

1. Soit $(A_n)_{n\in\mathbb{N}^*}$ une suite décroissante de sous-ensembles élémentaires de $[0,1]$. On suppose que $\bigcap_{n\in\mathbb{N}^*} A_n = \varnothing$. Montrer que $\lim_{n\to+\infty} \log A_n = 0$.
2. Soit $(f_n : [0,1] \to \mathbb{R})_{n\in\mathbb{N}^*}$ une suite décroissante de fonctions en escalier positives. On suppose qu'elle converge simplement vers 0 sur $[0,1]$. Montrer que $\lim_{n\to+\infty} \int_0^1 f_n = 0$.

Solution.

1. Nous ne démontrerons le résultat (qui n'est pas si simple…) que dans le cas où les parties A_n sont des *compacts* en admettant provisoirement le cas général. Ce cas général fait l'objet de l'exercice II.3.7 (*cf.* aussi l'exercice II.3.31).
   Supposons les A_n sont compacts. Si $\bigcap_{n\in\mathbb{N}^*} A_n = \varnothing$, il existe $N \in \mathbb{N}^*$ tel que, pour tout $n \geqslant N^*$, $A_n = \varnothing$. Sinon, d'après la compacité de $[0,1]$, on aurait $\bigcap_{n\in\mathbb{N}^*} A_n \neq \varnothing$.
   Ainsi, pour tout $n \geqslant N$, $\log A_n = 0$ donc $\lim_{n\to+\infty} \log A_n = 0$.
2. Soit $\varepsilon > 0$, on désigne par $C_n := \{x \in [0,1] \mid f_n(x) > \varepsilon/2\}$.
   Les C_n sont des sous-ensembles élémentaires de $[0,1]$ (puisque f_n est une fonction en escalier) et la suite $(C_n)_{n\in\mathbb{N}^*}$ est décroissante.
   Montrons que $\bigcap_{n\in\mathbb{N}^*} C_n = \varnothing$. Sinon il existerait $x \in \bigcap_{n\in\mathbb{N}^*} C_n$ et l'on aurait $f_n(x) > \varepsilon/2$ pour tout $n \in \mathbb{N}$ et, par suite $\lim_{n\to+\infty} f_n(x) \geqslant \varepsilon/2 > 0$, d'où une contradiction.
   La fonction en escalier f_1 est bornée par $M > 0$. Donc, pour tout $n \in \mathbb{N}^*$, $f_n \leqslant M$.

Pour tout $n \in \mathbb{N}^*$, $f_n = 1_{C_n} \times f_n + 1_{[0,1]\setminus C_n} \times f_n$. Pour tout $n \in \mathbb{N}^*$ et tout $x \in [0,1] \setminus C_n$ on a $f_n(x) \leqslant \varepsilon/2$, donc :

$$\int_0^1 f_n = \int_0^1 1_{C_n} f_n + \int_0^1 1_{[0,1]\setminus C_n} f_n \leqslant M \operatorname{long} C_n + \frac{\varepsilon}{2}.$$

D'après la question 1, $\lim\limits_{n\to+\infty} \operatorname{long} C_n = 0$, il existe donc $N' \in \mathbb{N}^*$, tel que, pour tout $n \geqslant N'$, $\operatorname{long} C_n < \varepsilon/2M$.

Donc, pour tout $n \geqslant N'$, on a $0 \leqslant \int_0^1 f_n < \varepsilon/2 + \varepsilon/2 = \varepsilon$.

1.3.7 Un exemple de fonction intégrable au sens de Riemann ayant un nombre infini non dénombrable de points de discontinuité

Il existe des fonctions intégrables au sens de Riemann qui ont « beaucoup » de discontinuités et qui, en particulier, ne sont pas continues par morceaux. En voici un exemple basé sur l'étude de *l'ensemble de Cantor* (*cf.* page 436).

Exemple. On reprend l'exemple de l'ensemble de Cantor $K_3 \subset [0,1]$, que nous noterons pour simplifier K. Rappelons le principe de la construction. Pour tout $n \in \mathbb{N}$ on définit un sous-ensemble fermé de $[0,1]$ formé par la réunion de 2^n segments de longueur $1/3^n$:

- $F_0 := [0,1]$, $F_1 := [0,1/3] \cup [2/3,1]$, $F_2 := [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1]$;
- F_n est formé de la réunion des segments $[3^{-n}t, 3^{-n}(t+1)]$, t décrivant l'ensemble des entiers de $[\![0, 3^n - 1]\!]$ dont l'écriture en base 3 ne comporte pas le chiffre 1.

La longueur de F_n est $(2/3)^n$.

On a $K := \bigcap\limits_{n\in\mathbb{N}} F_n$; K est un fermé non dénombrable de $[0,1]$.

Nous allons montrer que la fonction indicatrice 1_K de K est intégrable au sens de Riemann et que $\int_0^1 1_K = 0$.

Pour tout $n \in \mathbb{N}$, $0 \leqslant 1_K \leqslant 1_{F_n}$. La fonction nulle et la fonction 1_{F_n} sont des fonctions en escalier et l'on a $\int_0^1 1_{F_n} - \int_0^1 0 = (2/3)^n$. Puisque $(2/3)^n \to 0$ quand $n \to +\infty$, la fonction 1_K est intégrable au sens de Riemann et l'on a $\int_0^1 1_K = 0$.

La fonction 1_K est identiquement nulle sur $[0,1] \setminus K$ qui est une réunion d'intervalles ouverts, elle est donc continue sur $[0,1] \setminus K$.

Montrons que la fonction 1_K est discontinue en tout point de K. Soit $c \in K$, on a $1_K(c) = 1$. Tout intervalle ouvert contenant c contient un point de $[0,1] \setminus K$, sinon K contiendrait un intervalle fermé J de longueur $\ell > 0$, et l'on aurait $J \subset F_n$ pour tout $n \in \mathbb{N}$, ce qui est impossible, car, pour n assez grand, $\operatorname{long} F_n = (2/3)^n < \ell$. Ainsi tout intervalle ouvert contenant c contient un point où 1_K s'annule, par suite 1_K n'est pas continue en c. La fonction 1_K possède donc un *ensemble non dénombrable* de points de discontinuité.

1.3.8 Une autre définition de l'intégrabilité au sens de Riemann

Nous allons terminer ce paragraphe par une autre définition (équivalente à celle donnée ci-dessus) de l'intégrabilité au sens de Riemann. Ceci est utile pour diverses preuves et nous servira aussi plus loin à étendre la notion d'intégrabilité au sens de Riemann à des applications à valeurs dans un espace de Banach E.

Proposition 27. Soit $f : [a, b] \to \mathbb{R}$ une fonction. Les trois conditions suivantes sont équivalentes :

(i) f est intégrable au sens de Riemann ;

(ii) pour tout $\varepsilon > 0$ il existe des fonctions en escalier $\Phi_\varepsilon,\ \Psi_\varepsilon : [a, b] \to \mathbb{R}$ telles que :

$$|f - \Phi_\varepsilon| \leqslant \Psi_\varepsilon \quad \text{et} \quad \int_a^b \Psi_\varepsilon \leqslant \varepsilon;$$

(iii) il existe deux suites $\{\theta_n\}_{n\in\mathbb{N}}$ et $\{\eta_n\}_{n\in\mathbb{N}}$ de $\mathcal{E}\big([a, b]\big)$ telles que :

$$\forall n \in \mathbb{N}, \quad |f - \theta_n| \leqslant \eta_n \quad \text{et} \quad \lim_{n\to+\infty} \int_a^b \eta_n = 0.$$

Un couple de suites satisfaisant la condition (iii) est dit *couple associé* à la fonction intégrable f.

Démonstration. L'équivalence (ii) $\Leftrightarrow$ (iii) est facile (on utilise la suite $\big(\varepsilon_n := \frac{1}{n+1}\big)$). Montrons l'équivalence (i) $\Leftrightarrow$ (ii).

Supposons que f vérifie (ii). Pour tout $\varepsilon > 0$ il existe des fonctions en escalier $\Phi_{\varepsilon/2},\ \Psi_{\varepsilon/2} : [a, b] \to \mathbb{R}$ telles que :

$$|f - \Phi_{\varepsilon/2}| \leqslant \Psi_{\varepsilon/2} \quad \text{et} \quad \int_a^b \Psi_{\varepsilon/2} \leqslant \varepsilon/2.$$

Alors :

$$\Phi_{\varepsilon/2} - \Psi_{\varepsilon/2} \leqslant f \leqslant \Phi_{\varepsilon/2} + \Psi_{\varepsilon/2}.$$

Les fonctions $\varphi_\varepsilon := \Phi_{\varepsilon/2} + \Psi_{\varepsilon/2}$ et $\psi_\varepsilon := \Phi_{\varepsilon/2} - \Psi_{\varepsilon/2}$ sont des fonctions en escalier et l'on a :

$$\psi_\varepsilon \leqslant f \leqslant \varphi_\varepsilon \quad \text{et} \quad \int_a^b \varphi_\varepsilon - \int_a^b \psi_\varepsilon \leqslant 2 \int_a^b \Psi_{\varepsilon/2} \leqslant \varepsilon.$$

Ainsi f est intégrable au sens de Riemann sur $[a, b]$.

Inversement, supposons f intégrable au sens de Riemann sur $[a, b]$. Pour tout $\varepsilon > 0$, il existe des fonctions en escalier $\varphi_{2\varepsilon}$ et $\psi_{2\varepsilon}$ telles que :

$$\psi_{2\varepsilon} \leqslant f \leqslant \varphi_{2\varepsilon} \quad \text{et} \quad \int_a^b \varphi_{2\varepsilon} - \int_a^b \psi_{2\varepsilon} \leqslant 2\varepsilon.$$

Posons :

$$\Phi_\varepsilon := \frac{1}{2}\big(\varphi_{2\varepsilon} + \psi_{2\varepsilon}\big) \quad \text{et} \quad \Psi_\varepsilon := \frac{1}{2}\big(\varphi_{2\varepsilon} - \psi_{2\varepsilon}\big).$$

Les fonctions Φ_ε et Ψ_ε sont des fonctions en escalier et l'on a :

$$f - \Phi_\varepsilon \leqslant \varphi_{2\varepsilon} - \Phi_\varepsilon = \Psi_\varepsilon$$

$$\varphi_{2\varepsilon} - f \leqslant \Phi_\varepsilon - \psi_{2\varepsilon} = \Psi_\varepsilon,$$

donc

$$|f - \Phi_\varepsilon| \leqslant \Psi_\varepsilon \quad \text{et} \quad \int_a^b \Psi_\varepsilon \leqslant \varepsilon.$$

Ainsi f vérifie la condition (ii).

Proposition 28. Soit $f : [a,b] \to \mathbb{R}$ une fonction intégrable au sens de Riemann sur $[a,b]$. Pour tout couple (θ_n, η_n) associé à f, la suite réelle de terme général $\displaystyle\int_a^b \theta_n$ converge vers $\displaystyle\int_a^b f$.

Démonstration. On a $\left|\displaystyle\int_a^b \theta_n - \int_a^b f\right| \leqslant \int_a^b |\theta_n - f| \leqslant \int_a^b \eta_n$. d'où le résultat. ∎

1.3.9 Formule de la moyenne

Proposition 29 (Première formule de la moyenne).
Soient $f, g : [a,b] \to \mathbb{R}$ deux applications $\mathscr{R}$-intégrables sur $[a,b]$. On suppose g *positive*. On note m (resp. M) la borne inférieure (resp. supérieure) de f sur $[a,b]$. Alors il existe $k \in [m,M]$ tel que $\displaystyle\int_a^b fg = k\int_a^b g$.

Si, de plus, f est continue, il existe $c \in [a,b]$ tel que $\displaystyle\int_a^b fg = f(c)\int_a^b g$.

Démonstration. Le produit fg est $\mathscr{R}$-intégrable (*cf.* la proposition 15 de la page 559). La fonction g étant positive, on a $mg \leqslant fg \leqslant Mg$, donc

$$m\int_a^b g \leqslant \int_a^b fg \leqslant M\int_a^b g.$$

Si $\displaystyle\int_a^b g = 0$, tout $k \in [m,M]$ convient, sinon $\dfrac{\displaystyle\int_a^b fg}{\displaystyle\int_a^b g} \in [m,M]$. ∎

Si $g = 1$, on obtient $\displaystyle\int_a^b f = k(b-a)$ et, si f est continue, $\displaystyle\int_a^b f = f(c)(b-a)$.

2 Sommes de Darboux et de Riemann

2.1 Sommes de Darboux

Dans toute cette partie a et b désignent deux nombres réels tels que $a < b$.

2.1.1 Sommes de Darboux, théorème et critère de Darboux

Définition 5 (Sommes de Darboux). Soit $f : [a, b] \to \mathbb{R}$ une fonction *bornée*. À toute subdivision $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ de $[a, b]$ on associe les sommes :

$$d_{\mathbf{u}}(f) := \sum_{i=0}^{n-1} m_i(x_{i+1} - x_i) \quad \text{et} \quad D_{\mathbf{u}}(f) := \sum_{i=0}^{n-1} M_i(x_{i+1} - x_i),$$

où m_i et M_i désignent respectivement la borne inférieure et la borne supérieure de f sur le segment $[x_i, x_{i+1}]$. Les nombres $d_{\mathbf{u}}(f)$ et $D_{\mathbf{u}}(f)$ sont dits *sommes de Darboux inférieure et supérieure* de f pour la subdivision $\mathbf{u}$.

Soit $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ une subdivision de $[a, b]$. La fonction $f^d_{\mathbf{u}} : [a, b] \to \mathbb{R}$ définie par :

$$\begin{cases} \forall i \in [\![0, n-1]\!],\ \forall x \in]x_i, x_{i+1}[, \quad f^d_{\mathbf{u}}(x) := m_i \\ \forall i \in [\![0, n]\!], \quad f^d_{\mathbf{u}}(x_i) := f(x_i) \end{cases}$$

est un élément de $\mathcal{E}^-_f$ et la fonction $f^D_{\mathbf{u}} : [a, b] \to \mathbb{R}$ définie par :

$$\begin{cases} \forall i \in [\![0, n-1]\!],\ \forall x \in]x_i, x_{i+1}[, \quad f^D_{\mathbf{u}}(x) = M_i \\ \forall i \in [\![0, n]\!], \quad f^D_{\mathbf{u}}(x_i) := f(x_i) \end{cases}$$

est un élément de $\mathcal{E}^+_f$.

On a :

$$\int_a^b f^d_{\mathbf{u}} = d_{\mathbf{u}}(f) \quad \text{et} \quad \int_a^b f^D_{\mathbf{u}} = D_{\mathbf{u}}(f).$$

On en déduit :

$$d_{\mathbf{u}}(f) \leqslant \underset{*}{\int_a^b} f \leqslant \overset{*}{\int_a^b} f \leqslant D_{\mathbf{u}}(f). \tag{4}$$

Le lecteur vérifiera que pour toute subdivision $\mathbf{v}$ plus fine que la subdivision $\mathbf{u}$, l'on a :

$$d_{\mathbf{v}}(f) \leqslant d_{\mathbf{u}}(f) \leqslant D_{\mathbf{u}}(f) \leqslant D_{\mathbf{v}}(f). \tag{5}$$

Nous allons voir que, pour toute fonction bornée f sur $[a, b]$, l'intégrale supérieure $\overset{*}{\int_a^b} f$ (resp. inférieure $\underset{*}{\int_a^b} f$) est en un certain sens la « limite » des sommes de Darboux $D_{\mathbf{u}}(f)$ (resp. $d_{\mathbf{u}}(f)$) quand le pas $\delta(\mathbf{u})$ de la subdivision $\mathbf{u}$ tend vers 0.

Le point clé de la preuve est le lemme suivant.

Lemme 30. Soit $f : [a,b] \to \mathbb{R}$ une fonction *bornée* et $g \in \mathcal{E}\big([a,b]\big)$ une fonction en escalier telle que $g \leqslant f$. Pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour toute subdivision $\mathbf{u}$ de $[a,b]$ de pas inférieur à δ, l'on ait :

$$\int_a^b g - \varepsilon \leqslant d_{\mathbf{u}}(f) \leqslant \underset{*}{\int_a^b} f.$$

Démonstration. Soit $\mathbf{v} := (y_j)_{j\in[\![0,k]\!]}$ une subdivision adaptée à la fonction en escalier g fixée.

Soit $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ une subdivision quelconque de $[a,b]$. Nous allons majorer

$$\int_a^b g - d_{\mathbf{u}}(f) = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} (g - m_i).$$

Pour un indice $i \in [\![0, n-1]\!]$, on distingue deux cas (s'excluant mutuellement).

1. Il n'existe pas d'indice $j \in [\![0,k]\!]$ tel que $y_j \in [x_i, x_{i+1}]$. La fonction g est alors constante sur $[x_i, x_{i+1}]$. Puisque $g \leqslant f$, l'on a $g \leqslant m_i$ et par suite :
$$\int_{x_i}^{x_{i+1}} (g - m_i) \leqslant 0. \tag{6}$$

2. Il existe au moins un indice $j \in [\![0,k]\!]$ tel que $y_j \in [x_i, x_{i+1}]$. On a alors, en notant $M := \sup\limits_{x\in[a,b]} f(x)$ et $m := \inf\limits_{x\in[a,b]} f(x)$:
$$\int_{x_i}^{x_{i+1}} (g - m_i) \leqslant (M-m)\delta(\mathbf{u}). \tag{7}$$
En effet, pour tout $x \in [x_i, x_{i+1}]$:
$$g(x) \leqslant \sup_{x\in[x_i,x_{i+1}]} f(x) \leqslant M, \quad m_i = \inf_{x\in[x_i,x_{i+1}]} f(x) \geqslant m$$
et $x_{i+1} - x_i \leqslant \delta(\mathbf{u})$.

Le sous-ensemble de $[\![1, n-1]\!]$ formé des indices i pour lesquels on est dans le cas 2 est formé d'au plus $2(k-1)+2 = 2k$ éléments. En effet, si $j \neq 0, k$ (c'est-à-dire si $y_j \neq a$ et $y_j \neq b$), le point y_j appartient à deux intervalles *au plus* de la forme $[x_i, x_{i+1}]$ et si $j = 0$ ou $j = k$, on a $y_0 = a$ ou $y_k = b$ et y_j appartient à un et un seul intervalle de la forme $[x_i, x_{i+1}]$ ($[a, x_1]$ ou $[x_{n-1}, b]$).

En utilisant les inégalités (6) et (7), on en déduit :

$$\int_a^b g - d_{\mathbf{u}}(f) \leqslant 2k(M-m)\delta(\mathbf{u}). \tag{8}$$

On choisit $\delta > 0$ tel que $2k(M-m)\delta \leqslant \varepsilon$. Si $\delta(\mathbf{u}) \leqslant \delta$, l'on a $\displaystyle\int_a^b g - d_{\mathbf{u}(f)} \leqslant \varepsilon$. ∎

Théorème 31 (Théorème de Darboux). Soit $f : [a,b] \to \mathbb{R}$ une fonction *bornée*. Pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour toute subdivision $\mathbf{u}$ de $[a,b]$ de pas inférieur à δ, l'on ait :

$$0 \leqslant \int_{*a}^{b} f - d_{\mathbf{u}}(f) \leqslant \varepsilon \quad \text{et} \quad 0 \leqslant D_{\mathbf{u}}(f) - \overset{*}{\int_a^b} f \leqslant \varepsilon.$$

On dit aussi improprement que l'intégrale inférieure (resp. supérieure) $\int_{*a}^{b} f$ (resp. $\overset{*}{\int_a^b} f$) est « la limite » des sommes de Darboux inférieures (resp. supérieures) quand le pas de la subdivision tend vers 0. On peut donner un sens rigoureux à cette assertion en utilisant la notion de *base de filtre*, mais nous ne le ferons pas (*cf.* par exemple [R3]).

Démonstration. Il suffit de prouver le résultat pour l'intégrale inférieure (la preuve pour l'intégrale supérieure est similaire).

Soit $\varepsilon > 0$. D'après la définition de l'intégrale inférieure comme borne supérieure il existe une fonction en escalier $g \in \mathcal{E}\big([a,b]\big)$ telle que $g \leqslant f$ et

$$\int_{*a}^{b} f - \frac{\varepsilon}{2} \leqslant \int_a^b g. \tag{9}$$

D'après le lemme 30 (où l'on remplace ε par $\varepsilon/2$), il existe $\delta > 0$ tel que pour toute subdivision $\mathbf{u}$ de $[a,b]$ de pas inférieur à δ, l'on ait :

$$\int_a^b g - \frac{\varepsilon}{2} \leqslant d_{\mathbf{u}(f)}. \tag{10}$$

De (9) et (10) l'on déduit :

$$\int_{*a}^{b} f - \varepsilon \leqslant d_{\mathbf{u}(f)},$$

c'est-à-dire $\int_{*a}^{b} f - d_{\mathbf{u}(f)} \leqslant \varepsilon$. ∎

Notons $d(f) := \inf\limits_{\mathbf{u}} d_{\mathbf{u}}(f)$ et $D(f) := \sup\limits_{\mathbf{u}} d_{\mathbf{u}}(f)$ (les inf et sup étant pris sur toutes les subdivisions de $[a,b]$).

Corollaire 32. On a $d(f) = \int_{*a}^{b} f$ et $D(f) = \overset{*}{\int_a^b} f$.

Démonstration. On a évidemment $d(f) \geqslant \int_{*a}^{b} f$ et $D(f) \leqslant \overset{*}{\int_a^b} f$. On prouve les égalités en utilisant le théorème de Darboux. ∎

Corollaire 33 (Critère de Darboux). Soit $f : [a,b] \to \mathbb{R}$ une fonction. Les deux conditions suivantes sont équivalentes :
(i) f est intégrable au sens de Riemann ;
(ii) f est bornée et, pour tout $\varepsilon > 0$ il existe une subdivision $\mathbf{u} := \{x_i\}_{i\in[\![0,n]\!]}$ telle que

$$D_{\mathbf{u}}(f) - d_{\mathbf{u}}(f) = \sum_{i=0}^{n-1}(M_i - m_i)(x_{i+1} - x_i) \leqslant \varepsilon.$$

Démonstration. Montrons (ii) $\Rightarrow$ (i). Si (ii) est vérifiée, d'après (4), on a, pour tout $\varepsilon > 0$:

$$\overset{*}{\int_a^b} f - \underset{*}{\int_a^b} f \leqslant D_{\mathbf{u}}(f) - d_{\mathbf{u}}(f) \leqslant \varepsilon.$$

donc $\overset{*}{\int_a^b} f = \underset{*}{\int_a^b} f$ et f est intégrable.
Montrons (i) $\Rightarrow$ (ii). D'après le corollaire 32 ci-dessus, si f est intégrable, on a $d(f) = D(f)$. Pour tout $\varepsilon > 0$ il existe des subdivisions $\mathbf{u}$ et $\mathbf{v}$ telles que $0 \leqslant d(f) - d_{\mathbf{v}}(f) \leqslant \varepsilon/2$ et $0 \leqslant D_{\mathbf{u}}(f) - D(f) \leqslant \varepsilon/2$. Soit $\mathbf{w}$ une subdivision plus fine que $\mathbf{u}$ et $\mathbf{v}$, d'après (5) l'on a :

$$0 \leqslant D_{\mathbf{w}}(f) - d_{\mathbf{w}}(f) \leqslant D_{\mathbf{w}}(f) - D(f) + d(f) - d_{\mathbf{w}}(f) \leqslant D_{\mathbf{u}}(f) - D(f) + d(f) - d_{\mathbf{v}}(f) \leqslant \varepsilon.$$

2.1.2 Intégrabilité des fonctions monotones et des fonctions à variation bornée

Fonctions monotones

Le critère de Darboux permet de montrer l'intégrabilité des fonctions monotones.

Proposition 34. Toute fonction monotone $f : [a,b] \to \mathbb{R}$ est intégrable au sens de Riemann sur $[a,b]$.

Démonstration. Supposons f croissante (le cas décroissant se traite de façon similaire). Soit $\mathbf{u} := \{x_i\}_{i\in[\![0,n]\!]}$ une subdivision de $[a,b]$.

Pour tout $n \in \mathbb{N}^*$ on définit une subdivision $\mathbf{u}_n$ de $[a,b]$ par $x_i := a + i\frac{b-a}{n}$. Pour tout $i \in [\![0, n-1]\!]$ les bornes de f sur $[x_i, x_{i+1}]$ sont $m_i = f(x_i)$ et $M_i = f(x_{i+1})$. On a :

$$\sum_{i=0}^{n-1}(M_i - m_i)(x_{i+1} - x_i) = \sum_{i=0}^{n-1}\big(f(x_{i+1} - f(x_i)\big)(x_{i+1} - x_i) = \frac{b-a}{n}\big(f(b) - f(a)\big).$$

Pour tout $\varepsilon > 0$ il existe un entier $n > 0$ tel que $\frac{b-a}{n}\big(f(b) - f(a)\big) \leqslant \varepsilon$. Ainsi f est $\mathscr{R}$-intégrable.

Fonctions à variation bornée

De la $\mathscr{R}$-integrabilité des fonctions monotones on peut déduire celle des fonctions dites *à variation bornée*. Commençons par définir ces dernières.

Définition 6. Soient $(E, \|.\|)$ un espace vectoriel normé réel et $f : [a, b] \to E$ une application. Pour toute subdivision $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ on définit $V(f, \mathbf{u})$ par :

$$V(f, \mathbf{u}) := \sum_{i\in[\![0,n-1]\!]} \big\|f(x_{i+1} - f(x_i)\big\|.$$

On appelle variation (ou variation totale) de f l'élément de $\mathbb{R}^+ \cup \{+\infty\}$ défini par $V(f) := \sup_{\mathbf{u}} V(f, \mathbf{u})$. Si $V(f) < +\infty$, on dit que f est *à variation bornée*.

On a une définition analogue pour les fonctions à valeurs complexes en remplaçant la norme par la valeur absolue.

Exemple. Une fonction réelle monotone sur un segment est à variation bornée. Supposons, pour fixer les idées, que $f : [a, b] \to \mathbb{R}$ est croissante. On a :

$$V(f, \mathbf{u}) := \sum_{i\in[\![0,n-1]\!]} \big|f(x_{i+1} - f(x_i)\big| = \sum_{i\in[\![0,n-1]\!]} \big(f(x_{i+1} - f(x_i)\big) = f(b) - f(a),$$

donc $V(f) = f(b) - f(a) < +\infty$.

La notion d'aplication à variation bornée est liée à celle de *longueur d'un arc de courbe*. Si $(E, \|.\|)$ est un espace euclidien, on dit aussi que l'arc de courbe paramétrée défini par la fonction à variation bornée f est *rectifiable* et que $V(f)$ est sa *longueur*. Ceci généralise le cas d'un arc de courbe plane que nous avons étudié dans le module *Courbes paramétrées* de [L1]. On montre comme dans ce cas le résultat suivant.

Proposition 35. Soient $(E, \|.\|)$ un espace euclidien et $f : [a, b] \to E$ une application de classe $\mathcal{C}^1$. Alors f est à variation bornée et sa variation (c'est-à-dire la longueur de l'arc de courbe correspondant) est $V(f) = \displaystyle\int_a^b \big\|f'(t)\big\|\,\mathrm{d}t$. En particulier une fonction réelle de classe $\mathcal{C}^1$ est à variation bornée et sa variation est $V(f) = \displaystyle\int_a^b \big|f'(t)\big|\,\mathrm{d}t$.

Si $f : [a, b] \to E$ est une application à variation bornée et si $\mathbf{u}$ est la subdivision $(x_0 = a, x_1 = b)$ on a $V(f) \geqslant V(f, \mathbf{u}) = \big\|f(b) - f(a)\big\|$.

Exercice 8.

Soit $f : [a, b] \to \mathbb{R}$ une fonction k-lipschitzienne. Montrer qu'elle est à variation bornée et que $V(f) \leqslant k(b - a)$.

Solution. Pour toute subdivision $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ on a

$$V(f, \mathbf{u}) := \sum_{i\in[\![0,n-1]\!]} \big|f(x_{i+1} - f(x_i)\big| \leqslant k \sum_{i\in[\![0,n-1]\!]} (x_{i+1} - x_i) \leqslant k(b - a).$$

D'où $V(f) \leqslant k(b - a)$.

Si E est un espace vectoriel normé de dimension finie, on vérifie que f est à variation bornée si, et seulement si, ses cordonnées sont à variation bornée. On est ainsi ramené au cas des fonctions réelles à variation bornée.
Si $f, g : [a,b] \to E$ et $\lambda \in \mathbb{R}$, on a :

$$V(\lambda f, \mathbf{u}) = |\lambda|\, V(f, \mathbf{u}) \quad \text{et} \quad V(f+g, \mathbf{u}) \leqslant V(f, \mathbf{u}) + V(g, \mathbf{u}),$$

donc

$$V(\lambda f) = |\lambda|\, V(f) \quad \text{et} \quad V(f+g) \leqslant V(f) + V(g), \tag{11}$$

On en déduit le résultat suivant.

Proposition 36. L'ensemble $BV([a,b], E)$ des applications à variation bornée sur $[a,b]$ à valeurs dans $(E, \|.\|)$ est un sous-espace vectoriel de $\mathcal{F}([a,b], E)$.

Il est possible de munir $BV([a,b], E)$ de normes « raisonnables » qui en font un espace de Banach (*cf.* l'exercice II.3.10).

Proposition 37. Soient $(E, \|.\|)$ un espace vectoriel normé réel, $f : [a,b] \to E$ une application et $c \in]a,b[$. Alors f est à variation bornée si, et seulement si, $f_{|_{[a,c]}}$ et $f_{|_{[c,b]}}$ sont à variation bornée. S'il en est ainsi on a $V(f) = V\big(f_{|_{[a,c]}}\big) + V\big(f_{|_{[c,b]}}\big)$.

Démonstration. Soit $\mathbf{u}$ une subdivision de $[a,b]$. Soit $\mathbf{u}'$ la subdivision de $[a,b]$ obtenue en rajoutant c aux points de subdivision de $\mathbf{u}$ s'il n'y figure pas. Sinon $\mathbf{u}' = \mathbf{u}$. La subdivision $\mathbf{u}'$ induit des subdivisions $\mathbf{u}'_1$ et $\mathbf{u}'_2$ de $[a,c]$ et $[c,b]$ et l'on a :

$$V(f, \mathbf{u}) \leqslant V(f, \mathbf{u}') = V(f_{|_{[a,c]}}, \mathbf{u}'_1) + V(f_{|_{[c,b]}}, \mathbf{u}'_2).$$

Donc, si $f_{|_{[a,c]}}$ et $f_{|_{[c,b]}}$ sont à variation bornée, on a $V(f, \mathbf{u}) \leqslant V(f_{|_{[a,c]}}) + V(f_{|_{[c,b]}})$. Donc f est à variation bornée et l'on a :

$$V(f) \leqslant f_{|_{[a,c]}}) + V(f_{|_{[c,b]}}). \tag{12}$$

Inversement, si $\mathbf{u}_1$ est une subdivision de $[a,c]$, on peut la compléter en une subdivision $\mathbf{u}$ de $[a,b]$ en ajoutant le point b. On a $V(f_{|_{[a,c]}}, \mathbf{u}_1) \leqslant V(f, \mathbf{u}) \leqslant V(f)$. Donc si f est à variation bornée, il en est de même de $f_{|_{[a,c]}}$ et, par un argument similaire de $f_{|_{[c,b]}}$.

Supposons que $f_{|_{[a,c]}}$ et $f_{|_{[c,b]}}$ sont à variation bornée. Soit $\varepsilon > 0$. Il existe des subdivisions $\mathbf{u}_1$ et $\mathbf{u}_2$ de $[a,c]$ et $[c,b]$ respectivement telles que

$$V(f_{|_{[a,c]}}, \mathbf{u}_1) > V(f_{|_{[a,c]}}) - \varepsilon/2 \quad \text{et} \quad V(f_{|_{[c,b]}}, \mathbf{u}_2) > V(f_{|_{[c,b]}}) - \varepsilon/2.$$

Soit $\mathbf{u}$ la subdivision obtenue en « réunissant » $\mathbf{u}_1$ et $\mathbf{u}_2$. On a :

$$V(f) \geqslant V(f, \mathbf{u}) = V(f_{|_{[a,c]}}, \mathbf{u}_1) + V(f_{|_{[c,b]}}, \mathbf{u}_2) \geqslant V(f_{|_{[a,c]}}) + V(f_{|_{[c,b]}}) - \varepsilon.$$

L'inégalité ci-dessus étant vraie pour tout $\varepsilon > 0$, on en déduit $V(f) \geqslant V(f_{|_{[a,c]}}) + V(f_{|_{[c,b]}})$, puis, en utilisant (12), $V(f) = V(f_{|_{[a,c]}}) + V(f_{|_{[c,b]}})$. ∎

Proposition 38. Soit $f : [a, b] \to \mathbb{R}$. Les deux conditions suivantes sont équivalentes.
(i) La fonction f est à variation bornée.
(ii) La fonction f est la somme de deux fonctions monotones.

La condition (ii) peut être remplacée par la condition équivalente suivante :
(ii') La fonction f est la différence de deux fonctions croissantes.

Démonstration. L'implication (ii) $\Rightarrow$ (i) se déduit de ce qui précède (cas des fonctions monotones et proposition 36).
Montrons l'implication (i) $\Rightarrow$ (ii). Supposons que la fonction f soit à variation bornée. D'après la proposition 37, sa restriction à tout segment $[a, t]$ ($t \in [a, b]$) est à variation bornée et la fonction $g : t \in [a, t] \to V\big(f_{|_{[a,t]}}\big) \in \mathbb{R}$ est *croissante* (la variation d'une fonction est toujours positive).
Si $a \leqslant t < t' \leqslant b$:

$$g(t') - g(t) = V((f_{|_{[t,t']}}) \geqslant \big|f(t') - f(t)\big| \geqslant f(t') - f(t),$$

donc $f(t') - g(t') \leqslant f(t) - g(t)$, autrement dit la fonction $h := f - g$ est *décroissante*.
Ainsi $f = g + h$ est la somme de deux fonctions monotones ∎

En utilisant la proposition 36, on en déduit le résultat suivant.

Corollaire 39. Le sous-espace de $\mathcal{F}([a, b], \mathbb{R})$ engendré par les fonctions monotones est l'espace des fonctions à variation bornée.

Corollaire 40. L'ensemble des points de discontinuité d'une fonction à variation bornée est au plus dénombrable.

Démonstration. Montrons d'abord le résultat pour une fonction monotone. Quitte à remplacer f par $-f$, on peut la supposer croissante.
Nous avons vu dans le module *Fonctions réelles* de [L1] qu'une fonction monotone $f : [a, b] \to \mathbb{R}$ admet une limite à droite $f^+(c)$ et une limite à gauche $f^-(c)$ en tout point $c \in]a, b[$.
Soient $c, d \in]a, b[$, avec $c < d$, tels que f ne soit pas continue en c et d. On a :

$$f^-(c) < f^+(c) \leqslant f^-(d) < f^+(d).$$

Les intervalles ouverts $]f^-(c), f^+(c)[$ et $]f^-(d), f^+(d)[$ sont donc *disjoints*.
Notons $(x_i)_{i\in I}$ la famille des points de discontinuité de f sur $]a, b[$. Les intervalles $]f^-(x_i), f^+(x_i)[$, $i \in I$, sont *deux à deux disjoints*. Chacun de ces intervalles contient *au moins un* rationnel. Il existe donc une application *injective* $\theta : I \to \mathbb{Q}$ (telle que $\theta(i) \in]f^-(x_i), f^+(x_i)[$), ce qui prouve que I est au plus dénombrable.
Passons au cas d'une fonction à variation bornée f. On écrit $f = g + h$, où g et h sont monotones. L'ensemble $S(f)$ des points de discontinuité de f est *contenu* dans la réunion de l'ensemble $S(g)$ des points de discontinuité de g et de l'ensemble $S(h)$ des points de discontinuité de h. L'ensemble $S(g) \cup S(h)$ est au plus dénombrable, il en est donc de même de $S(f)$. ∎

La somme de deux fonctions monotones est $\mathscr{R}$-intégrable. On en déduit le résultat suivant.

Corollaire 41. Une fonction réelle à variation bornée sur un segment est intégrable au sens de Riemann sur ce segment.

Une fonction continue n'est pas en général à variation bornée.

Exercice 9.
Montrer que la fonction continue f, définie sur $[0, 2/\pi]$ par :

$$f : x \in [0, 2/\pi] \to \begin{cases} x \sin 1/x & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases}$$

n'est pas à variation bornée.
Solution. Pour tout $N \in \mathbb{N}^*$ on définit une subdivision $\mathbf{u}_N$ de $[0, 2/\pi]$ par :

$$x_0 := 0 \quad \text{et} \quad \forall \quad i \in [\![1, N]\!], \quad x_i := \frac{2}{(2N - 2i + 1)\pi}.$$

On a $f(x_0) = 0$ et, pour tout $i \in [\![1, N]\!]$, $f(x_i) = (-1)^{N+i} \frac{2}{(2N-2i+1)\pi}$, donc :

$$\begin{aligned} V(f, \mathbf{u}_N) &= \sum_{i \in [\![0, N-1]\!]} \big| f(x_{i+1}) - f(x_i) \big| \\ &= \frac{2}{(2N-1)\pi} + \frac{2}{\pi} \sum_{i \in [\![1, N-1]\!]} \Big(\frac{1}{2N - 2i - 1} + \frac{1}{2N - 2i + 1} \Big). \end{aligned}$$

On a :

$$1 + \frac{1}{3} + \cdots + \frac{1}{2p+1} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{2p} + \frac{1}{2p+1} - \frac{1}{2}\Big(1 + \frac{1}{2} + \cdots + \frac{1}{p}\Big).$$

En utilisant $\ln N \sim 1 + \cdots + \frac{1}{N}$, on en déduit $V(f, \mathbf{u}_N) \sim \frac{2}{\pi} \ln N$ quand $N \to +\infty$, ce qui montre que f n'est pas à variation bornée.

2.2 Sommes de Riemann d'une fonction

Définition 7. Soient E un $\mathbb{K}$-espace vectoriel et $f : [a, b] \to E$. Pour toute subdivision pointée $\mathcal{D} := \Big\{ \big([x_i, x_{i+1}], \xi_i \big) \Big\}_{i \in [\![0, n-1]\!]}$ de $[a, b]$ on appelle *somme de Riemann* de f pour $\mathcal{D}$:

$$S_{\mathcal{D}}(f) := \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i) \in E.$$

Un peu d'Histoire

La notion de fonction à variation bornée est due au mathématicien français Camille Jordan (1881). Il était initialement motivé par l'étude des séries de Fourier à la suite des travaux de Dirichlet (les conditions dites de Dirichlet sur la convergence des séries de Fourier n'étaient pas stables par addition). Plus tard (paradoxalement...), Jordan a découvert que la notion de fonction à variation bornée est liée à la rectification des courbes.
Stieltjes a utilisé (en 1894) la notion de fonction à variation bornée pour définir l'intégrale qui porte son nom (intégrale de Stieltjes ou de Riemann-Stieltjes). L'idée est de remplacer dans les sommes de Riemann les longueurs $x_{i+1} - x_i$ par les $g(x_{i+1}) - g(x_i)$, g étant une fonction à variation bornée. L'intégrale de Stieljes est utile en arithmétique (théorie analytique des nombres), en probabilités et en analyse fonctionnelle : en 1909, Frédéric Riesz a interprété les fonctionnelles linéaires sur les espaces de fonctions continues sur un segment en termes d'intégrales de Stieltjes. Basée sur la généralisation de *mesure* d'un intervalle, l'intégrale de Stieltjes a été un premier pas vers l'intégrale de Lebesgue.
À partir d'un travail de Tonelli (en 1926), la théorie des fonctions à variation bornée a été étendue à plusieurs variables avec de nombreuses applications en théorie des équations aux dérivées partielles, en calcul des variations, en physique, chimie, technologie et plus récemment en traitement d'images...

Théorème 42. Soit $f : [a, b] \to \mathbb{R}$. Les conditions suivantes sont équivalentes.
(i) La fonction f est intégrable au sens de Riemann sur $[a, b]$.
(ii) Il existe un réel I tel que pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour toute subdivision pointée $\mathcal{D}$ de $[a, b]$ de pas inférieur à δ, l'on ait $\left|S_{\mathcal{D}}(f) - I\right| \leqslant \varepsilon$.
(iii) La fonction f est bornée et l'on a $d(f) = D(f)$.
Si ces conditions sont satisfaites, on a :

$$I = \int_a^b f = {}_*\!\int_a^b f = {}^*\!\int_a^b f = d(f) = D(f).$$

La condition (ii) est la définition originellement donnée par Riemann. On dit aussi improprement que $I = \displaystyle\int_a^b f$ est « la limite » des sommes de Riemann quand le pas de la subdivision tend vers 0. On peut donner un sens rigoureux à cette assertion en utilisant la notion de *base de filtre)*, mais nous ne le ferons pas (*cf.* par exemple [R3]).

Démonstration. Montrons l'équivalence (i)⇔(ii).
Supposons que f soit intégrable au sens de Riemann. Ses intégrales supérieure et inférieure sont égales. D'après le théorème de Darboux (théorème 31 de la page 572), pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour toute subdivision $\mathbf{u}$ de $[a, b]$ de pas inférieur à δ, l'on ait :

$$0 \leqslant \int_a^b f - d_{\mathbf{u}}(f) \leqslant \varepsilon \quad \text{et} \quad 0 \leqslant D_{\mathbf{u}}(f) - \int_a^b f \leqslant \varepsilon.$$

Si $\mathcal{D}$ est une subdivision pointée de $[a, b]$ de pas inférieur à δ, on a $d_{\mathbf{u}}(f) \leqslant S_{\mathcal{D}}(f) \leqslant D_{\mathbf{u}}(f)$ ($\mathbf{u}$ désignant la subdivision non pointée sous-jacente à $\mathcal{D}$), donc $\left|S_{\mathcal{D}}(f) - \displaystyle\int_a^b f\right| \leqslant \varepsilon$.

Inversement, supposons que la condition (ii) soit satisfaite. Montrons que la fonction f est *bornée*. En utilisant la condition (ii) avec $\varepsilon := 1$, on obtient $\delta_1 > 0$ tel que pour toute subdivision pointée $\mathcal{D}$ de pas inférieur à δ_1 l'on ait $\big|S_{\mathcal{D}}(f) - I\big| \leqslant 1$. Fixons une subdivision $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ de pas inférieur à δ_1 et $k \in [\![0, n-1]\!]$. On considère les subdivisions pointées $\mathcal{D} := \Big\{\big([x_i, x_{i+1}], \xi_i\big)\Big\}_{i\in[\![0,n-1]\!]}$ telles que $\xi_i = x_i$ si $i \neq k$, $\xi_k \in [x_k, x_{k+1}]$ étant *arbitraire*. On a :

$$\big|S_{\mathcal{D}}(f) - I\big| = \big|(x_{k+1} - x_k)f(\xi_k) + \sum_{i\neq k}(x_{i+1} - x_i)f(x_i) - I\big| \leqslant 1,$$

d'où :

$$(x_{k+1} - x_k)\big|f(\xi_k)\big| \leqslant \big|\sum_{i\neq k}(x_{i+1} - x_i)f(x_i)\big| + |I| + 1.$$

On en déduit que f est bornée sur $[x_k, x_{k+1}]$, puis, $k \in [\![0, n-1]\!]$ étant arbitraire, que f est bornée sur $[a, b]$.

D'après la condition (ii) et le théorème de Darboux (f est bornée), pour tout $\varepsilon > 0$ il existe $\delta > 0$ tel que pour toute subdivision pointée $\mathcal{D}$ de pas inférieur à δ l'on ait :

$$\underset{*}{\int_a^b} f - d_{\mathbf{u}}(f) \leqslant \varepsilon \quad \text{et} \quad \big|S_{\mathcal{D}}(f) - I\big| \leqslant \varepsilon, \tag{13}$$

$\mathbf{u}$ désignant toujours la subdivision non pointée sous-jacente à $\mathcal{D}$.

Fixons une subdivision non-pointée $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ de pas inférieur à δ.

Puisque, pour tout $i \in [\![0, n-1]\!]$, $m_i = \inf\limits_{\xi\in[x_i,x_{i+1}]} f(\xi)$, il existe $\zeta_i \in [x_i, x_{i+1}]$ tel que $0 \leqslant f(\zeta_i) - m_i \leqslant \varepsilon/(b-a)$.

Considérons la subdivision pointée $\mathcal{D}_0 := \Big\{\big([x_i, x_{i+1}], \zeta_i\big)\Big\}_{i\in[\![0,n-1]\!]}$, on a :

$$0 \leqslant S_{\mathcal{D}_0}(f) - d_{\mathbf{u}}(f) = \sum_{i\in[\![0,n-1]\!]} (x_{i+1} - x_i)\big(f(\zeta_i) - m_i\big) \leqslant \frac{\varepsilon}{b-a} \sum_{i\in[\![0,n-1]\!]} (x_{i+1} - x_i) = \varepsilon.$$

Ainsi $\big|S_{\mathcal{D}_0}(f) - d_{\mathbf{u}}(f)\big| \leqslant \varepsilon$.

Par ailleurs, d'après (13), on a $\underset{*}{\int_a^b} f - d_{\mathbf{u}}(f) \leqslant \varepsilon$ et $\big|S_{\mathcal{D}_0}(f) - I\big| \leqslant \varepsilon$. On en déduit que $\left|\underset{*}{\int_a^b} f - I\right| \leqslant 3\varepsilon$, puis, $\varepsilon > 0$ étant arbitraire, que $\underset{*}{\int_a^b} f = I$. On montre de même que $\overset{*}{\int_a^b} f = I$, donc $\overset{*}{\int_a^b} f = \underset{*}{\int_a^b} f = I$. Ainsi f est intégrable et $I = \int_a^b f$.

L'équivalence (i) $\Leftrightarrow$ (iii) se déduit du corollaire 32. ∎

Le théorème ci-dessus permet, comme l'avait fait initialement Riemann, de définir la $\mathscr{R}$-intégrabilité à l'aide des sommes de Riemann (condition (ii)). Une variante de cette définition, basée sur la notion de *jauge* et de *subdivision adaptée à une jauge* (*cf.* plus loin la définition 10), permet de définir une nouvelle notion d'intégrale, dite *de Henstock-Kurzweil*, plus générale que celle de Riemann et possédant de bien meilleures propriétés. Nous étudierons en détails cette intégrale dans le module *intégration* du second volume [L2-L3].

Corollaire 43.
Soit $f : [a,b] \to \mathbb{R}$ une fonction $\mathscr{R}$-intégrable. Pour toute suite de subdivisions pointées $(\mathcal{D}_n)_{n\in\mathbb{N}}$, telle que $\lim_{n\to+\infty} \delta(\mathcal{D}_n) = 0$, la suite $\big(S_{\mathcal{D}_n}(f)\big)$ converge vers $\displaystyle\int_a^b f$.
En particulier les deux suites, dites sommes de Riemann régulières à gauche et à droite :

$$\Big(\frac{b-a}{p}\sum_{i=0}^{p-1} f\big(a+i\frac{b-a}{p}\big)\Big)_{p\in\mathbb{N}^*} \quad \text{et} \quad \Big(\frac{b-a}{p}\sum_{i=1}^{p} f\big(a+i\frac{b-a}{p}\big)\Big)_{p\in\mathbb{N}^*}$$

sont convergentes et admettent pour limite $\displaystyle\int_a^b f$.

Pour f continue on retrouve la notion d'intégrale introduite par Cauchy : les sommes de Riemann régulières à gauche sont aussi appelées *sommes de Cauchy*.
Pour une fonction f de classe $\mathcal{C}^2$ on peut étudier l'erreur faite en remplaçant l'intégrale $\displaystyle\int_a^b f$ par une somme de Riemann régulière (*cf.* l'exercice II.3.12).

3 Intégrale de Riemann d'une application à valeurs dans un espace vectoriel de dimension finie

Dans cette section nous étudions l'intégrale de Riemann d'une application à valeurs dans un espace vectoriel réel ou complexe *de dimension finie* E. Dans cette étude, très élémentaire, on se ramène essentiellement au cas scalaire réel, le seul point délicat étant l'inégalité $\left\|\displaystyle\int_a^b f\right\| \leqslant \int_a^b \|f\|$ quand E est muni d'une norme $\|.\|$ (que nous prouvons ici à partir des sommes de Riemann).
En première lecture, on pourra se limiter à cette section 3, suffisante pour traiter nombre d'applications, par exemple en géométrie ou en physique. Nous revenons sur la question dans la section suivante 4 où nous traitons le cas plus général des applications à valeurs dans un espace de Banach en utilisant une approche *différente*. En lisant la section 4 le lecteur est invité à « oublier » la section 3 jusqu'à la proposition 62 de la page 587 qui fournit l'équivalence entre les définitions 8 et 56. Le lecteur intrépide peut aussi choisir de ne pas lire la section 3 en passant directement à la section 4 qui en est indépendante et contient tous les résultats…

3.1 Définitions

Définition 8. Soit $f : [a,b] \to \mathbb{C}$. On dit que f est intégrable au sens de Riemann (ou $\mathscr{R}$-intégrable) sur $[a,b]$ si sa partie réelle $\operatorname{Re} f$ et sa partie imaginaire $\operatorname{Im} f$ le sont.
On définit alors l'intégrale $\displaystyle\int_a^b f$ par : $\displaystyle\int_a^b f := \int_a^b \operatorname{Re} f + \Big(\int_a^b \operatorname{Im} f\Big)\, i$.

Proposition et définition 44. Soient E un $\mathbb{K}$-espace vectoriel de dimension finie et $f : [a,b] \to E$. On dit que f est intégrable au sens de Riemann (ou $\mathscr{R}$-intégrable) sur $[a,b]$ si ses fonctions coordonnées $f_1, \dots, f_m$ sur une base $\mathcal{B} := (e_1, \dots, e_m)$ de E le sont. Cette propriété est indépendante de la base ainsi que le vecteur $\Big(\int_a^b f_1\Big)e_1 + \cdots + \Big(\int_a^b f_m\Big)e_m \in E$.
Ce vecteur est appelé intégrale de f sur $[a,b]$ et noté $\displaystyle\int_a^b f$.

Démonstration. Soit $\mathcal{B}' := (\varepsilon_1, \dots, \varepsilon_m)$. On note $A := (a_{ij}) = \mathrm{Pass}(\mathcal{B}, \mathcal{B}')$ la matrice de changement de base. Soient $(g_1, \dots, g_m)$ les fonctions coordonnées de f dans la base $\mathcal{B}'$:

$$f = f_1 e_1 + \cdots + f_m e_m = g_1\varepsilon_1 + \cdots + g_m\varepsilon_m.$$

On a, pour tout $i \in [\![1,m]\!]$, $f_i = \displaystyle\sum_{j\in[\![1,m]\!]} a_{ij} g_j$. Donc, si les g_j, $j \in [\![1,m]\!]$, sont $\mathscr{R}$-intégrables il en est de même des f_i, $i \in [\![1,m]\!]$, et l'on a $\displaystyle\int_a^b f_i = \sum_{j\in[\![1,m]\!]} a_{ij} \int_a^b g_j$.

On a :

$$\sum_{j\in[\![1,m]\!]} \Big(\int_a^b g_j\Big)\varepsilon_j = \sum_{j\in[\![1,m]\!]} \sum_{i\in[\![1,m]\!]} a_{ij}\Big(\int_a^b g_j\Big)e_i = \sum_{i\in[\![1,m]\!]} \Big(\sum_{j\in[\![1,m]\!]} a_{ij}\int_a^b g_j\Big)e_i$$
$$= \Big(\int_a^b f_1\Big)e_1 + \cdots + \Big(\int_a^b f_m\Big)e_m$$

3.2 Propriétés

Les preuves des résultats suivants sont très simples et laissées au lecteur.

Proposition 45. Soit E un $\mathbb{K}$-espace vectoriel de dimension finie, l'ensemble $\mathscr{R}\big([a,b],E\big)$ des fonctions intégrables au sens de Riemann sur $[a,b]$ est un sous-espace vectoriel de $\mathcal{F}\big([a,b],E\big)$. L'application $f \in \mathscr{R}\big([a,b],E\big) \mapsto \displaystyle\int_a^b f \in E$ est $\mathbb{K}$-linéaire.

Proposition 46. Soient E et F deux $\mathbb{K}$-espaces vectoriels de dimension finie et $u : E \to F$ une application linéaire. Si $f : [a,b] \to E$ est une application intégrable au sens de Riemann, $u \circ f$ est une application intégrable au sens de Riemann de $[a,b]$ dans F et l'on a $\displaystyle\int_a^b u \circ f = u\Big(\int_a^b f\Big)$.

Proposition 47. Soient E, F et G trois $\mathbb{K}$-espaces vectoriels de dimension finie et $\Phi : E \times F \to G$ une application bilinéaire.
Si $f : [a,b] \to E$ et $g : [a,b] \to F$ sont deux applications intégrables au sens de Riemann sur $[a,b]$, $\Phi \circ (f,g) : t \in [a,b] \mapsto \Phi\big(f(t), g(t)\big) \in G$ est une application intégrable au sens de Riemann sur $[a,b]$.

Corollaire 48. Soient E un $\mathbb{K}$-espace vectoriel de dimension finie, $f : [a,b] \to \mathbb{R}$ et $g : [a,b] \to E$ une fonction et une application intégrables au sens de Riemann sur $[a,b]$. Alors $fg : [a,b] \to E$ est intégrable au sens de Riemann sur $[a,b]$.

Formule de Chasles

On étend facilement aux applications à valeurs dans un espace vectoriel de dimension finie E la proposition 16 de la page 560 (on utilise des coordonnées sur E).

Proposition 49 (Formule de Chasles). Soient E un $\mathbb{K}$-espace vectoriel de dimension finie, $f : [a,b] \to E$ et $c \in]a,b[$. Pour que f soit $\mathscr{R}$-intégrable sur $[a,b]$ il faut et il suffit que ses restrictions à $[a,c]$ et $[c,b]$ soient $\mathscr{R}$-intégrables respectivement sur $[a,c]$ et $[c,b]$. Alors, en notant f ces restrictions on a :

$$\int_a^b f = \int_a^c f + \int_c^b f. \tag{14}$$

Si $b \leqslant a$, on généralise la définition de l'intégrale $\displaystyle\int_a^b f$ d'une application f à valeurs dans E en utilisant des conventions similaires à celles adoptées pour les fonctions (*cf.* page 561). Comme dans le cas des fonctions on peut alors étendre la formule de Chasles

Proposition 50. Soient $a,b,c \in \mathbb{R}$ et $f \in \mathscr{R}\big([\min(a,b,c),\max(a,b,c)],E\big)$. Alors f est $\mathscr{R}$-intégrable sur chacun des intervalles compacts déterminés par deux quelconques des points a,b,c, et l'on a :

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

3.3 Sommes de Riemann

Dans ce qui précède nous n'avons pas utilisé de norme sur l'espace vectoriel de dimension finie E : la définition de la $\mathscr{R}$-intégrabilité donnée plus haut (*cf.* la définition 44) n'utilise que la structure d'espace vectoriel de E. Dans ce paragraphe nous considérons par contre le cas d'un $\mathbb{K}$-espace vectoriel normé.

Théorème 51. Soient $(E,\|.\|)$ un $\mathbb{K}$-espace vectoriel normé *de dimension finie* et $f : [a,b] \to E$.
Les conditions suivantes sont équivalentes.
(i) La fonction f est intégrable au sens de Riemann sur $[a,b]$.
(ii) Il existe $I \in E$ tel que, pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour toute subdivision pointée $\mathcal{D}$ de $[a,b]$ de pas inférieur à δ, l'on ait $\big\|S_{\mathcal{D}}(f) - I\big\| \leqslant \varepsilon$.
Si ces conditions sont satisfaites, on a :

$$I = \int_a^b f.$$

Démonstration. Sur E toutes les normes sont équivalentes. Il suffit donc de prouver le résultat pour $E = \mathbb{K}^n$ et $\big\|(v_1, \dots, v_n)\big\| := \sup\{|v_1|, \dots, |v_n|\}$.
Montrons (i) $\Rightarrow$ (ii).

Supposons f $\mathscr{R}$-intégrable. On va montrer (ii) avec $I := \displaystyle\int_a^b f$.

On a, en notant $f = (f_1, \dots, f_n)$:

$$\Big\|S_{\mathcal{D}}(f) - \int_a^b f\Big\| = \sup_{i\in[\![1,n]\!]} \Big\|S_{\mathcal{D}}(f_i) - \int_a^b f_i\Big\|.$$

On obtient le résultat en appliquant le théorème 42 de la page 578 aux fonctions f_i ($i \in [\![1, n]\!]$) qui sont $\mathscr{R}$-intégrables.
Montrons (ii) $\Rightarrow$ (i).
On note $I = (I_1, \dots, I_n)$. Soit $\varepsilon > 0$, il existe $\delta > 0$, tel que pour toute subdivision pointée $\mathcal{D}$ de $[a, b]$ de pas inférieur à δ, l'on ait $\big\|S_{\mathcal{D}}(f) - I\big\| \leqslant \varepsilon$. Alors, pour tout $i \in [\![1, n]\!]$, on a $|S_{\mathcal{D}}(f_i) - I_i| \leqslant \varepsilon$. On en déduit, en utilisant le théorème 42 de la page 578, que les fonctions f_i sont $\mathscr{R}$-intégrables. Par suite f est $\mathscr{R}$-intégrable. ∎

Proposition 52. Soient $(E, \|.\|)$ un espace normé de dimension finie sur $\mathbb{K}$ et $f : [a, b] \to E$ une application $\mathscr{R}$-intégrable sur $[a, b]$. La fonction $\|f\| : [a, b] \to \mathbb{R}$ est $\mathscr{R}$-intégrable sur $[a, b]$ et l'on a :

$$\left\|\int_a^b f\right\| \leqslant \int_a^b \|f\| \leqslant (b-a) \sup_{x\in[a,b]} \big\|f(x)\big\|. \tag{15}$$

Démonstration. Il suffit de montrer le résultat pour $E := \mathbb{K}^n$.
On note $N(v) = N(v_1, \dots, v_n) := \sup\{|v_1|, \dots, |v_n|\}$. Il existe $C > 0$ tel que, pour tout $v \in \mathbb{K}^n$, $\|v\| \leqslant CN(v)$.

1. Montrons que si f est $\mathscr{R}$-intégrable, $\|f\|$ est $\mathscr{R}$-intégrable. On utilise pour cela la proposition 27 de la page 568. On note $(f_i)_{i\in[\![1,n,]\!]}$ les fonctions coordonnées de f.
   Soit $\varepsilon > 0$, les fonctions f_i ($i \in [\![1, n]\!]$) étant $\mathscr{R}$-intégrables, il existe des fonctions en escalier Φ_i et Ψ_i telles que :

$$|f_i - \Phi_i| \leqslant \Psi_i \quad \text{et} \quad \int_a^b \Psi_i \leqslant \varepsilon/C.$$

   D'où, en notant :

$$\Phi := (\Phi_1, \dots, \Phi_n) : [a, b] \to \mathbb{K}^n \quad \text{et} \quad \Psi := \sup_{i\in[\![1,n]\!]} \Psi_i,$$

$$N(f - \Phi) \leqslant \Psi \quad \text{et} \quad \|f - \Phi\| \leqslant CN(f - \Phi) \leqslant C\Psi.$$

   Les applications Φ et Ψ sont en escalier, donc la fonction $\|\Phi\|$ est en escalier et l'on a :

$$\big|\|f\| - \|\Phi\|\big| \leqslant \|f - \Phi\| \leqslant C\Psi \quad \text{et} \quad \int_a^b C\Psi \leqslant \varepsilon$$

En utilisant la proposition 27, on en déduit que $\|f\|$ est $\mathscr{R}$-intégrable sur $[a,b]$.

2. Soit $S_{\mathcal{D}}(f) := \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1}-x_i)$ une somme de Riemann. On a :

$$\big\|S_{\mathcal{D}}(f)\big\| = \Big\|\sum_{i=0}^{n-1} f(\xi_i)(x_{i+1}-x_i)\Big\| \leqslant \sum_{i=0}^{n-1} \big\|f(\xi_i)\big\|(x_{i+1}-x_i) = S_{\mathcal{D}}\big(\|f\|\big).$$

En utilisant la proposition 51 ci-dessus et la $\mathscr{R}$-intégrabilité de $\|f\|$, on en déduit $\left\|\int_a^b f\right\| \leqslant \int_a^b \|f\|$. Par ailleurs on a $\int_a^b \|f\| \leqslant (b-a) \sup_{x\in[a,b]} \big\|f(x)\big\|$. ∎

Remarque. On vérifie que si f est *continue par morceaux*, $\|f\|$ est continue par morceaux (il suffit de le faire pour la norme « sup »), donc $\mathscr{R}$-intégrable. Dans ce cas la preuve 1 ci-dessus est donc inutile.

4 Intégrale de Riemann d'une application à valeurs dans un espace de Banach

4.1 Applications en escalier

Dans la partie 1.2 nous avons défini les applications en escalier à valeurs dans un espace de Banach et étudié leurs premières propriétés. Nous allons donner ici quelques compléments. Les preuves sont faciles et laissées au lecteur.

Proposition 53. Soient E et F deux $\mathbb{K}$-espaces vectoriels et $u : E \to F$ une application linéaire. Si $f : [a,b] \to E$ est une application en escalier, $u \circ f$ est une application en escalier de $[a,b]$ dans F et l'on a $\displaystyle\int_a^b u \circ f = u\Big(\int_a^b f\Big)$.

Proposition 54. Soient E, F et G trois $\mathbb{K}$-espaces vectoriels et $\Phi : E \times F \to G$ une application bilinéaire. Si $f : [a,b] \to E$ et $g : [a,b] \to F$ sont deux applications en escalier, $\Phi \circ (f,g) : t \in [a,b] \mapsto \Phi\big(f(t), g(t)\big) \in G$ est une application en escalier.

Proposition 55. Soient E un $\mathbb{K}$-espace vectoriel normé et $f : [a,b] \to E$ une application en escalier. Alors $\|f\| : t \in [a,b] \mapsto \big\|f(t)\big\| \in \mathbb{R}$ est une application en escalier et l'on a $\left\|\displaystyle\int_a^b f\right\| \leqslant \int_a^b \|f\|$.

4.2 Applications intégrables au sens de Riemann

Proposition et définition 56. Soient E un espace de Banach sur $\mathbb{K}$ et $f : [a,b] \to E$ une application. Les deux conditions suivantes sont équivalentes :

(i) Pour tout $\varepsilon > 0$ il existe des applications en escalier $\Phi_\varepsilon : [a,b] \to E$ et $\Psi_\varepsilon : [a,b] \to \mathbb{R}$ telles que :

$$\forall x \in [a,b], \quad \big\|f(x) - \Phi_\varepsilon(x)\big\| \leqslant \Psi_\varepsilon(x) \quad \text{et} \quad \int_a^b \Psi_\varepsilon \leqslant \varepsilon;$$

(ii) il existe une suite $\{\theta_n\}_{n\in\mathbb{N}}$ d'applications en escalier de $[a,b]$ dans E et une suite $\{\eta_n\}_{n\in\mathbb{N}}$ de $\mathcal{E}\big([a,b]\big)$ telles que :

$$\forall n \in \mathbb{N},\ \forall t \in [a,b], \quad \big\|f(x) - \theta_n(x)\big\| \leqslant \eta_n(x) \quad \text{et} \quad \lim_{n\to+\infty} \int_a^b \eta_n = 0.$$

Si ces conditions sont satisfaites, on dit que f est *intégrable au sens de Riemann* sur $[a,b]$.

Dans le cas où $E = \mathbb{R}$ cette définition est équivalente à la définition 3 de la page 553 (*cf.* la proposition 27 de la page 568).
Soit N une norme sur E *équivalente* à $\|.\|$. Il existe $C > 0$ tel que, pour tout $v \in E$, l'on ait $N(v) \leqslant C\|v\|$. Alors, pour tout $x \in [a,b]$:

$$N\big(f(x) - \Phi_\varepsilon(x)\big) \leqslant C\,\big\|f(x) - \Phi_\varepsilon(x)\big\| \leqslant C\,\Psi_\varepsilon(x).$$

On en déduit que si f est $\mathscr{R}$-intégrable en tant qu'application dans $\big(E, \|.\|\big)$, elle l'est aussi en tant qu'application dans $\big(E, N\big)$.

4.3 Intégrale d'une application intégrable

4.3.1 Définition de l'intégrale

Proposition et définition 57.
Soient E un espace de Banach sur $\mathbb{K}$ et $f : [a,b] \to E$ une application intégrable.
Pour tout couple de suites $\Big(\{\theta_n\}_{n\in\mathbb{N}}, \{\eta_n\}_{n\in\mathbb{N}}\Big)$ associé à f, la suite $\Big(\int_a^b \theta_n\Big)$ d'éléments de E converge. Sa limite est indépendante du choix du couple de suites associé, on l'appelle intégrale de f sur $[a,b]$ et on la note $\displaystyle\int_a^b f$.

Si f est une application en escalier, on peut choisir, pour tout $n \in \mathbb{N}$, $\varphi_n := f$ et $\psi_n := 0$, l'intégrale définie ci-dessus coïncide donc avec celle définie antérieurement. Si f est une fonction, l'intégrale définie ci-dessus coïncide avec celle définie antérieurement d'après la proposition 28 de la page 569.

Démonstration. Pour montrer que la suite $\Big(\int_a^b \theta_n\Big)$ converge il suffit de montrer qu'elle

est de Cauchy.

Soit $\varepsilon > 0$. On a $\lim\limits_{n\to+\infty} \int_a^b \eta_n = 0$, il existe donc $N \in \mathbb{N}$ tel que $\int_a^b \eta_n \leqslant \varepsilon/2$ si $n \geqslant N$.

Si $m, n \in \mathbb{N}$:

$$\|\theta_m - \theta_n\| \leqslant \|\theta_m - f\| + \|f - \theta_n\| \leqslant \eta_m + \eta_n.$$

Donc, si $m, n \geqslant N$:

$$\left\| \int_a^b \theta_m - \int_a^b \theta_n \right\| \leqslant \int_a^b \|\theta_m - \theta_n\| \leqslant \int_a^b (\eta_m + \eta_n) = \int_a^b \eta_m + \int_a^b \eta_n \leqslant \varepsilon.$$

Montrons que la limite ne dépend pas du choix du couple de suites associé. Soit $\left(\{\theta_n^1\}_{n\in\mathbb{N}}, \{\eta_n^1\}_{n\in\mathbb{N}}\right)$ un autre couple de suites associé. On a :

$$\|\theta_n - \theta_n^1\| \leqslant \|\theta_n - f\| + \|f - \theta_n^1\| \leqslant \eta_n + \eta_n^1.$$

On en déduit que $\lim\limits_{n\to\infty} \int_a^b \|\theta_n - \theta_n^1\| = 0$ et donc a fortiori $\lim\limits_{n\to\infty} \left\| \int_a^b \theta_n - \int_a^b \theta_n^1 \right\| = 0$.

Les deux suites convergentes $\left(\int_a^b \theta_n\right)$ et $\left(\int_a^b \theta_n^1\right)$ ont donc la même limite. ∎

4.3.2 Propriétés de l'intégrale

Les preuves des résultats suivants sont très simples et laissées au lecteur.

Proposition 58. Soit E un $\mathbb{K}$-espace de Banach, l'ensemble $\mathscr{R}\big([a,b], E\big)$ des fonctions intégrables au sens de Riemann sur $[a,b]$ est un sous-espace vectoriel de $\mathcal{F}\big([a,b], E\big)$. L'application $f \in \mathscr{R}\big([a,b], E\big) \mapsto \int_a^b f \in E$ est $\mathbb{K}$-linéaire.

Proposition 59. Soient E et F deux $\mathbb{K}$-espaces de Banach et $u : E \to F$ une application linéaire continue. Si $f : [a,b] \to E$ est une application intégrable au sens de Riemann, $u \circ f$ est une application intégrable au sens de Riemann de $[a,b]$ dans F et l'on a $\int_a^b u \circ f = u\Big(\int_a^b f\Big)$.

Proposition 60.
Soient E, F et G trois $\mathbb{K}$-espaces de Banach et $\Phi : E \times F \to G$ une application bilinéaire continue. Si $f : [a,b] \to E$ et $g : [a,b] \to F$ sont deux applications intégrables au sens de Riemann sur $[a,b]$, $\Phi \circ (f,g) : t \in [a,b] \mapsto \Phi\big(f(t), g(t)\big) \in G$ est une application intégrable au sens de Riemann sur $[a,b]$.

Corollaire 61.
Soient E un $\mathbb{K}$-espace de Banach, $f : [a,b] \to \mathbb{R}$ et $g : [a,b] \to E$ une fonction et une application intégrables au sens de Riemann sur $[a,b]$. Alors $fg : [a,b] \to E$ est intégrable au sens de Riemann sur $[a,b]$.

Proposition 62.
Soient E un $\mathbb{K}$-espace de Banach *de dimension finie*, $f : [a,b] \to \mathbb{R}$ et $(e_1, \dots, e_m)$ une base de E. On note $f_1, \dots, f_m$ les fonctions coordonnées de f sur cette base. Alors f est $\mathscr{R}$-intégrable si, et seulement si, les f_i, $i = 1, \dots, m$, sont $\mathscr{R}$-intégrables.

On obtient ainsi l'équivalence entre les deux définitions de la $\mathscr{R}$-intégrabilité données respectivement dans les parties 3 et 4 (*cf.* la définition 8 de la page 580 et la proposition-définition 27 de la page 568).

Proposition 63. Soient E un espace de Banach sur $\mathbb{K}$ et $f : [a,b] \to E$ une application $\mathscr{R}$-intégrable sur $[a,b]$. La fonction $\|f\| : [a,b] \to \mathbb{R}$ est $\mathscr{R}$-integrable sur $[a,b]$ et l'on a :

$$\left\| \int_a^b f \right\| \leqslant \int_a^b \|f\| \leqslant (b-a) \sup_{x \in [a,b]} \big\| f(x) \big\|. \tag{16}$$

Démonstration. Soit (θ_n, η_n) un couple de suites associé à f. On a :

$$\forall n \in \mathbb{N}, \quad \|f - \theta_n\| \leqslant \eta_n \quad \text{et} \quad \lim_{n \to +\infty} \int_a^b \eta_n = 0.$$

Pour tout $n \in \mathbb{N}$ l'on a $\big| \|f\| - \|\theta_n\| \big| \leqslant \eta_n$, donc :

$$\forall n \in \mathbb{N}, \quad \big| \|f\| - \|\theta_n\| \big| \leqslant \eta_n \quad \text{et} \quad \lim_{n \to +\infty} \int_a^b \eta_n = 0.$$

Les applications $\|\theta_n\|$ et η_n sont en escalier, donc $\big(\|\theta_n\|, \eta_n\big)$ est un couple de suites associé à la fonction $\|f\|$, ainsi cette application est $\mathscr{R}$-intégrable.

L'inégalité $\left\| \int_a^b f \right\| \leqslant \int_a^b \|f\|$ est vraie si f est une application en escalier. Elle le reste pour f intégrable par passage à la limite en utilisant la définition de l'intégrale :

$$\int_a^b f = \lim_{n \to +\infty} \int_a^b \theta_n.$$

Exercice 10.
Soit $I := [a,b]$ un segment de $\mathbb{R}$.

i) Soit $\varphi : I \mapsto \mathbb{C}$ une fonction en escalier sur I, montrer : $\lim\limits_{x \to +\infty} \int_a^b e^{ixt}\, \varphi(t)\, \mathrm{d}t = 0$.

ii) Soit $f : I \mapsto \mathbb{C}$ une fonction intégrable sur I, montrer : $\lim\limits_{x \to +\infty} \int_a^b e^{ixt}\, f(t)\, \mathrm{d}t = 0$.

Solution. On peut supposer $x > 0$.

i) Soit $\mathbf{u} := (a_k)_{k\in[\![0,n]\!]}$ une subdivision de I adaptée à φ, pour tout $k \in [\![0, n-1]\!]$, on note c_k la valeur de φ sur $]a_k, a_{k+1}[$.
On a :

$$\int_a^b e^{ixt}\,\varphi(t)\,\mathrm{d}t = \sum_{k=0}^{n-1} c_k \int_{a_k}^{a_{k+1}} e^{ixt}\,\mathrm{d}t = \frac{1}{ix}\sum_{k=0}^{n-1} c_k(e^{ixa_{k+1}} - e^{ixa_k}).$$

Pour tout $k \in [\![0, n-1]\!]$, on a $|e^{ixa_{k+1}} - e^{ixa_k}| \leqslant 2$, par suite :

$$\lim_{x\to+\infty} \frac{1}{ix}\, c_k(e^{ixa_{k+1}} - e^{ixa_k}) = 0 \quad \text{et} \quad \lim_{x\to+\infty} \int_a^b e^{ixt}\,\varphi(t)\,\mathrm{d}t = 0.$$

ii) Supposons f intégrable sur I. D'après la proposition 62 et la proposition-définition 27, pour tout $\varepsilon > 0$, il existe une application en escalier $\varphi_\varepsilon : I \mapsto \mathbb{C}$ et une fonction en escalier $\psi_\varepsilon : I \mapsto \mathbb{R}$ telles que :

$$|f - \varphi_\varepsilon| \leqslant \psi_\varepsilon \text{ et } \int_a^b \psi_\varepsilon \leqslant \varepsilon.$$

On écrit $\displaystyle\int_a^b e^{ixt} f = \int_a^b e^{ixt}\big(f - \varphi_\varepsilon\big) + \int_a^b e^{ixt}\varphi_\varepsilon$.

Pour tout $x > 0$, $\displaystyle\left|\int_a^b e^{ixt}\big(f - \varphi_\varepsilon\big)\right| \leqslant \int_a^b |f - \varphi_\varepsilon| \leqslant \varepsilon$ et $\displaystyle\lim_{x\to+\infty}\int_a^b e^{ixt}\,\varphi_\varepsilon = 0$, d'où le résultat.

4.3.3 Formule de Chasles

On étend aux applications à valeurs dans E la proposition 16 de la page 560.

Proposition 64 (Formule de Chasles). Soient $f : [a,b] \to E$ et $c \in]a,b[$. Pour que f soit $\mathscr{R}$-intégrable sur $[a,b]$ il faut et il suffit que ses restrictions à $[a,c]$ et $[c,b]$ soient $\mathscr{R}$-intégrables respectivement sur $[a,c]$ et $[c,b]$. Alors, en notant f ces restrictions on a :

$$\int_a^b f = \int_a^c f + \int_c^b f. \tag{17}$$

Démonstration.

1. Supposons f $\mathscr{R}$-intégrable sur $[a,b]$. Soit $\Big(\{\theta_n\}_{n\in\mathbb{N}}, \{\eta_n\}_{n\in\mathbb{N}}\Big)$ un couple de suites associé à f. La formule de Chasles est vraie pour les fonctions de $\mathcal{E}\big([a,b]\big)$ donc $\displaystyle\int_a^b \eta_n = \int_a^c \eta_n + \int_c^b \eta_n$.
On a $\eta_n \geqslant 0$, donc $\displaystyle\int_a^c \eta_n \leqslant \int_a^b \eta_n$ et $\displaystyle\int_c^b \eta_n \leqslant \int_a^b \eta_n$. On en déduit :

$$\lim_{n\to+\infty}\int_a^c \eta_n = \lim_{n\to+\infty}\int_c^b \eta_n = 0.$$

Par ailleurs, pour tout $x \in [a,c]$ (resp $x \in [c,b]$), on a $\big\|f(x)-\theta_n(x)\big\| \leqslant \eta_n(x)$. Ainsi les restrictions de θ_n et η_n à $[a,c]$ (resp. $[c,b]$) déterminent un couple de suites associé à $f_{|_{[a,c]}}$ (resp. $f_{|_{[c,b]}}$). Ainsi $f_{|_{[a,c]}}$ (resp. $f_{|_{[c,b]}}$) est intégrable sur $[a,c]$ (resp. $[c,a]$) et, la formule de Chasles étant vraie pour les fonctions de $\mathcal{E}\big([a,b],E\big)$, l'on a :

$$\int_a^b f = \lim_{n\to+\infty}\int_a^b \theta_n = \lim_{n\to+\infty}\int_a^c \theta_n + \lim_{n\to+\infty}\int_c^b \theta_n = \int_a^c f + \int_c^b f.$$

2. Inversement supposons les restrictions $f_{|_{[a,c]}}$ et $f_{|_{[c,b]}}$ intégrables respectivement sur les segments $[a,c]$ et $[c,b]$. On peut alors leur associer respectivement des couples de suites $\Big(\{\theta_n^1\}_{n\in\mathbb{N}},\{\eta_n^1\}_{n\in\mathbb{N}}\Big)$ et $\Big(\{\theta_n^2\}_{n\in\mathbb{N}},\{\eta_n^2\}_{n\in\mathbb{N}}\Big)$. On définit des fonctions $(\theta_n,\eta_n)_{n\in\mathbb{N}}$ sur $[a,b]$:

$$\theta_n : x \to \begin{cases}\theta_n^1 & \text{si} \quad x\in[a,c]\\ \theta_n^2 & \text{si} \quad x\in\,]c,b]\end{cases} \quad \text{et} \quad \eta_n : x \to \begin{cases}\eta_n^1 & \text{si} \quad x\in[a,c]\\ \eta_n^2 & \text{si} \quad x\in\,]c,b].\end{cases}$$

Ces fonctions sont en escalier. La formule de Chasles étant vraie pour les fonctions de $\mathcal{E}\big([a,b]\big)$, on a donc, pour tout $n\in\mathbb{N}$:

$$\int_a^b \eta_n = \int_a^c \eta_n + \int_c^b \eta_n = \int_a^c \eta_n^1 + \int_c^b \eta_n^2,$$

puisque $(\eta_n)_{|_{[c,b]}}$ et η_n^2 ne différent (éventuellement) qu'au point c.

Ainsi l'on a $\|f-\theta_n\| \leqslant \eta_n$ et :

$$\lim_{n\to+\infty}\int_a^b \eta_n = \lim_{n\to+\infty}\int_a^b \eta_n^1 + \lim_{n\to+\infty}\int_a^b \eta_n^2 = 0.$$

Il s'en suit que $\Big(\{\theta_n\}_{n\in\mathbb{N}},\{\eta_n\}_{n\in\mathbb{N}}\Big)$ est un couple de suites associé à f et que f est $\mathscr{R}$-intégrable sur $[a,b]$.

3. Montrons l'égalité (17).

La formule de Chasles est vraie pour les fonctions de $\mathcal{E}\big([a,b],E\big)$, on a donc :

$$\int_a^b f = \lim_{n\to+\infty}\int_a^b \theta_n = \lim_{n\to+\infty}\int_a^c \theta_n + \lim_{n\to+\infty}\int_c^b \theta_n = \int_a^c f + \int_c^b f. \qquad \blacksquare$$

Si $b \leqslant a$, on généralise la définition de l'intégrale $\displaystyle\int_a^b f$ d'une application f à valeurs dans E en utilisant des conventions similaires à celles adoptées pour les fonctions (*cf.* page 561). Comme dans le cas des fonctions on peut alors étendre la formule de Chasles

Proposition 65. Soient $a,b,c \in \mathbb{R}$ et $f \in \mathscr{R}\big([\min(a,b,c),\max(a,b,c)],E\big)$. Alors f est $\mathscr{R}$-intégrable sur chacun des intervalles compacts déterminés par deux quelconques des points a,b,c, et l'on a :

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

4.4 Sommes de Riemann d'une application à valeurs dans un espace de Banach

Proposition 66. Soient $(E, \|.\|)$ un $\mathbb{K}$-espace de Banach et $f : [a,b] \to E$ une fonction intégrable au sens de Riemann sur $[a,b]$.
Alors, pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour toute subdivision pointée $\mathcal{D}$ de $[a,b]$ de pas inférieur à δ, l'on ait $\left\| S_{\mathcal{D}}(f) - \displaystyle\int_a^b f \right\| \leqslant \varepsilon$.

Démonstration. Nous avons prouvé ce résultat dans le cas où E est *de dimension finie* (*cf.* la proposition 51 de la page 582). Nous l'admettrons dans le cas général, on trouvera par exemple une preuve dans [R3], 6.2.3 (page 204) dont l'argument est voisin de celui de la preuve du lemme 30 de la page 571. ∎

Quand E est *de dimension infinie*, les sommes de Riemann peuvent « converger » sans que la fonction soit intégrable : *cf.* l'exercice II.3.14.

5 Applications réglées

5.1 Définition et premières propriétés

Définition 9. Soit E un espace de Banach réel.
On dit qu'une application *bornée* $f : [a,b] \to E$ est *réglée* si elle peut être approchée uniformément, c'est-à-dire au sens de la norme $\|.\|_\infty$ sur $\mathcal{B}([a,b], E)$, par des applications en escalier à valeurs dans E.

Autrement, dit f est réglée si, et seulement si, pour tout $\varepsilon > 0$, il existe une fonction en escalier $g \in \mathcal{E}([a,b], E)$ telle que :

$$\|g - f\|_\infty := \sup_{x \in [a,b]} \|g(x) - f(x)\| \leqslant \varepsilon.$$

Une application *bornée* $f : [a,b] \to E$ est *réglée* si, et seulement si, elle est la limite uniforme sur $[a,b]$ d'une *suite* d'applications en escalier à valeurs dans E.
Si $E = \mathbb{R}$ on parle de *fonction réglée*. En première lecture on pourra se limiter à ce cas pour la suite de ce paragraphe. C'est par exemple suffisant pour les applications aux séries de Fourier.
On note $\mathcal{R}([a,b], E)$ l'ensemble des applications réglées de $[a,b]$ dans E (ne pas confondre $\mathcal{R}$ et $\mathscr{R}$). C'est *l'adhérence* de $\mathcal{E}([a,b], E)$ dans l'espace de Banach $\mathcal{B}([a,b], E)$.
On vérifie que $\mathcal{R}([a,b], E)$ est un sous-espace vectoriel fermé de $\mathcal{B}([a,b], E)$.
On prouve facilement les résultats suivants.

Proposition 67. Soit $(f_n)_{n \in \mathbb{N}}$ une suite de $\mathcal{R}([a,b], E)$ convergeant *uniformément* sur $[a,b]$ vers f. Alors $f \in \mathcal{R}([a,b], E)$.

Proposition 68. Soit $f \in \mathcal{R}\big([a,b],E\big)$.
Alors $\|f\| \in \mathcal{R}\big([a,b]\big)$. Si de plus $E=\mathbb{R}$, alors f^+ et f^- sont dans $\mathcal{R}\big([a,b]\big)$.

La définition et la proposition ci-dessous vont nous permettre de donner une caractérisation des applications réglées. Par ailleurs elles joueront un rôle important dans la théorie de l'intégrale de Henstock-Kurzweil que nous présenterons dans [L2-L3].

Définition 10. Une *jauge* sur un segment $[a,b]$ est une fonction strictement positive $\delta : [a,b] \to]0,+\infty[$. Une subdivision pointée $\mathcal{D}$ de $[a,b]$ est δ-*fine* si :

$$\forall (J,\xi) \in \mathcal{D}, \quad \operatorname{long}(J) \leqslant \delta(\xi).$$

L'appellation « jauge » pour une fonction à valeurs strictement positives peut paraître inutile. Le concept vraiment intéressant est celui de subdivision pointée δ-fine (l'idée est que la jauge « contrôle » la longueur des intervalles de la subdivision). Historiquement ces notions ont été introduites dans le cadre de l'intégrale de Henstock-Kurzweil dite aussi *intégrale de jauge*. Ici ces notions ne seront utilisées que dans la preuve de la proposition 70 et l'on pourrait s'en passer à un changement de vocabulaire près.

> ***Remarque***. Pour simplifier, nous parlerons de « subdivision δ-fine » en omettant le terme « pointée ».

Le résultat qui suit est une variante du théorème de Borel-Lebesgue prouvée (à une modification de vocabulaire près…) par Pierre Cousin en 1895 (antérieurement aux travaux de Borel et Lebesgue) :

Proposition 69 (Lemme de Cousin).
Si δ est une jauge sur $[a,b]$, alors il existe une subdivision de $[a,b]$ qui est δ-fine.

Démonstration. La démonstration repose sur le théorème des segments emboîtés.
On procède par dichotomie. Supposons par l'absurde que le segment $J_0 := [a,b]$ n'admette pas de subdivision δ-fine. Si les deux segments $\big[a,(a+b)/2\big]$ et $\big[(a+b)/2,b\big]$ admettaient une subdivision δ-fine, la réunion de ces subdivisions serait une subdivision δ-fine de $[a,b]$. Par conséquent, au moins l'un des deux segments n'admet pas de subdivision δ-fine. Notons $J_1 \subset J_0$ un tel segment et observons que :

$$\operatorname{long}(J_1) = \frac{1}{2}\operatorname{long}(J_0).$$

En réitérant cet argument, on construit une suite (J_k) de segments emboîtés, telle que J_k n'admette pas de subdivision δ-fine et :

$$\operatorname{long}(J_k) = \frac{1}{2^k}\operatorname{long}(J_0) \underset{k\to+\infty}{\longrightarrow} 0.$$

D'après le théorème des segments emboîtés, l'intersection des J_k contient un point ξ. Si k est assez grand, $\operatorname{long}(J_k) \leqslant \delta(\xi)$ et $\big\{(J_k,\xi)\big\}$ est une subdivision δ-fine de J_k. D'où une contradiction. ∎

Proposition 70. Soit E un espace de Banach. Une application $f : [a,b] \to E$ est *réglée* si, et seulement si, elle admet une limite à droite en tout point de $[a,b[$ et une limite à gauche en tout point de $]a,b]$.
Toute application réglée est *continue* en dehors d'une partie au plus dénombrable de $[a,b]$.

Démonstration.

1. Soit $f : [a,b] \to E$ une application réglée. Il existe une suite $\big(g_n \in \mathcal{E}([a,b],E)\big)_{n\in\mathbb{N}^*}$ convergeant vers f pour la norme uniforme de $\mathcal{B}([a,b],E)$.
   Soit $c \in [a,b[$. Pour tout $n \in \mathbb{N}^*$ la fonction en escalier $t \mapsto g_n(t)$ admet une limite à droite α_n en c. L'espace E étant *complet*, d'après le théorème de la double limite (*cf.* le module *Suites et séries de fonctions*, théorème 12 de la page 482), la suite $(\alpha_n)_{n\in\mathbb{N}^*}$ de E a une limite α et l'application f admet α pour limite à droite en c.
   On peut raisonner de façon analogue pour les limites à gauche aux points de $]a,b]$.
   Si l'on note S_n l'ensemble *fini* des points de discontinuité de g_n, l'ensemble $S := \bigcup_{n\in\mathbb{N}} S_n$ est au plus dénombrable. Soit $c \in]a,b[\setminus S$, pour tout $n \in \mathbb{N}$ les limites à droite et à gauche de g_n sont égales à $g_n(c)$. On a donc aussi $\lim\limits_{c^+} f = \lim\limits_{c^-} f = f(c)$.
   Ainsi f est *continue* en dehors d'une partie au plus dénombrable de $[a,b]$.
2. Soit $f : [a,b] \to E$ une application. On suppose qu'elle admet une limite à droite en tout point de $[a,b[$ et une limite à gauche en tout point de $]a,b]$.
   Nous allons montrer, en utilisant la proposition 69, que f est limite uniforme d'une suite $(h_n)_{n\in\mathbb{N}^*}$ de fonctions en escalier.
   Notons, pour tout $x \in [a,b[$ (resp. $x \in]a,b]$), $f(x^+)$ (resp. $f(x^-)$) la limite à droite (resp. à gauche) de f en x.
   Soit $n \in \mathbb{N}^*$. D'après la définition des limites à droite et à gauche, pour tout $x \in [a,b[$ il existe des réels strictement positifs $\alpha_n(x)$ et $\beta_n(x)$ tel que :
   – $\forall x \in [a,b[\,,\ x+\alpha_n(x) < b$ et $\forall y \in]x, x+\alpha_n(x)]\,,\ \big\|f(y)-f(x^+)\big\| \leqslant 1/n$;
   – $\forall x \in]a,b]\,,\ a < x-\beta_n(x)$ et $y \in [x-\beta_n(x), x[\,,\ \big\|f(y)-f(x^-)\big\| \leqslant 1/n$.
   En posant $\delta_n := \inf\{\alpha_n, \beta_n\}$, on en déduit qu'il existe une *jauge* δ_n sur $[a,b]$ telle que :
   – $\forall x \in [a,b[\,,\ x+\delta_n(x) < b$ et $y \in]x, x+\delta_n(x)]\,,\ \big\|f(y)-f(x^+)\big\| \leqslant 1/n$;
   – $\forall x \in]a,b]\,,\ a < x-\delta_n(x)$ et $y \in [x-\delta_n(x), x[\,,\ \big\|f(y)-f(x^-)\big\| \leqslant 1/n$.
   D'après la proposition 69 de la page précédente, on peut choisir une subdivision pointée δ_n-fine $\mathcal{D} := \big\{[a_k, a_{k+1}], x_k\big\}_{k\in[\![0,N-1]\!]}$ de $[a,b]$. On définit une application $h_n : [a,b] \to E$ par :

$$\forall k \in [\![0,N]\!], \quad h_n(a_k) := f(a_k)$$
$$\forall k \in [\![0,N-1]\!], \quad h_n(x_k) := f(x_k)$$
$$\forall k \in [\![0,N-1]\!],\ \forall x \in]a_k, x_k[, \quad h_n(x) := f(x_k^-)$$

$$\forall k \in [\![0, N-1]\!],\ \forall x \in]x_k, a_{k+1}[, \quad h_n(x) := f(x_k^+)$$

On vérifie que h_n est une application en escalier associée à la subdivision

$$(a_0, x_0, a_1, x_1, \ldots, a_{k-1}, x_{k-1}, a_k, \ldots, a_{N-1}, x_{N-1}, a_N)$$

et que, pour tout $x \in [a, b]$, on a $\big\|f(x) - h_n(x)\big\| \leqslant 1/n$. ■

Corollaire 71. Toute application *continue* d'un segment $[a, b]$ dans un espace de Banach est réglée.

Corollaire 72. Toute fonction réelle *monotone*, ou plus généralement *à variation bornée*, sur un segment est réglée.

Dans les exercices ci-dessous on donne deux exemples de fonctions intégrables au sens de Riemann qui ne sont pas réglées.

Exercice 11.

Montrer que la fonction $f : [-1, 1] \to \mathbb{R}$ définie par :

$$f(x) := \begin{cases} \sin(1/x) & \text{si} \quad x \neq 0 \\ 0 & \text{si} \quad x = 0. \end{cases}$$

n'est pas réglée.

Solution. La fonction f n'admet pas de limite à droite en 0. Elle est intégrable au sens de Riemann d'après l'exercice 21 de la page 562.

Exercice 12.

Soit 1_K la fonction indicatrice de l'ensemble de Cantor sur $[0, 1]$ (*cf.* la partie 1.3.7). Montrer qu'elle n'est pas réglée.

Solution. La fonction 1_K n'est pas réglée car elle admet un ensemble *non dénombrable* de points de discontinuité. Nous avons vu qu'elle est intégrable au sens de Riemann.

Exercice 13.

Soit f la fonction définie sur $[0, 1]$ par :

$$f(x) := \begin{cases} 0 & \text{si} \quad x \notin \mathbb{Q} \\ \frac{1}{q} & \text{si} \quad x = \frac{p}{q} \quad \text{(rationnel irréductible)}. \end{cases}$$

1. Montrer que f est une fonction réglée sur $[0, 1]$.
   Quels sont ses points de discontinuité ?
2. Déterminer une suite de fonctions en escalier sur $[0, 1]$ convergeant uniformément vers f sur $[0, 1]$.
3. Calculer $\displaystyle\int_0^1 f$.

Solution.

1. Soit $n \in \mathbb{N}^*$, considérons $A_n := \{x \in [0,1] \mid f(x) > \frac{1}{n}\}$.
   On vérifie que $A_n \subset \left\{x = \frac{p}{q} \in [0,1] \cap \mathbb{Q} \mid 1 \leqslant q < n\right\}$, A_n est donc un ensemble *fini*.
   Soit $a \in [0,1]$. Pour tout $n \in \mathbb{N}^*$, il existe $\alpha > 0$ tel que :
   $$A_n \cap]a-\alpha, a[= \varnothing \qquad \text{et} \qquad A_n \cap]a, a+\alpha[= \varnothing.$$
   Ainsi, pour tout $x \in]a-\alpha, a[\cap [0,1]$ (resp. $]a, a+\alpha[\cap [0,1]$), on a $f(x) \leqslant \frac{1}{n}$.
   On en déduit $\lim\limits_{x \to a, x \neq a} f(x) = 0$. La fonction f est donc réglée, elle est continue en tout point irrationnel et discontinue en tout point rationnel.
2. Soit $(f_n)_{n \in \mathbb{N}}$ la suite de fonctions en escalier définies sur $[0,1]$ par :
   $$f_n(x) := \begin{cases} 0 & \text{si} \quad x \notin A_n \\ f(x) & \text{si} \quad x \in A_n. \end{cases}$$
   On vérifie :
   $$\forall x \in [0,1], \quad f_n(x) \leqslant f(x) \leqslant f_n(x) + \frac{1}{n}. \tag{18}$$
   On a $\sup\limits_{x \in [0,1]} \left|f_n(x) - f(x)\right| \leqslant \frac{1}{n}$, donc la suite $(f_n)_{n \in \mathbb{N}}$ converge uniformément vers f sur $[0,1]$.
3. D'après les inégalités (18) et la définition de l'intégrale, on a $\displaystyle\int_0^1 f = 0$. On peut aussi utiliser la proposition 75 ci-dessous.

5.2 Intégrabilité des fonctions réglées

Proposition 73. Toute application *réglée* d'un segment $[a,b]$ dans un espace de Banach est intégrable au sens de Riemann.

Autrement dit $\mathcal{R}\big([a,b], E\big) \subset \mathscr{R}\big([a,b], E\big)$. L'inclusion est stricte d'après l'exercice 12 de la page précédente.

Démonstration. On utilise le critère d'intégrabilité de la proposition 27 de la page 568.
Soient E un espace de Banach et $f : [a,b] \to E$ une application *réglée*. Par définition, pour tout $\varepsilon > 0$ il existe une application en escalier $\Phi : [a,b] \to E$ telle que $\sup\limits_{x \in [a,b]} \big\|f(x) - \Phi(x)\big\| \leqslant \varepsilon/(b-a)$. Soit Ψ la fonction constante $x \mapsto \varepsilon/(b-a)$, c'est une fonction en escalier sur $[a,b]$ et l'on a :
$$\|f - \Phi\| \leqslant \Psi \quad \text{et} \quad \int_a^b \Psi \leqslant \varepsilon.$$
Ainsi f est intégrable au sens de Riemann. ∎
En utilisant le corollaire 71 ci-dessus on obtient le résultat suivant.

Corollaire 74. Toute application *continue* d'un segment $[a,b]$ dans un espace de Banach est intégrable au sens de Riemann.

Une fonction monotone étant réglée, on retrouve le résultat de la proposition 34 de la page 573 : une fonction monotone sur un segment est intégrable au sens de Riemann sur ce segment.

Proposition 75. Pour que l'intégrale d'une fonction réglée (resp. continue) *positive* $f : [a,b] \to \mathbb{R}$ soit nulle il faut et il suffit que l'ensemble $\{x \in [a,b] \mid f(x) \neq 0\}$ soit au plus dénombrable (resp. vide).

Démonstration.

Le résultat pour f continue a été prouvé dans le module *Intégration* de [L1].

Supposons l'ensemble :

$$S := \{x \in [a,b] \mid f(x) \neq 0\} = \{x \in [a,b] \mid f(x) > 0\}$$

au plus dénombrable. Soit φ une fonction en escalier majorée par f, on ne peut avoir $\varphi(x) > 0$ que si $x \in S$. L'ensemble des points d'un intervalle ouvert non vide étant *non dénombrable* la fonction φ est négative sur tous les intervalles ouverts non vides où elle est constante et l'on a $\displaystyle\int_a^b \varphi \leqslant 0$. On en déduit $\displaystyle\underset{*}{\int_a^b} f \leqslant 0$. Par ailleurs, puisque f est positive, l'on a $\displaystyle\overset{*}{\int_a^b} f \geqslant 0$ et, f étant intégrable, $\displaystyle\int_a^b f = \underset{*}{\int_a^b} f = \overset{*}{\int_a^b} f = 0$.

Inversement, supposons que $\displaystyle\int_a^b f = 0$. Il existe une suite $(\varphi_n)_{n\in\mathbb{N}^*}$ de fonctions en escalier telle que, pour tout $n \in \mathbb{N}^*$, l'on ait $\displaystyle\sup_{x\in[a,b]} \big|f(x) - \varphi_n(x)\big| \leqslant 1/n$.

Posons $\psi_n := \varphi_n - 1/n$. On a $\displaystyle\sup_{x\in[a,b]} \big|f(x) - \psi_n(x)\big| \leqslant 2/n$ et $\psi_n \leqslant f$.

Les fonctions ψ_n sont des fonctions en escalier, quitte à les remplacer par les fonctions $\psi_n^+ := \sup(\psi_n, 0)$, qui sont aussi en escalier, on peut supposer que $\psi_n \geqslant 0$. On a alors $0 \leqslant \displaystyle\int_a^b \psi_n \leqslant \int_a^b f = 0$ donc $\displaystyle\int_a^b \psi_n = 0$. Ainsi les $\psi_n \geqslant 0$ sont nulles sauf peut-être sur un ensemble *fini* T_n. L'ensemble $T := \displaystyle\bigcup_{n\in\mathbb{N}^*} T_n$ est au plus *dénombrable* et si $x \in [a,b] \setminus T$, $f(x) = \displaystyle\lim_{n\to+\infty} \psi_n(x) = 0$. Donc $\{x \in [a,b] \mid f(x) \neq 0\} \subset T$ est au plus dénombrable. ∎

On peut généraliser la notion de fonction réglée à un intervalle quelconque en utilisant les notions de limites à droite et à gauche.

Définition 11. Soient $I \subset \mathbb{R}$ un intervalle d'intérieur non vide et E un espace de Banach. Une application $f : I \to E$ est *réglée* si, et seulement si, elle admet une limite à droite en tout point de I et une limite à gauche en tout point de I (avec la convention évidente aux extrémités éventuelles).

Une telle application est réglée sur I si, et seulement si, ses restrictions aux segments de I le sont. On en déduit que si f est réglée elle admet des limites à droite et à gauche en tout point et qu'elle est *continue* en dehors d'une partie au plus dénombrable de I.

5.3 Caractérisation des fonctions intégrables au sens de Riemann

Nous avons vu ci-dessus que l'ensemble des points de discontinuité d'une fonction réglée est au plus dénombrable. On peut, plus généralement s'interroger sur les propriétés de l'ensemble des points de discontinuité d'une fonction intégrable au sens de Riemann. La réponse à cette question a conduit Lebesgue à une une caractérisation « simple » des fonctions intégrables au sens de Riemann que nous énoncerons plus loin. C'est un premier pas vers les notions de mesure et d'intégrale de Lebesgue. Dans ce module nous n'irons pas plus loin dans cette direction et ce paragraphe peut être omis en première lecture.

Nous avons vu (*cf.* l'exemple 12 de la page 593) que l'ensemble des points de discontinuité d'une fonction $\mathscr{R}$-intégrable peut ne pas être dénombrable. Nous allons voir qu'il doit tout de même être « assez petit ». Le premier pas consiste à préciser en quel sens grâce à la notion de *sous-ensemble négligeable* de $\mathbb{R}$ (dont un exemple « non trivial » est l'ensemble de Cantor).

5.3.1 Sous-ensembles négligeables de $\mathbb{R}$

La notion de sous-ensemble négligeable est nécessaire pour formuler le théorème de Lebesgue (*cf.* le théorème 78). Nous reviendrons sur cette notion à la fin du module *Intégration* de [L2-L3] en relation avec l'étude de la mesure de Lebesgue.

Définition 12. Une partie A de $\mathbb{R}$ est dite *négligeable* si pour tout $\varepsilon > 0$, il existe une suite $(I_n)_{n\in\mathbb{N}}$ d'intervalles compacts telle que :
(i) $A \subset \bigcup_{n\in\mathbb{N}} I_n$;
(ii) $\sum_{n\in\mathbb{N}} \text{long } I_n \leqslant \varepsilon$ (c'est à dire que $\sum_{n\in\mathbb{N}} \text{long } I_n$ est convergente et de somme inférieure à ε).

On vérifie facilement que l'on obtient une définition équivalente en remplaçant les intervalles *compacts* par des intervalles *ouverts*.

En utilisant la définition « naïve » de longueur introduite dans la partie 1.3.6 de la page 562 on peut aussi dire qu'une partie A est négligeable si, pour tout $\varepsilon > 0$, il existe un ouvert de longueur inférieure à ε contenant A.

Tout sous-ensemble d'un sous-ensemble négligeable est évidemment négligeable.

Un sous-ensemble réduit à un point est négligeable. Un intervalle n'est pas négligeable sauf s'il est réduit à un point. Un sous-ensemble négligeable est d'intérieur vide.

Proposition 76. Une réunion au plus *dénombrable* de parties négligeables est une partie négligeable.

Démonstration. Le cas d'une réunion finie est trivial. Considérons celui d'une réunion dénombrable. Soit $(A_k)_{k\in\mathbb{N}}$ une suite de parties négligeables. Soit $\varepsilon > 0$. Par hypothèse, pour tout $k \in \mathbb{N}$ il existe une suite d'intervalles compacts $I_{k,n}$ telle que :
(i) $A_k \subset \bigcup_{n\in\mathbb{N}} I_{k,n}$;
(ii) $\sum_{n\in\mathbb{N}} \text{long } I_{k,n} \leqslant \varepsilon/2^{k+1}$.

La famille $(I_{k,n})_{(k,n)\in\mathbb{N}^2}$ est une réunion dénombrable d'ensemble dénombrables, elle est donc *dénombrable*. On la ré-indexe par $\mathbb{N}$ et on la note $(J_m)_{m\in\mathbb{N}}$. On a alors :

(i) $\bigcup_{k\in\mathbb{N}} A_k \subset \bigcup_{m\in\mathbb{N}} J_m$;

(ii) $\sum_{m\in\mathbb{N}} \operatorname{long} J_m \leqslant \sum_{k\in\mathbb{N}} \varepsilon/2^{k+1} = \varepsilon$. ∎

Corollaire 77. Tout sous-ensemble dénombrable est négligeable.

Exemple. On considère $\mathbb{Q} \subset \mathbb{R}$. Soit $\varepsilon > 0$, $\mathbb{Q}$ étant dénombrable, il existe une réunion dénombrable disjointe d'intervalles ouverts de longueur inférieure à ε recouvrant $\mathbb{Q}$. Cette réunion est dense ($\mathbb{Q}$ l'est), mais, si ε est « petit », « elle n'occupe que peu de place ». Nous invitons le lecteur à méditer sur cette situation.

On vérifie que l'ensemble de Cantor est négligeable (il est l'intersection d'une famille décroissante de parties qui sont réunions finies de segments et dont la longueur tend vers 0). Ainsi une partie négligeable n'est pas nécessairement dénombrable.
On dit qu'une propriété portant sur les points d'un intervalle I est vraie *presque partout* si elle vraie sur le complémentaire d'un ensemble négligeable contenu dans I.

5.3.2 Le théorème de Lebesgue

Théorème 78. Soient I un segment et $f : I \to \mathbb{R}$. Les deux conditions suivantes sont équivalentes.
(i) La fonction f est intégrable au sens de Riemann sur I.
(ii) La fonction f est bornée sur I et son ensemble de points de discontinuité est négligeable.

Autrement dit une fonction est intégrable au sens de Riemann sur I si, et seulement si, elle est bornée et presque partout continue sur I.

Démonstration. Nous allons montrer l'implication (i) $\Rightarrow$ (ii). Nous admettrons la réciproque, nous la prouverons dans le module *Intégration* de [L2-L3] (on pourra aussi trouver une preuve dans [R3], 6.5.3, page 224).

Une fonction intégrable au sens de Riemann est bornée, il reste à montrer que l'ensemble de ses points de discontinuité est négligeable.

Un point $y \in I$ est un point de discontinuité de f si, et seulement si, la condition suivante est satisfaite :

$$\exists\, h > 0,\ \forall \eta > 0,\ \exists\, x \in I, |x - y| < \eta \quad \text{et} \quad \big|f(x) - f(y)\big| > h.$$

Notons, pour tout $h > 0$:

$$A_h := \{y \in I \,\big|\, \forall \eta > 0,\ \exists\, x \in I \quad \text{tel que} \quad |x - y| < \eta \quad \text{et} \quad \big|f(x) - f(y)\big| > h\}. \tag{19}$$

Si $h' \leqslant h$, $A_h \subset A_{h'}$. L'ensemble des points de discontinuité de f est $A := \bigcup_{h>0} A_h = \bigcup_{n\in\mathbb{N}^*} A_{1/n}$.

La deuxième réunion est *dénombrable* ; pour montrer que A est négligeable, il suffit donc de prouver que, pour tout $h > 0$, A_h est négligeable.

Fixons $h > 0$. Soit $\varepsilon > 0$. Nous allons montrer que A_h est contenu dans une réunion finie d'intervalles dont la somme des longueurs est majorée par ε.

Puisque f est intégrable, le critère de Darboux est satisfait (*cf.* le corollaire 33 de la page 573). Il existe donc une subdivision $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ de I telle que

$$D_{\mathbf{u}}(f) - d_{\mathbf{u}}(f) = \sum_{i=0}^{n-1} (M_i - m_i)(x_{i+1} - x_i) < \varepsilon h/2.$$

Notons $J \subset [\![0, n-1]\!]$ le sous-ensemble des indices i tels que $]x_i, x_{i+1}[\cap A_h \neq \varnothing$. On a :

$$A_h \setminus \{x_0, \dots, x_n\} = \bigcup_{i\in[\![0,n-1]\!]} \big(]x_i, x_{i+1}[\cap A_h\big) \subset \bigcup_{i\in J}]x_i, x_{i+1}[.$$

Par ailleurs, si $i \in J$, il existe $y \in]x_i, x_{i+1}[\cap A_h$, donc d'après (19), il existe $z \in]x_i, x_{i+1}[$ tel que $\big|f(z) - f(y)\big| > h$. Par suite : $M_i - m_1 \geqslant h$, donc :

$$h \sum_{i\in J} (x_{i+1} - x_i) \leqslant \sum_{i\in J} (M_i - m_i)(x_{i+1} - x_i) < \varepsilon h/2.$$

$$\sum_{i\in J} (x_{i+1} - x_i) < \varepsilon/2.$$

Ainsi $A_h \setminus \{x_0, \dots, x_n\}$ est contenu dans une réunion finie d'intervalles dont la somme des longueurs est majorée par $\varepsilon/2$. Par ailleurs, il est clair que l'ensemble *fini* $\{x_0, \dots, x_n\}$ est aussi contenu dans une réunion finie d'intervalles dont la somme des longueurs est majorée par $\varepsilon/2$, donc A_h est contenu dans une réunion finie d'intervalles dont la somme des longueurs est majorée par ε. ∎

Exercice 14.
Soit $K \subset [0,1]$ un compact. Donner une condition nécessaire et suffisante portant sur la frontière ∂K de K pour que la fonction caractéristique 1_K soit intégrable au sens de Riemann sur $[0,1]$.

Solution. On a une partition : $[0,1] = \operatorname{Int} K \cup \big([0,1] \setminus K\big) \cup \partial K$. La fonction 1_K est constante et égale à 1 (resp. 0) sur l'ouvert $\operatorname{Int} K$ (resp. $\operatorname{Int}\big([0,1] \setminus K\big)$), donc elle est continue sur $K \cup \operatorname{Int}\big([0,1] \setminus K\big)$. Par ailleurs, si $a \in \partial K$, tout voisinage de a rencontre $\operatorname{Int} K$ et $\big([0,1] \setminus K\big)$ donc la fonction 1_K n'est pas continue en a. Ainsi ∂K est l'ensemble des points de discontinuité de 1_K. D'après le théorème de Lebesgue 1_K est intégrable au sens de Riemann sur $[0,1]$ si, et seulement si, ∂K est négligeable.

Remarque. Il existe des compacts de $[0,1]$ qui ne sont pas négligeables mais qui sont d'intérieur vide (on peut en construire en utilisant une variante de la construction de la construction de l'ensemble de Cantor, on obtient ainsi par exemple l'ensemble de Smith-Volterra-Cantor, *cf.* l'exercice II.3.6).
Pour un tel compact K on a $\partial K = K$, donc ∂K n'est pas négligeable. D'après le résultat de l'exercice ci-dessus il en résulte que 1_K n'est pas intégrable au sens de Riemann sur $[0,1]$.

6 Propriétés de l'intégrale fonction de sa borne supérieure

Dans cette section, I désigne un intervalle réel non vide et E un espace de Banach sur $\mathbb{K}$. En première lecture, on pourra se limiter au cas des fonctions en remplaçant E par $\mathbb{R}$ dans les énoncés.

6.1 Compléments sur la dérivation

Nous allons voir qu'en un certain sens l'opération d'intégration est l'inverse de celle de dérivation. Dans le cas des fonctions à valeurs dans un espace de Banach, nous aurons besoin de quelques compléments sur la dérivation.

Dans le module *Introduction aux fonctions vectorielles d'une variable réelle* de [L1], nous avons introduit la notion de dérivée d'une application $f : I \subset \mathbb{R} \to \mathbb{R}^n$. Nous allons la généraliser au cas des fonctions à valeurs dans un espace vectoriel normé réel E (pour le cas où E est un $\mathbb{C}$-espace vectoriel on utilise la structure réelle sous-jacente). Nous énoncerons, sans détailler les démonstrations, un certain nombre de résultats dont les preuves sont similaires, *mutatis mutandis*, à celles du cas où $E = \mathbb{R}$ (*cf.* le module *Fonctions réelles, dérivabilité* de [L1]).

Définition 13. Soient I un intervalle réel, E un espace vectoriel normé réel, $a \in I$ et $f : I \to E$ une application.

(i) On dit que f est dérivable au point a si l'application $F : x \in I \setminus \{a\} \mapsto \frac{f(x)-f(a)}{x-a}$ admet une limite quand x tend vers a. S'il en est ainsi, on note $f'(a) \in E$ cette limite et l'on dit que c'est la dérivée de f en a.

(ii) Si $I' := I \cap [a, +\infty[\neq \{a\}$ et si la restriction g de f à I' est dérivable en a, on dit que f est dérivable *à droite* en a et que $f'_d(a) := g'(a)$ est sa dérivée à droite en a. On définit de même la dérivablilité *à gauche* en remplaçant I' par $I \cap]-\infty, a]$. On note $f'_g(a)$ la dérivée à gauche en a.

Si $a \in I$ est un point intérieur, f est dérivable en a si, et seulement si, elle est dérivable à droite et à gauche en a et si $f'_d(a) = f'_g(a)$ et l'on a alors $f'(a) = f'_d(a) = f'_g(a)$.

Une application $f : I \to E$ est dite dérivable si elle l'est en tout point de I, $x \in I \mapsto f'(x)$ est la dérivée de f. Si f' est continue, on dit que f est de classe $\mathcal{C}^1$.

Proposition 79. Soient I un intervalle réel, E un espace vectoriel normé réel, $a \in I$ et $f : I \to E$ une application. Si f est dérivable (resp. dérivable à droite, resp. dérivable à gauche) en a, alors elle est continue (resp. continue à droite, resp. continue à gauche) en a,

Proposition 80. Soient I un intervalle réel, $E := E_1 \times \cdots \times E_n$ un produit d'espaces vectoriels normés réels et $f := (f_1, \ldots, f_n) : I \to E$ une application. Alors f est dérivable (resp. dérivable à droite, resp. dérivable à gauche) en $a \in I$ si, et seulement si, les f_i, $i = 1, \ldots, n$, le sont. On a alors $f'(a) = \big(f'_1(a), \ldots, f'_n(a)\big)$.

Proposition 81. Soient I un intervalle réel, E un espace vectoriel normé réel, $a \in I$ et $f, g : I \to E$ deux applications. Si f et g sont dérivables (resp. dérivables à droite, resp. dérivables à gauche) en a, alors, pour tous λ, $\mu \in \mathbb{R}$, l'application $\lambda f + \mu g$ est dérivable (resp. dérivable à droite, resp. dérivable à gauche) en a et l'on a $(\lambda f + \mu g)'(a) = \lambda f'(a) + \mu g'(a)$.

Proposition 82.
Soient I un intervalle réel, $E_1, \dots, E_n, F$ des espaces vectoriels normés réels, $u : E_1 \times \cdot \times E_n \to F$ une application n-linéaire *continue* et $f := (f_1, \dots, f_n) : I \to E$. On note $g := u \circ f : I \to F$ l'application composée. Si chaque f_i $(i = 1, \dots, n)$ est dérivable en a, alors g est dérivable en a et l'on a :

$$g'(a) = \sum_{i=1}^{n} u\big(f_1(a), \dots, f_{i-1}(a), f_i'(a), f_{i+1}(a), \dots, f_n(a)\big).$$

Démonstration. En utilisant la n linéarité de u, on a, pour tout $x \in I \setminus \{a\}$,

$$\frac{g(x) - g(a)}{x - a} = \sum_{i=1}^{n} u\Big(f_1(a), \dots, f_{i-1}(a), \frac{f_i(x) - f_i(a)}{x - a}, f_{i+1}(a), \dots, f_n(a)\Big).$$

On déduit alors le résultat du théorème de continuité des fonctions composées (*cf.* le module *Espaces vectoriels normés*, proposition 13 de la page 410). ∎

On peut appliquer la proposition ci-dessus (avec $n = 2$) aux « produits usuels » (produit d'une fonction scalaire par une application à valeurs vectoriels, produits scalaires…).

Proposition 83 (Dérivation des fonctions composées).
Soient I, J deux intervalles réels, $\varphi : I \to J$, $a \in I$, E un espace vectoriel normé réel et $f : J \to E$. Si φ est dérivable en a et f dérivable en $\varphi(a)$, alors $f \circ \varphi$ est dérivable en a et l'on a $(f \circ \varphi)'(a) = \varphi'(a)\, f'\big(\varphi(a)\big)$.

On définit les dérivées d'ordre supérieur comme dans le cas des fonctions à valeurs scalaires en étendant les notations.
On a une généralisation de la formule de Leibniz. La preuve, par récurrence en utilisant la proposition 82 ci-dessus et la relation $\binom{m-1}{k} + \binom{m-1}{k-1} = \binom{m}{k}$, est similaire ; elle est laissée au lecteur.

Proposition 84 (Formule de Leibniz). Soient I un intervalle réel, E, F, G des espaces vectoriels normés réels, $u : E \times F \to G$ une application bilinéaire *continue* et $f : I \to E$, $g : I \to F$ deux applications. Si f et g admettent des dérivées $m^{\text{èmes}}$ en $a \in I$, l'application $h := u \circ (f, g) : I \to G$ admet une dérivée $m^{\text{ème}}$ en a et l'on a :

$$h^{(m)}(a) = \sum_{k=0}^{m} \binom{m}{k} u\big(f^{(k)}(a), g^{(m-k)}(a)\big).$$

6.2 L'application intégrale

Nous utiliserons, dans cette partie, les conventions de la page 561 et la relation de Chasles sous la forme de la proposition 19.

Définition 14. Soit $f : I \to E$ on dit qu'elle est $\mathscr{R}$*-localement intégrable* sur I si elle est $\mathscr{R}$-intégrable sur tout *sous-segment* de I.

On dira souvent *localement intégrable* au lieu de $\mathscr{R}$*-localement intégrable* pour alléger la rédaction.

> ***Remarque***. Toute fonction continue par morceaux, monotone ou plus généralement réglée est $\mathscr{R}$-localement intégrable.

Le but de ce paragraphe est l'étude de l'application $x \mapsto \displaystyle\int_a^x f$ (dite *application intégrale*) quand f est localement intégrable sur I et $a \in I$. Le *Théorème fondamental de l'analyse* dit « en gros » que les deux opérations fondamentales de l'analyse c'est-à-dire la dérivation et l'intégration sont inverses l'une de l'autre ; il apparaît (sous une forme évidemment non rigoureuse) dès les débuts du calcul infinitésimal chez Leibniz et Newton. Les énoncés corrects sont malheureusement plus délicats comme nous allons le voir. Nous nous placerons dans le cadre de l'intégrale de Riemann. Nous verrons dans le module *Intégration* du livre [L2-L3] que, dans le cas des fonctions, on obtient des énoncés beaucoup plus satisfaisants dans le cadre de l'intégrale de Henstock-Kurzweil.

Proposition 85. Soit $f : I \to E$ localement intégrable sur I. Pour tout $a \in I$ l'application intégrale $x \mapsto \displaystyle\int_a^x f$ est *continue*.

Démonstration.
Soit $x_0 \in I$. Il existe un segment $[c,d] \subset I$ qui est un voisinage de x_0 *dans* I. La fonction f est $\mathscr{R}$-intégrable sur $[c,d]$, elle est donc bornée sur $[c,d]$: $\sup\limits_{x\in[c,d]} \|f(x)\| \leqslant M$.
Pour tout $x \in [c,d]$, on a, en utilisant la relation de Chasles :

$$\left\| \int_a^x f - \int_a^{x_0} f \right\| \leqslant \left\| \int_{x_0}^x f \right\| \leqslant \sup_{x\in[c,d]} \|f(x)\| \, |x - x_0| = M\,|x - x_0|.$$

La fonction $x \mapsto \displaystyle\int_a^x$ est lipshitzienne et donc continue en x_0. ∎

Proposition 86. Soient $f : I \to E$ localement intégrable sur I et $a \in I$.
(i) Si f admet une limite à droite (resp. à gauche) au point $x_0 \in I$, l'application intégrale $x \mapsto \displaystyle\int_a^x f$ admet $f(x_0^+)$ (resp. $f(x_0^-)$) pour dérivée à droite (resp. à gauche) en x_0.
(ii) Si f est continue en $x_0 \in I$, l'application intégrale $x \mapsto \displaystyle\int_a^x f$ est dérivable en x_0 et sa dérivée en ce point est $f(x_0)$.

Démonstration. Notons, pour tout $x \in I$, $F(x) := \displaystyle\int_a^x f$.
(i) Supposons que f admette une limite *à droite* au point $x_0 \in I$.

Il existe $\alpha > 0$ tel que $x_0 + \alpha \in I$. En modifiant si nécessaire f en x_0, on peut supposer que $f(x_0) = f(x_0^+)$ (l'application intégrale ne change pas). Soit $\varepsilon > 0$; il existe $0 < \eta \leqslant \alpha$ tel que, pour tout $x \in [x_0, x_0 + \eta]$, $\|f(x) - f(x_0^+)\| \leqslant \varepsilon$.

Par suite, pour tout $x \in]x_0, x_0 + \eta]$:

$$\|F(x)-F(x_0)-(x-x_0)f(x_0^+)\| = \left\| \int_{x_0}^{x} (f - f(x_0^+)) \right\| \leqslant \int_{x_0}^{x} \|f-f(x_0^+\| \leqslant \varepsilon(x-x_0)$$

et $\left\|\frac{F(x)-F(x_0)}{x-x_0} - f(x_0^+)\right\| \leqslant \varepsilon$.

On raisonnerait de façon similaire pour la dérivée à gauche.

(ii) C'est une conséquence immédiate de (i). ∎

Des deux propositions ci-dessus, l'on déduit le résultat suivant.

Corollaire 87. Soient $f : I \to E$ une *fonction réglée* et $a \in I$. L'application intégrale $F : x \mapsto \displaystyle\int_a^x f$ est continue sur I et dérivable sur I sauf sur un ensemble au plus dénombrable. Sur le complémentaire de cet ensemble on a $F' = f$.

6.3 Le théorème des accroissements finis

Nous allons donner des généralisations du théorème des accroissements finis traité dans [L1] qui nous seront utiles dans la partie 6.4.

Théorème 88 (Théorème des accroissements finis).
Soient E un $\mathbb{K}$-espace vectoriel normé et $S \subset [a, b]$ un sous-ensemble *au plus dénombrable*. Soient $f : [a, b] \to E$ et $g : [a, b] \to \mathbb{R}$ deux applications *continues* sur $[a, b]$ et *dérivables à droite* en tout point de $[a, b[\setminus S$. Si, pour tout $x \in [a, b[\setminus S$, on a $\|f'_d(x)\| \leqslant g'_d(x)$, alors :

$$\|f(b) - f(a)\| \leqslant g(b) - g(a). \tag{20}$$

Démonstration.

1. Nous allons d'abord prouver le résultat dans le cas où $S = \varnothing$. On pourra se limiter à ce cas en première lecture.

   On va montrer que, pour tout $\varepsilon > 0$:

   $$\|f(b) - f(a)\| \leqslant g(b) - g(a) + \varepsilon(b - a + 1), \tag{21}$$

   ce qui entraînera (20).

   Soit $\varepsilon > 0$ fixé. On définit une application :

   $$\varphi : x \in [a, b] \mapsto \|f(x) - f(a)\| - g(x) + g(a) - \varepsilon(x - a) \in \mathbb{R}.$$

   Cette application est continue.

   Soit $D := \{x \in [a, b] \mid \varphi(x) \leqslant \varepsilon\} \subset [a, b]$. Cet ensemble n'est pas vide ($\varphi(a) = 0$, donc $a \in D$), notons $c \in [a, b]$ sa borne supérieure. L'application φ étant continue, on a $c > a$ (puisque $\varphi(a) = 0$) et $c \in D$.

Nous allons montrer que $c = b$, ce qui entraînera (21) et terminera la preuve. On raisonne par l'absurde et l'on suppose que $c < b$.

Les fonctions f et g étant dérivables à droite en c, il existe $x \in]c, b]$ tel que :

$$\left\|\frac{f(x)-f(c)}{x-c}\right\| \leqslant \left\|f'_d(c)\right\| + \varepsilon/2 \quad \text{et} \quad g'_d(c) \leqslant \frac{g(x)-g(c)}{x-c} + \varepsilon/2,$$

d'où :

$$\begin{aligned}\left\|f(x)-f(c)\right\| &\leqslant (x-c)\left\|f'_d(c)\right\| + \frac{\varepsilon}{2}(x-c)\\ &\leqslant (x-c)\,g'_d(c) + \frac{\varepsilon}{2}(x-c) \leqslant g(x)-g(c)+\varepsilon(x-c).\end{aligned}$$

Puisque $c \in D$, $\left\|f(c)-f(a)\right\| \leqslant g(c)-g(a)+\varepsilon(c-a)+\varepsilon$. Alors :

$$\begin{aligned}\left\|f(x)-f(a)\right\| &\leqslant \left\|f(x)-f(c)\right\| + \left\|f(c)-f(a)\right\|\\ &\leqslant g(x)-g(a)+\varepsilon(x-a)+\varepsilon,\end{aligned}$$

c'est-à-dire $\varphi(x) \leqslant \varepsilon$ et $x \in D$, ce qui contredit $c < x$.

2. Revenons au cas général. La preuve est une modification (un peu délicate…) du cas précédent. Puisque $a < b$, on peut, quitte à augmenter S si nécessaire, le supposer *dénombrable*. En utilisant une bijection entre $\mathbb{N}$ et S on écrit alors $S := \{\sigma_n\}_{n\in\mathbb{N}}$.

On va montrer que, pour tout $\varepsilon > 0$:

$$\left\|f(b)-f(a)\right\| \leqslant g(b)-g(a)+\varepsilon(b-a+2), \tag{22}$$

ce qui entraînera (20).

Soit $A \subset [a, b]$ le sous-ensemble formé des points x tels que, pour tout $a \leqslant y < x$:

$$\left\|f(y)-f(a)\right\| \leqslant g(y)-g(a)+\varepsilon(y-a)+\varepsilon \sum_{\{n|\sigma_n<y\}} 2^{-n}. \tag{23}$$

L'ensemble $\{n \mid \sigma_n < y\}$ peut être infini, $\sum_{\{n|\sigma_n<y\}} 2^{-n}$ est alors la somme d'une série à termes positifs majorée par 2.

L'ensemble A contient évidemment a, il n'est donc pas vide. Soit c sa borne supérieure. Si $a \leqslant z \leqslant x$ et si $x \in A$, alors, d'après la définition de A, $z \in A$, A est donc un *intervalle*. Si $a \leqslant y < c$, il existe $x \in A$ tel que $y < x < c$, donc l'inégalité (23) est vérifiée, ainsi $c \in A$ et $A = [a, c]$.

Pour tout $a \leqslant y < c$, $\sum_{\{n|\sigma_n<y\}} 2^{-n} \leqslant \sum_{\{n|\sigma_n<c\}} 2^{-n}$, donc :

$$\left\|f(y)-f(a)\right\| \leqslant g(y)-g(a)+\varepsilon(y-a)+\varepsilon \sum_{\{n|\sigma_n<c\}} 2^{-n},$$

et, d'après la continuité de f et g :

$$\left\|f(c)-f(a)\right\| \leqslant g(c)-g(a)+\varepsilon(c-a)+\varepsilon \sum_{\{n|\sigma_n<c\}} 2^{-n}. \tag{24}$$

Puisque $\sum_{\{n|\sigma_n<b\}} 2^{-n} \leqslant 2$, pour obtenir (22) il suffit de montrer que $c = b$.

Supposons que $c < b$. On va distinguer le cas où $c \notin S$ et le cas $c \in S$ et montrer que dans chaque cas on aboutit à une contradiction.

(a) *Cas où* $c \notin S$.
D'après la définition de la dérivée à droite en c, il existe $\eta > 0$ tel que $c+\eta \leqslant b$ et, que, pour tout $c \leqslant y < c+\eta$:

$$\big\|f(y)-f(c)-(y-c)f'_d(c)\big\| \leqslant \frac{\varepsilon}{2}(y-c) \quad \text{et} \quad \big|g(y)-g(c)-(y-c)g'_d(c)\big| \leqslant \frac{\varepsilon}{2}(y-c).$$

On en déduit :

$$\begin{aligned}\big\|f(y)-f(c)\big\| &\leqslant \big\|f'_d(c)\big\|(y-c)+\frac{\varepsilon}{2}(y-c) \leqslant g'_d(c)(y-c)+\frac{\varepsilon}{2}(y-c)\\ &\leqslant g(y)-g(c)+\varepsilon(y-c).\end{aligned}$$

De (24) on déduit :

$$\begin{aligned}\big\|f(y)-f(a)\big\| &\leqslant \big\|f(y)-f(c)\big\|+\big\|f(c)-f(a)\big\|\\ &\leqslant g(y)-g(a)+\varepsilon(y-a)+\varepsilon\sum_{\{n|\sigma_n<c\}}2^{-n}\\ &\leqslant g(y)-g(a)+\varepsilon(y-a)+\varepsilon\sum_{\{n|\sigma_n<y\}}2^{-n},\end{aligned}$$

donc $c+\eta \in A$, ce qui contredit la définition de c.

(b) *Cas où* $c \in S$.
Il existe $m \in \mathbb{N}$ tel que $c = \sigma_m$. D'après la continuité de f et g, il existe $\eta > 0$ tel que $c+\eta \leqslant b$ et que, pour tout $c \leqslant y < c+\eta$:

$$\big\|f(y)-f(c)\big\| \leqslant \frac{\varepsilon}{2}\,2^{-m} \quad \text{et} \quad \big|g(y)-g(c)\big| \leqslant \frac{\varepsilon}{2}\,2^{-m}.$$

De (24) on déduit :

$$\begin{aligned}\big\|f(y)-f(a)\big\| &\leqslant \big\|f(y)-f(c)\big\|+\big\|f(c)-f(a)\big\|\\ &\leqslant g(c)-g(a)+\varepsilon(c-a)+\varepsilon\sum_{\{n|\sigma_n<c\}}2^{-n}+\frac{\varepsilon}{2}\,2^{-m}\\ &= g(y)-g(a)+\varepsilon(c-a)+\varepsilon\sum_{\{n|\sigma_n<c\}}2^{-n}+\frac{\varepsilon}{2}\,2^{-m}+g(c)-g(y)\\ &\leqslant g(y)-g(a)+\varepsilon(c-a)+\varepsilon\sum_{\{n|\sigma_n\leqslant c\}}2^{-n}\\ &\leqslant g(y)-g(a)+\varepsilon(c-a)+\varepsilon\sum_{\{n|\sigma_n<y\}}2^{-n},\end{aligned}$$

donc $c+\eta \in A$, ce qui contredit la définition de c. ∎

Corollaire 89 (Inégalité des accroissements finis).
Soient E un $\mathbb{K}$-espace vectoriel normé et $S \subset [a,b]$ un sous-ensemble *au plus dénombrable*. Soit $f : [a,b] \to E$ une applications *continue* sur $[a,b]$ et *dérivable à droite* en tout point de $[a,b[\,\setminus S$. Si, pour tout $x \in [a,b[\,\setminus S$, on a $\big\|f'_d\big\| \leqslant C$, alors : $\big\|f(b)-f(a)\big\| \leqslant C(b-a)$.

Démonstration.
On applique le théorème des accroissements finis avec $g : x \in [a,b] \mapsto Cx \in \mathbb{R}$. ∎

Corollaire 90. Soient E un $\mathbb{K}$-espace vectoriel normé et $S \subset I$ un sous-ensemble *au plus dénombrable*. Soit $f : I \to E$ une application continue sur I et dérivable en tout point de $I \setminus S$. Si, pour tout $x \in I \setminus S$, on a $f' = 0$, alors f est *constante* sur I.

Le résultat reste vrai si l'on remplace « dérivable » par « dérivable à droite » en ajoutant, si nécessaire la borne supérieure de I à S.

Démonstration. Soient $a, b \in I$, $a < b$. On peut appliquer le théorème des accroissements finis (*cf.* le théorème 88 ci-dessus) à la restriction de f à $[a, b]$, avec $g := 0$. On obtient $\big\|f(b) - f(a)\big\| \leqslant g(b) - g(a) = 0$, donc $f(a) = f(b)$. ∎

6.4 Primitives et théorème fondamental

Dans cette partie E est un $\mathbb{K}$-espace vectoriel normé. On pourra en première lecture se limiter au cas $E = \mathbb{K}$.

6.4.1 Primitives

Définition 15. Soit $f : I \to E$ une application. On appelle *primitive* de f sur I toute application $F : I \to E$ qui est dérivable et admet f pour application dérivée.

On vérifie facilement le résultat suivant.

Proposition 91. Si F est une primitive de $f : I \to E$ sur I, $G : I \to E$ est une primitive de f si, et seulement s'il existe une constante $C \in E$ telle que $G = F + C$.

Définition 16. Soit $f : I \to E$ une application. On appelle *primitive généralisée* de f sur I toute application $F : I \to E$ qui est *continue*, dérivable sur le complémentaire d'un ensemble *au plus dénombrable* S et admet f pour application dérivée sur $I \setminus S$.

La proposition suivante est une généralisation de la proposition 91 ci-dessus.

Proposition 92. Si F est une primitive généralisée de f sur I, $G : I \to E$ est une primitive généralisée de f si, et seulement s'il existe une constante $C \in E$ telle que $G = F + C$.

Démonstration. Si F est une primitive généralisée de f et si $C \in E$, $G := F + C$ est évidemment une primitive généralisée de f. Inversement, soient F et G deux primitives généralisées de f, la fonction $G - F$ est dérivable et de dérivée nulle sur $I \setminus S$, où S est un ensemble au plus dénombrable. D'après le corollaire 90, $G - F$ est constante. ∎

6.4.2 Le théorème fondamental

Il existe deux versions « idéales » du théorème fondamental de l'analyse, exprimant que *dérivation* et *intégration* sont deux opérations « inverses l'une de l'autre », dans la ligne des idées des fondateurs du calcul différentiel :

1. si $\dfrac{\mathrm{d}F}{\mathrm{d}x} = f$, alors $\displaystyle\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$;

2. $\dfrac{\mathrm{d}}{\mathrm{d}x}\left(\displaystyle\int_a^x f(t)\,\mathrm{d}t\right) = f(x)$.

Malheureusement tout n'est pour le mieux que dans le meilleur des mondes *possibles* et il faut ajouter des conditions pour que l'assertion 1 (resp. l'assertion 2) soit vraie. De plus le résultat dépend de la théorie de l'intégration choisie.

Théorème 93 (Théorème fondamental, première version).
Soit $f : I \to E$ une application *continue*.
(i) L'application $F : x \in I \mapsto \displaystyle\int_a^x f \in E$ est une *primitive* de f.
(ii) Une application $G : I \to E$ est une primitive de f si, et seulement s'il existe une constante $C \in E$ telle que $G = F + C$.
Pour tous $a, b \in I$ on a alors $\displaystyle\int_a^b f = G(b) - G(a)$.

Corollaire 94. Soit $f : I \to E$ une application de classe $\mathcal{C}^1$.
Pour tous $a, b \in I$, l'on a $f(b) - f(a) = \displaystyle\int_a^b f'$.

On peut se demander si le résultat ci-dessus reste vrai si l'on suppose seulement que la fonction f est *dérivable*. En fait l'énoncé lui même pose problème car f' n'est pas nécessairement $\mathscr{R}$-intégrable sur les sous-segments de I (*cf.* l'exemple de la page ci-contre). Nous verrons plus loin que si l'on suppose que f' est $\mathscr{R}$-intégrable sur les sous-segments de I le résultat reste vrai. Tout se passe encore mieux si l'on remplace l'intégrale de Riemann par l'intégrale de Henstock-Kurzweil ; on peut alors montrer que f' est toujours intégrable en ce sens et que $f(b) - f(a) = \displaystyle\int_a^b f'$. Nous en donnerons une preuve dans le module *Intégration* de [L2-L3].

Théorème 95 (Théorème fondamental pour les applications réglées).
Soit $f : I \to E$ une application *réglée*.
(i) L'application $F : x \in I \mapsto \displaystyle\int_a^x f \in E$ est une *primitive généralisée* de f.
(ii) Une application $G : I \to E$ est une primitive généralisée de f si, et seulement s'il existe une constante $C \in E$ telle que $G = F + C$.
Pour tous $a, b \in I$ on a alors $\displaystyle\int_a^b f = G(b) - G(a)$.

Démonstration. C'est une conséquence immédiate de résultats établis antérieurement.
- L'assertion (i) est une reformulation du corollaire 87 de la page 602.
- L'assertion (ii) résulte du corollaire 90 de la page précédente. ∎

Comme nous l'avons annoncé plus haut, on peut généraliser le corollaire 94 ci-dessus au cas où la dérivée de f est seulement $\mathscr{R}$-intégrable sur les sous-segments de I au lieu de continue.

Théorème 96. Soit $f : I \to E$ une application *dérivable* dont la dérivée est $\mathscr{R}$-intégrable sur les sous-segments de I. Pour tous $a, b \in I$ l'on a $f(b) - f(a) = \displaystyle\int_a^b f'$.

Démonstration. Nous prouverons plus loin ce résultat à l'aide du théorème d'Arzela (*cf.* le corollaire 122 de la page 632). ∎

Comme le montre l'exemple suivant, il existe des fonctions dérivables sur un segment I dont la dérivée n'est pas $\mathscr{R}$-intégrable sur I.

Exemple. Soit $f : [0,1] \to \mathbb{R}$ définie par $f(x) = \begin{cases} 0 & \text{si } x = 0 \\ x^2 \sin \frac{1}{x^2} & \text{si } 0 < x \leqslant 1. \end{cases}$

On a $f'(0) = 0$ et $f'(x) = 2x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2}$ si $x > 0$. La fonction f' n'est pas bornée sur $[0,1]$, elle n'est donc pas $\mathscr{R}$-intégrable sur cet intervalle.
Nous verrons plus loin que f' est intégrable au sens de Riemann généralisé sur $]0,1]$ et qu'en ce sens $\displaystyle\int_0^1 f' = f(1) - f(0)$ (*cf.* l'exercice 20 de la page 622).

On peut se demander si, quand f' est *bornée* (ce qui n'est pas le cas dans l'exemple ci-dessus), f' est $\mathscr{R}$-intégrable. Un exemple classique dû à Vito Volterra, basé sur un jeu qui utilise à la fois un ensemble de type Cantor et la fonction de l'exemple ci-dessus, montre que f' peut être bornée et non $\mathscr{R}$-intégrable. Dans ce cas f' est intégrable au sens de Lebesgue (c'est le point de départ des recherches de Lebesgue sur l'intégration).

6.5 Changement de variables, intégration par parties

Les formules de changement de variables et d'intégration par parties sont des outils de calcul essentiels, ils sont aussi très importants sur le plan théorique.
Les théorèmes 97 et 98 ci-dessous généralisent ceux qui ont été établis dans le module *Intégration* de [L1] pour les fonctions à valeurs réelles. Les preuves sont similaires *mutatis-mutandis* et nous ne les détaillerons pas.

6.5.1 Changement de variables

Théorème 97 (Changement de variables). Soient E un espace de Banach, I et J deux intervalles réels, $\varphi : I \to J$ une fonction *continûment dérivable* et $f : J \to E$ une application *continue*. Alors, pour tous $\alpha,\ \beta \in I$, on a :

$$\int_\alpha^\beta (f \circ \varphi)\, \varphi' = \int_{\varphi(\alpha)}^{\varphi(\beta)} f.$$

Notation. Le théorème ci-dessus justifie la notation traditionnelle de l'intégrale $\int_a^b f(x)\mathrm{d}x := \int_a^b f$, due à Leibniz. On a alors :

$$\int_\alpha^\beta f\big(\varphi(t)\big)\,\varphi'(t)\,\mathrm{d}t = \int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)\,\mathrm{d}x,$$

que l'on peut écrire :

$$\int_\alpha^\beta f\big(\varphi\big)\ \mathrm{d}\varphi = \int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)\,\mathrm{d}x,$$

pour retenir la formule et faciliter les calculs.

6.5.2 Intégration par parties

Formule d'intégration par parties

Théorème 98 (Intégration par parties). Soient E un espace de Banach, I un intervalle réel, $u : I \to \mathbb{R}$ et $v : I \to E$ deux applications continûment différentiables. Alors, pour tous $a,\ b \in I$, on a :

$$\int_a^b u(x)v'(x)\,\mathrm{d}x = \Big[u(x)v(x)\Big]_a^b - \int_a^b u'(x)v(x)\,\mathrm{d}x.$$

Formule de Taylor avec reste intégral

Le résultat ci-dessous généralise celui donné pour les fonctions à valeurs réelles dans le module *Approximation* de [L1]. Comme dans le cas scalaire on le prouve par récurrence sur n en utilisant le théorème d'intégration par parties.

Proposition 99 (Formule de Taylor avec reste intégral). Soient E un espace de Banach, $n \in \mathbb{N}$ et $f : [a, b] \to E$ une application de classe $\mathcal{C}^{n+1}$. Alors :

$$f(b) = f(a) + \cdots + \frac{(b-a)^n}{n!} f^{(n)}(a) + \int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\,\mathrm{d}x. \qquad (25)$$

Remarque. En utilisant le changement de variables $t := \frac{x-a}{b-a}$, on écrit le « reste intégral » $\displaystyle\int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\,\mathrm{d}x$ sous la forme suivante :

$$\int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\,\mathrm{d}x = \frac{(b-a)^{n+1}}{n!} \int_0^1 (1-t)^n f^{(n+1)}\big((1-t)a + tb\big)\,\mathrm{d}t.$$

Deuxième formule de la moyenne

Proposition 100 (Deuxième formule de la moyenne).
Soient $f : [a,b] \to \mathbb{R}^+$ une fonction *décroissante* (au sens large) et $g : [a,b] \to \mathbb{R}$ une fonction *réglée*. Alors il existe $c \in [a,b]$ tel que :

$$\int_a^b fg = f(a) \int_a^c g$$

Démonstration.
On peut supposer $f(a) \neq 0$ (sinon le résultat est évident). Quitte à multiplier f par une constante, on peut aussi supposer $f(a) = 1$. Nosons $G(x) := \displaystyle\int_a^x g(t)\,\mathrm{d}t$. La fonction G est continue sur $[a,b]$, donc bornée, et elle prend toute valeur comprise entre sa borne inférieure m et sa borne supérieure M. Il suffit donc de montrer que $m \leqslant \displaystyle\int_a^b fg \leqslant M$.

Un cas particulier
Il y a une preuve très simple de la proposition quand f est *de classe* $\mathcal{C}^1$ et g *continue*. On pourra se limiter à ce cas en première lecture.
D'après la formule d'intégration par parties (G est dérivable de dérivée g), on a :

$$\int_a^b f(x)g(x)\,\mathrm{d}x = \big[f(x)G(x)\big]_a^b - \int_a^b f'(x)G(x)\,\mathrm{d}x = f(b)G(b) + \int_a^b (-f'(x))G(x)\,\mathrm{d}x.$$

On a $-f' \geqslant 0$, donc, d'après la première formule de la moyenne (*cf.* la proposition 29 de la page 569), il existe $c \in [a,b]$ tel que :

$$\int_a^b (-f'(x))G(x)\,\mathrm{d}x = G(c)\int_a^b (-f'(x))\,\mathrm{d}x = G(c)\big(f(a) - f(b)\big).$$

Donc :

$$\int_a^b f(x)g(x)\,\mathrm{d}x = f(b)G(b) + \big(f(a) - f(b)\big)G(c) = f(b)G(b) + \big(1 - f(b)\big)G(c).$$

On a $0 \leqslant f(b) \leqslant 1$, donc, par convexité :

$$f(b)G(b) + \big(1 - f(b)\big)G(c) \in [m,M] \quad \text{et} \quad m \leqslant \int_a^b f(x)g(x)\,\mathrm{d}x \leqslant M.$$

Le cas général

1. Montrons d'abord le résultat quand g est une fonction en escalier. Soit $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ une subdivision adaptée à g : $x_0 := a < x_1 < \ldots < x_n := b$. Pour tout $i \in [\![1,n]\!]$, la fonction g est constante et égale à c_i sur $]x_{i-1}, x_i[$.

   Pour tout $i \in [\![1,n]\!]$, on note $\mu_i := \frac{1}{x_i - x_{i-1}} \displaystyle\int_{x_{i-1}}^{x_i} f$. En utilisant la décroissance de f, on obtient : $f(x_{i+1}) \leqslant \mu_i = \frac{1}{x_i - x_{i-1}} \displaystyle\int_{x_{i-1}}^{x_i} f \leqslant f(x_{i-1})$, et par suite :

   $$\mu_0 := f(a) = 1 \geqslant \mu_1 \geqslant f(x_1) \geqslant \mu_2 \geqslant f(x_2) \geqslant \ldots \geqslant \mu_n \geqslant f(b) \geqslant 0.$$

On a, de plus, pour tout $i \in [\![1,n]\!]$, $G(x_i) - G(x_{i-1}) = c_i(x_i - x_{i-1})$, donc (en utilisant $G(x_0) = 0$) :

$$\int_a^b fg = \sum_{i\in[\![1,n]\!]} c_i \int_{x_{i-1}}^{x_i} f = \sum_{i\in[\![1,n]\!]} \mu_i\big(G(x_i) - G(x_{i-1})\big)$$
$$= \Big(\sum_{i\in[\![0,n-1]\!]} G(x_i)(\mu_i - \mu_{i+1})\Big) + G(x_n)\mu_n.$$

On reconnait dans la dernière ligne le *barycentre* des nombres réels :

$$G(x_0), \dots, G(x_{n-1}), G(x_n) \in [m, M]$$

affectés des coefficients *positifs* $\mu_0 - \mu_1, \dots, \mu_{n-1} - \mu_n, \mu_n$, de somme $\mu_0 = 1$. Par convexité ce barycentre appartient au segment $[m, M]$, donc $\displaystyle\int_a^b fg \in [m, M]$.

2. Traitons maintenant le cas général en supposant que g est une fonction réglée. Soit $\varepsilon > 0$, il existe une fonction en escalier $h \in \mathcal{E}([a,b])$ telle que $\|g - h\| \leqslant \frac{\varepsilon}{2(b-a)}$ ($\|.\|$ désignant la norme sup sur $[a,b]$). Posons $H(x) := \displaystyle\int_a^x h(t)\,\mathrm{d}t$. La fonction H est continue sur $[a,b]$, donc bornée. Notons m' sa borne inférieure et M' et sa borne supérieure. D'après 1 : $m' \leqslant \displaystyle\int_a^b fh \leqslant M'$. Par ailleurs $\|G - H\| \leqslant (b-a)\frac{\varepsilon}{2(b-a)} = \frac{\varepsilon}{2}$, donc $m - \varepsilon/2 \leqslant m' \leqslant M' \leqslant M + \varepsilon/2$. On a :

$$\Big|\int_a^b fg - \int_a^b fh\Big| \leqslant \int_a^b |g-h| \leqslant (b-a)\frac{\varepsilon}{2(b-a)} = \frac{\varepsilon}{2}.$$

On en déduit $m - \varepsilon \leqslant \displaystyle\int_a^b fg \leqslant M + \varepsilon$, puis, $\varepsilon > 0$ étant arbitraire, $m \leqslant \displaystyle\int_a^b fg \leqslant M$. ∎

Remarque. Il existe une version plus générale de la proposition ci-dessus. On peut supposer seulement que g est $\mathscr{R}$-intégrable (*cf.* par exemple [R3], 6.3.6).

Exercice 15.

Soit f une fonction *positive décroissante* sur $[a,b]$. Montrer que, pour tout $x \neq 0$, $\Big|\displaystyle\int_a^b f(t)e^{ixt}\,\mathrm{d}t\Big| \leqslant 4f(a)/|x|$.

Solution. Montrons $\Big|\displaystyle\int_a^b f(t)\cos(xt)\,\mathrm{d}t\Big| \leqslant 2f(a)/|x|$.

On montrerait de façon analogue $\Big|\displaystyle\int_a^b f(t)\sin(xt)\,\mathrm{d}t\Big| \leqslant 2f(a)/|x|$).

D'après la deuxième formule de la moyenne (avec $g := \cos(xt)$)

$$\int_a^b f(t)\cos(xt)\,\mathrm{d}t = f(a)\int_a^c \cos(xt)\,\mathrm{d}t = \frac{f(a)}{x}\big[\sin(xt)\big]_a^b = \frac{f(a)}{x}\big(\sin(xb) - \sin(xa)\big).$$

Le résultat s'en déduit.

---

7 Intégration sur un intervalle quelconque

Dans cette partie, E est un espace de Banach et $I \subset \mathbb{R}$ est un intervalle d'intérieur non vide. Nous noterons a la borne inférieure de I et b sa borne supérieure dans $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ (il se peut que l'intervalle I ne soit pas borné et dans ce cas l'on a $a = -\infty$ ou $b = +\infty$).
En première lecture on pourra se limiter au cas où E est de dimension finie.
Dans la théorie de Riemann, si $I = [a, b[$, on dit qu'une fonction $f : I \to \mathbb{R}$ est « intégrable » si pour tout $c \in [a, b[$ la fonction f est intégrable sur $[a, c]$ et si $\displaystyle\int_a^c f$ admet une limite finie quand $c \to b^-$. Plus précisément on dit alors que f admet une *intégrale impropre* sur I et l'on définit :

$$\int_I f = \lim_{c \to b^-} \int_a^c f.$$

Il existe des fonctions sur $\mathbb{R}$ qui admettent une intégrale impropre au sens de Riemann mais qui ne sont pas intégrables au sens de Lebesgue (par exemple $x \mapsto \sin(x)/x$ comme nous le verrons plus loin). C'est une lacune de l'intégrale de Lebesgue qui ne considère comme intégrables que des fonctions dont la *valeur absolue* est intégrable. Une vision de l'intégration exclusivement centrée sur l'intégrale de Lebesgue conduit beaucoup d'auteurs à écarter de telles fonctions (dites semi-intégrables), ce qui est dommageable pour les applications...

7.1 Intégrales impropres au sens de Riemann

Dans cette partie E est un $\mathbb{K}$-espace de Banach. On pourra, en première lecture, se limiter au cas où $E = \mathbb{R}$.

Définition 17. Soit $f : [a, b[\to E$ (respectivement sur $f :]a, b] \to E$) une fonction $\mathscr{R}$-localement intégrable (*cf.* la définition de 14 de la page 600). On dit que *l'intégrale impropre* $\displaystyle\int_a^b f$ est *convergente* ou que la fonction f est intégrable sur $[a, b[$ (resp. $]a, b]$) si l'intégrale $\displaystyle\int_a^x f$ (respectivement $\displaystyle\int_x^b f$) a une limite finie $A \in E$ quand $x \to b^-$ (respectivement $x \to a^+$). Dans ce cas, on dit que A est la valeur de cette intégrale impropre convergente et l'on pose : $\displaystyle\int_a^b f := A$.
Dans le cas contraire, on dit que l'intégrale impropre est *divergente*.

Si f est $\mathscr{R}$-intégrable sur $[a, b]$ les intégrales impropres de ses restrictions à $[a, b[$ et $]a, b]$ sont convergentes et leur valeur est l'intégrale de Riemann de f sur $[a, b]$: $\displaystyle\int_a^b f$.
Notons que si $f : [a, b[\to E$ est localement intégrable, alors pour tout $c \in]a, b[$, on a, d'après la relation de Chasles :

$$\int_a^c f + \int_c^x f = \int_a^x f.$$

On en déduit, par passage à la limite, que l'intégrale impropre $\int_a^b f$ est convergente si, et seulement si, l'intégrale impropre $\int_c^b f$ est convergente. Dans ce cas,

$$\int_a^c f + \int_c^b f = \int_a^b f.$$

Un raisonnement analogue montre que l'on a le même résultat si $f :]a, b] \to \mathbb{R}$ est localement intégrable. On dispose donc de la relation de Chasles pour les intégrales impropres.

Définition 18. Lorsque $f :]a, b[\to E$ est localement intégrable, l'intégrale impropre $\int_a^b f$ est convergente si pour un (et donc pour tout) $c \in]a, b[$, les intégrales impropres $\int_a^c f$ et $\int_c^b f$ sont convergentes. On pose alors : $\int_a^b f := \int_a^c f + \int_c^b f$.

La relation de Chasles implique que la valeur de l'intégrale impropre $\int_a^b f$ ne dépend pas du choix de $c \in]a, b[$.

Si l'intégrale impropre $\int_a^b f$ est convergente on dit aussi que f est *intégrable au sens de Riemann généralisé* sur l'intervalle $]a, b[$.

Les résultats qui suivent sont énoncés et démontrés pour des fonctions définies sur des intervalles de la forme $[a, b[$ avec $a \in \mathbb{R}$ et $b \in \mathbb{R} \cup \{+\infty\}$. Le lecteur les généralisera sans problème au cas des fonctions définies sur des intervalles de la forme $]a, b]$ avec $a \in \mathbb{R} \cup \{-\infty\}$ et $b \in \mathbb{R}$.

Les exemples suivants sont très utiles.

Exemples.

1. La fonction $x \mapsto \ln x$ est intégrable au sens de Riemann généralisé sur $]0, 1]$ mais elle n'est pas bornée sur cet intervalle. La fonction $x \mapsto x \ln x - x$ est une primitive de $x \mapsto \ln x$ sur $]0, +\infty[$, donc, pour tout $a \in]0, 1]$, $\int_a^1 \ln x = -1 - a \ln a + a$ et $\lim\limits_{a \to 0^+} \int_a^1 \ln x = -1$.
2. La fonction $f : t \mapsto e^{-t}$ admet une intégrable impropre convergente sur $[0, +\infty[$. En effet elle est continue sur $[0, +\infty[$ et, de plus :

$$\int_0^c e^{-t} \, \mathrm{d}t = \left[-e^{-t}\right]_0^c = 1 - e^{-c} \underset{c \to +\infty}{\longrightarrow} 1.$$

On vérifie facilement que l'ensemble des applications de l'intervalle I dans E dont l'intégrale sur I converge est un sous-espace vectoriel de $\mathcal{F}(I, E)$.

Une intégrale impropre $\displaystyle\int_I (f+g)$ peut converger tandis que les intégrales impropres $\displaystyle\int_I f$ et $\displaystyle\int_I g$ divergent.

Exemple. Soit $\varphi : x \in [0,1] \to \frac{1}{\sqrt{1+x^2}+\sqrt{1-x^2}} \in \mathbb{R}$.
Cette fonction est continue, donc intégrable sur $[0,1]$. Sa restriction ψ à $]0,1]$ est intégrable au sens généralisé et l'on a $\displaystyle\int_0^1 \psi = \int_0^1 \varphi$.
La fonction ψ s'écrit $x \in]0,1] \to \psi(x) = \frac{\sqrt{1+x^2}-\sqrt{1-x^2}}{2x^2} = \frac{\sqrt{1+x^2}}{2x^2} - \frac{\sqrt{1-x^2}}{2x^2}$, ce qui conduit à considérer les deux intégrales impropres $\displaystyle\int_0^1 \frac{\sqrt{1+x^2}}{2x^2}\,\mathrm{d}x$ et $\displaystyle\int_0^1 \frac{\sqrt{1-x^2}}{2x^2}\,\mathrm{d}x$.
Ces intégrales sont divergentes, *on ne peut donc pas écrire* :

$$\int_0^1 \frac{\mathrm{d}x}{\sqrt{1+x^2}+\sqrt{1-x^2}} = \int_0^1 \frac{\sqrt{1+x^2}}{2x^2}\,\mathrm{d}x - \int_0^1 \frac{\sqrt{1-x^2}}{2x^2}\,\mathrm{d}x.$$

Proposition 101 (Intégrales de référence). Soit p un nombre réel.

(i) L'intégrale impropre $\displaystyle\int_1^{+\infty} \frac{1}{x^p}\,\mathrm{d}x$ est convergente si, et seulement si, $p > 1$.

(ii) L'intégrale impropre $\displaystyle\int_0^1 \frac{1}{x^p}\,\mathrm{d}x$ est convergente si, et seulement si, $p < 1$.

Démonstration. La fonction $x \mapsto 1/x^p$ est continue sur $]0,+\infty[$. Pour $c > 0$:

$$\int_1^c f = \begin{cases} \ln c & \text{si } p = 1 \\ \dfrac{c^{1-p}-1}{1-p} & \text{si } p \neq 1. \end{cases}$$

Il y a une limite finie quand $c \to +\infty$ si, et seulement si, $p > 1$. Il y a une limite finie quand $c \to 0^+$ si, et seulement si, $p < 1$.
Dans les deux cas, cette limite est alors égale à $1/(p-1)$. ∎

Remarque. L'intégrale impropre $\displaystyle\int_0^{+\infty} \frac{1}{x^p}\,\mathrm{d}x$ est toujours divergente.

Un certain nombre de propriétés découlent directement de la relation de Chasles.

Proposition 102.
Soient $f, g : [a,b[\to \mathbb{R}$ des fonctions *positives*, localement intégrables.

(i) L'intégrale impropre $\displaystyle\int_a^b f$ est convergente si, et seulement s'il existe $M \in \mathbb{R}$ tel que pour tout $c \geqslant a$, on ait $\displaystyle\int_a^c f \leqslant M$.

(ii) Si $f \leqslant g$ et si l'intégrale impropre $\displaystyle\int_a^b g$ est convergente, il en est de même de $\displaystyle\int_a^b f$.

Démonstration.

(i) Si $a < c < c' < b$, alors $\displaystyle\int_a^{c'} f = \int_a^c f + \int_c^{c'} f \geqslant \int_a^c f$. La fonction $c \mapsto \displaystyle\int_a^c f$ est donc croissante. Si elle est majorée par M, elle admet une limite quand $c \to b^-$.

(ii) Si $a < c < b$, $\displaystyle\int_a^c f \leqslant \int_a^c g \leqslant M := \int_a^b g$. ∎

Exemple.

La fonction $f : t \mapsto e^{-t^2}$ admet une intégrable impropre convergente sur $[0, +\infty[$.
En effet :

– elle est continue sur $[0, +\infty[$;

– si $c > 1$, $\displaystyle\int_0^c e^{-t^2} = \int_0^1 e^{-t^2} + \int_1^c e^{-t^2}$ et, pour tout $t \geqslant 1$, $e^{-t^2} \leqslant e^{-t}$.

On utilise alors la proposition ci-dessus et l'exemple 2 de la page 612.

On montrera plus loin que $\displaystyle\int_0^{+\infty} e^{-x^2} = \frac{\sqrt{\pi}}{2}$ en utilisant diverses méthodes (*cf.* les exercices 26 et 27 de la page 634 ainsi que l'exercice II.3.33 de la page 657.)

Définition 19. Soit I un intervalle réel d'intérieur non vide. Une *suite exhaustive de segments* de I est une suite croissante de segments inclus dans I, dont la réunion est égale à I.

On déduit facilement le résultat suivant de la proposition ci-dessus.

Corollaire 103. Soit $f : [a, b[\to \mathbb{R}$ une fonction positive, localement intégrable et $(J_n)_{n\in\mathbb{N}}$ une suite exhaustive de segments de I. Les conditions suivantes sont équivalentes.

(i) L'intégrale impropre $\displaystyle\int_a^b f$ est convergente.

(ii) La suite $\Big(\displaystyle\int_{J_n} f\Big)_{n\in\mathbb{N}}$ est bornée.

(iii) La suite $\Big(\displaystyle\int_{J_n} f\Big)_{n\in\mathbb{N}}$ est convergente.

Si ces conditions sont satisfaites, on a $\displaystyle\int_I f = \sup_{n\in\mathbb{N}} \int_{J_n} f = \lim_{n\to+\infty} \int_{J_n} f$.

Exercice 16.

1. Montrer que, pour tout $x > 0$, l'intégrale $\displaystyle\int_0^{+\infty} t^{x-1}e^{-t}\,\mathrm{d}t$ est convergente. On définit la fonction Gamma d'Euler par :

$$\Gamma : x \in \mathbb{R}_+^* \mapsto \Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\,\mathrm{d}t.$$

2. À l'aide d'une intégration par parties, montrer que pour tout $x > 0$, on a la relation $\Gamma(x+1) = x\,\Gamma(x)$.
3. Calculer $\Gamma(1)$ puis déterminer $\Gamma(n)$ pour tout $n \geqslant 1$.

Solution.

1. Soit $x > 0$, posons $f(t) = t^{x-1}e^{-t}$. La fonction f est continue et positive sur $]0,+\infty[$. Notons que pour tout $t > 0$, on a $f(t) \leqslant t^{x-1}$.
   De plus, $t^2 f(t) = t^{x+1}e^{-t} \to 0$ quand $t \to +\infty$. Par conséquent, il existe $A > 0$ tel que $f(t) \leqslant A/t^2$ dès que $t \geqslant 1$. Alors, pour tout $c \in]0,1[$ et tout $d \in]1,+\infty[$, on a :
$$\begin{aligned}\int_c^d f(t)\,\mathrm{d}t &\leqslant \int_c^1 t^{x-1}\,\mathrm{d}t + \int_1^d \frac{A}{t^2}\,\mathrm{d}t\\ &= \left[\frac{t^x}{x}\right]_c^1 + \left[-\frac{A}{t}\right]_1^d\\ &= \frac{1-c^x}{x} + A - \frac{A}{d} \leqslant \frac{1}{x} + A.\end{aligned}$$
   L'intégrale $\displaystyle\int_c^d f$ est donc majorée quand $c \to 0$ et $d \to +\infty$. Comme la fonction f est positive, l'intégrale impropre $\displaystyle\int_0^{+\infty} f$ est donc convergente.
2. Une intégration par parties donne, pour $0 < c < d < +\infty$:
$$\begin{aligned}\int_c^d t^x e^{-t}\,\mathrm{d}t &= \left[t^x \cdot (-e^{-t})\right]_c^d - \int_c^d (xt^{x-1}) \cdot (-e^{-t})\,\mathrm{d}t\\ &= \underbrace{c^x e^{-c}}_{\substack{\longrightarrow 0\\ c\to 0}} - \underbrace{d^x e^{-d}}_{\substack{\longrightarrow 0\\ d\to+\infty}} + x\int_c^d t^{x-1}e^{-t}\,\mathrm{d}t.\end{aligned}$$
   En faisant tendre c vers 0 et d vers $+\infty$, on obtient l'égalité requise :
$$\Gamma(x+1) = \int_0^{+\infty} t^x e^{-t}\,\mathrm{d}t = x\int_0^{+\infty} t^{x-1}e^{-t}\,\mathrm{d}t = x\Gamma(x).$$
3. Comme $\Gamma(1) = \displaystyle\int_0^{+\infty} e^{-t}\,\mathrm{d}t = 1$, on obtient par récurrence :
$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = (n-1)!\Gamma(1) = (n-1)!.$$

Exercice 17.

La fonction Beta d'Euler est définie par :
$$(r,s) \in]0,+\infty[\times]0,+\infty[\mapsto \mathrm{B}(r,s) := \int_0^1 t^{r-1}(1-t)^{s-1}\,\mathrm{d}t.$$

1. Vérifier que $\mathrm{B}(r,s)$ est bien définie.
2. À l'aide du changement de variable $t = \sin^2\theta$, montrer que $\mathrm{B}(1/2,1/2) = \pi$.

Solution.

1. Soit $(r,s) \in]0,+\infty[\times]0,+\infty[$, posons $f(t) = t^{r-1}(1-t)^{s-1}$. La fonction f est continue et positive sur $]0,1[$. Il existe une constante M telle que $t^{r-1} \leqslant M$ si $t \in [1/2,1]$ et $(1-t)^{s-1} \leqslant M$ si $t \in [0,1/2]$. Si $0 < c < 1/2 < d < 1$, on a donc :
$$\begin{aligned}\int_c^d f(t)\,\mathrm{d}t &\leqslant \int_c^{1/2} Mt^{r-1}\,\mathrm{d}t + \int_{1/2}^d M(1-t)^{s-1}\,\mathrm{d}t \\ &\leqslant \left[\frac{Mt^r}{r}\right]_c^{1/2} - \left[\frac{M(1-t)^s}{s}\right]_{1/2}^d \\ &\leqslant M\frac{1}{r2^r} + M\frac{1}{s2^s}\cdot\end{aligned}$$
L'intégrale $\displaystyle\int_c^d f$ est donc majorée quand $c \to 0$ et $d \to 1$. Comme la fonction f est positive, l'intégrale impropre $\displaystyle\int_0^d f$ est donc convergente.

2. Pour $r = s = 1/2$, on a $\mathrm{B}(1/2,1/2) = \displaystyle\int_0^1 \frac{\mathrm{d}t}{\sqrt{t(1-t)}}\cdot$
Le changement de variable proposé donne :
$$\begin{aligned}\int_\varepsilon^{1-\varepsilon'} \frac{\mathrm{d}t}{\sqrt{t(1-t)}} &= \int_{\arcsin\sqrt{\varepsilon}}^{\arcsin\sqrt{1-\varepsilon'}} \frac{2\sin\theta\cos\theta}{\sin\theta\cos\theta}\,d\theta \\ &= 2\big(\arcsin\sqrt{1-\varepsilon'} - \arcsin\sqrt{\varepsilon}\big) \underset{\substack{\varepsilon\to 0\\ \varepsilon'\to 0}}{\longrightarrow} 2(\pi/2-0) = \pi.\end{aligned}$$
Donc $\mathrm{B}(1/2,1/2) = \pi$.

Proposition 104. Soit $f : [a,+\infty[\to \mathbb{R}$ une fonction localement intégrable, telle que $\lim\limits_{x\to+\infty} f(x) = \ell$.

(i) Si l'intégrale impropre $\displaystyle\int_a^{+\infty} f(t)\,\mathrm{d}t$ converge alors $\ell = 0$.

(ii) Si $\ell \neq 0$ alors l'intégrale impropre $\displaystyle\int_a^{+\infty} f(t)\,\mathrm{d}t$ diverge.

Démonstration. La première assertion est la contraposée de la seconde. Il suffit donc de montrer la seconde. Supposons que $\ell > 0$ (le cas $\ell < 0$ est similaire). Il existe alors $c_0 \geqslant a$ tel que pour $t \geqslant c_0$, on ait $f(t) \geqslant \ell/2 > 0$. Alors, pour $c \geqslant c_0$, on a :
$$\int_a^c f = \int_a^{c_0} f + \int_{c_0}^c f \geqslant \int_a^{c_0} f + \frac{\ell}{2}(c-c_0) \underset{c\to+\infty}{\longrightarrow} +\infty.$$

Un argument similaire montre que lorsque l'intégrale impropre $\displaystyle\int_a^{+\infty} f$ converge, alors f ne peut tendre ni vers $+\infty$ ni vers $-\infty$ quand $x \to +\infty$.

Nous ne prétendons aucunement que si l'intégrale impropre $\displaystyle\int_a^{+\infty} f$ converge, alors f admet une limite en $+\infty$ (*cf.* l'exercice II.3.24). Nous affirmons seulement que si cette limite existe, alors elle est nulle. De même, la condition $f(x) \underset{x\to+\infty}{\longrightarrow} 0$ n'implique pas que l'intégrale impropre $\displaystyle\int_a^{+\infty} f$ converge. Un contre-exemple est fourni par la fonction $f : t \mapsto 1/t$ sur $[1, +\infty[$. Enfin, le résultat est faux si l'on considère une intégrale impropre sur $[a, b[$ avec $b < +\infty$.

Proposition 105 (Comparaison série-intégrale). Soit $f : [a, +\infty[\to \mathbb{R}$ une fonction décroissante. L'intégrale impropre $\displaystyle\int_a^{+\infty} f$ converge si, et seulement si, la série $\displaystyle\sum_{k\geqslant 0} f(a+k)$ est convergente.

Une extension du résultat ci-dessus est proposée dans l'exercice II.3.26.

Démonstration. Comme f est décroissante, si l'intégrale impropre $\displaystyle\int_a^{+\infty} f$ converge ou si la série $\displaystyle\sum_{k\geqslant a} f(k)$ est convergente, on voit que $f(x) \underset{x\to+\infty}{\longrightarrow} 0$ et donc f est *positive*. La fonction $F : c \mapsto \displaystyle\int_a^c f$ et la suite $S_n := \displaystyle\sum_{k=0}^{n} f(a+k)$ sont alors *croissantes*.

Comme f est décroissante sur $[a, +\infty[$, on voit que pour tout $k \geqslant 0$, on a :

$$f(a+k+1) \leqslant \int_{a+k}^{a+k+1} f \leqslant f(a+k). \tag{26}$$

D'après la relation de Chasles, on a donc :

$$\forall n \in \mathbb{N}, \quad \sum_{k=1}^{n+1} f(a+k) \leqslant \int_a^{a+n+1} f \leqslant \sum_{k=0}^{n} f(a+k). \tag{27}$$

Si l'intégrale impropre $\displaystyle\int_a^{+\infty} f$ converge, la fonction F admet une limite en $+\infty$ et la suite S_n est majorée, donc convergente. Si la série $\displaystyle\sum_{k\geqslant 0} f(a+k)$ est convergente, la suite S_n converge et la fonction F est majorée, donc convergente. ∎

Proposition 106. Soit $f : [a, +\infty[\to \mathbb{R}^+$ une fonction décroissante. On note, pour tout $n \in \mathbb{N}$:

$$X_n := \sum_{k=0}^{n} f(a+k) - \int_a^{a+n+1} f.$$

Alors la suite $(X_n)_{n\in\mathbb{N}}$ est croissante et convergente.

Démonstration. D'après (26) la suite $(X_n)_{n\in\mathbb{N}}$ est croissante. D'après (27) on a ;

$$0 \leqslant X_n \leqslant \sum_{k=0}^{n} f(a+k) - \sum_{k=1}^{n} f(a+k+1) = f(a) - f(a+k+1) \leqslant f(a). \qquad \blacksquare$$

Série harmonique et constante d'Euler. La fonction $x \in]0, +\infty[\mapsto 1/x \in \mathbb{R}$ est décroissante, donc la série harmonique $\sum \dfrac{1}{n}$ et l'intégrale $\displaystyle\int_1^{+\infty} \frac{\mathrm{d}x}{x}$ sont de même nature.

On a, pour $c > 1$, $\displaystyle\int_1^c \frac{\mathrm{d}x}{x} = \ln c$, donc l'intégrale et la série divergent. (Nous avons vu la divergence de cette dernière par une autre méthode dans le module *Séries numériques* de [L1].)

D'après la proposition 106 la suite de terme général $1 + \frac{1}{2} + \cdots + \frac{1}{n} - \ln n$ est croissante et convergente. Sa limite $\gamma \approx 0,577215$ est appelée constante d'Euler-Mascheroni.

Séries de Bertrand.

On considère les séries de terme général $\left(u_n := \frac{1}{n^\alpha (\ln n)^\beta}\right)_{n\geqslant 2}$, $\alpha,\ \beta \in \mathbb{R}$.

Nous allons d'abord voir que ces séries convergent pour $\alpha > 1$ et divergent pour $\alpha < 1$. Nous traiterons ensuite le cas $\alpha = 1$ en utilisant la proposition 105 ci-dessus.

1. Supposons $\alpha > 1$. Soit $1 < \gamma < \alpha$, on a $n^\gamma u_n = n^{\gamma-\alpha}(\ln n)^{-\beta}$, donc $n\gamma u_n \to 0$ quand $n \to +\infty$ et la série $\sum u_n$ est convergente.
2. Supposons $\alpha < 1$. On a $nu_n = n^{1-\alpha}(\ln n)^{-\beta}$, donc $nu_n \to +\infty$ quand $n \to +\infty$ et la série $\sum u_n$ est divergente.
3. Notons $f_\beta : t \in]1, +\infty[\to \frac{1}{t(\ln t)^\beta}$.

   Cette fonction est dérivable et $f'_\beta(t) = -\frac{\ln t + \beta}{t^2 (\ln t)^{\beta+1}}$.

   Donc f_β est décroissante sur $\left[e^{-\beta}, +\infty\right[$.

   Supposons d'abord $\beta \geqslant 1$. Alors $e^{-\beta} \leqslant e^{-1} \leqslant 1$ et f_β est décroissante sur $]1, +\infty[$.

   On en déduit que la série de Bertrand de terme général $u_n := \dfrac{1}{n(\ln n)^\beta}$ et l'intégrale $\displaystyle\int_2^{+\infty} \frac{1}{t(\ln t)^\beta}$ sont de même nature. Soit $c > 2$, le changement de variables $x := \ln t$ transforme l'intégrale $\displaystyle\int_2^{+\infty} \frac{\mathrm{d}t}{t(\ln t)^\beta}$ en l'intégrale $\displaystyle\int_{\ln 2}^{+\infty} \frac{\mathrm{d}x}{x^\beta}$.

   Cette dernière converge pour $\beta > 1$ et diverge pour $\beta = 1$, donc la série de terme général $u_n := \frac{1}{n(\ln n)^\beta}$ converge pour $\beta > 1$ et diverge pour $\beta = 1$.

   Si $\beta < 1$, $\frac{1}{n(\ln n)^\beta} > \frac{1}{n \ln n}$, donc la série de terme général $u_n := \frac{1}{n(\ln n)^\beta}$ est divergente.

Pour aller plus loin, nous utiliserons le critère de Cauchy.

Théorème 107 (Critère de Cauchy). Soit $f : [a, b[\to E$ une application localement intégrable. L'intégrale impropre $\displaystyle\int_a^b f$ est convergente si, et seulement si, l'intégrale $\displaystyle\int_c^{c'} f$ tend vers 0 quand c et c' tendent simultanément vers b^-.

Démonstration. Si l'on pose $F(c) := \displaystyle\int_a^c f$, la condition signifie que pour tout $\varepsilon > 0$, il existe $c_0 \in [a, b[$ tel que $\big|F(c') - F(c)\big| \leqslant \varepsilon$ dès que $c \geqslant c_0$ et $c' \geqslant c_0$. En vertu du critère de Cauchy pour la limite des fonctions, cette condition est bien nécessaire et suffisante pour que $F(c)$ ait une limite quand c tend vers b^-. ∎

Corollaire 108. Soit $f : [a, b[\to E$ une application localement intégrable.
Si l'intégrale impropre $\displaystyle\int_a^b \|f\|$ est convergente, alors l'intégrale impropre $\displaystyle\int_a^b f$ est convergente.

Démonstration. Si f est localement intégrable, il en est de même de $\|f\|$ et l'on a :

$$\left\| \int_c^{c'} f \right\| \leqslant \int_c^{c'} \|f\|.$$

On peut donc appliquer le critère de Cauchy. ∎

Ce qui précède conduit naturellement à la définition suivante.

Proposition et définition 109. Soit $f :]a, b[\to \mathbb{R}$. Les conditions suivantes sont équivalentes :

(i) f est localement intégrable et les intégrales impropres $\displaystyle\int_a^b |f|$ et $\displaystyle\int_a^b f$ sont convergentes ;

(ii) f est localement intégrable et l'intégrale impropre $\displaystyle\int_a^b |f|$ est convergente.

Si elles sont satisfaites on dit que l'intégrale impropre $\displaystyle\int_a^b f$ est *absolument convergente* ou que f est *absolument intégrable* sur $]a, b[$.

On prendra garde au fait que la convergence de $\displaystyle\int_a^b |f|$ n'implique pas le fait que f est localement intégrable.
La terminologie, bien que classique, est un peu dangereuse…

Remarque. Lorsqu'une intégrale impropre est convergente sans être absolument convergente, on dit qu'elle est semi-convergente. C'est par exemple le cas pour l'intégrale $\displaystyle\int_0^{+\infty} \frac{\sin t}{t}\,\mathrm{d}t$ (*cf.* l'exercice 18 ci-dessous).

Corollaire 110. Soit $f : [a, b[\to \mathbb{R}$ une fonction localement intégrable. S'il existe une fonction localement intégrable $g : [a, b[\to \mathbb{R}$ telle que $|f| \leqslant g$ et telle que l'intégrale impropre $\displaystyle\int_a^b g$ soit convergente, alors l'intégrale impropre $\displaystyle\int_a^b f$ est absolument convergente.

Démonstration. Si $a < c < c' < b$, alors $\displaystyle\int_c^{c'} |f| \leqslant \int_c^{c'} g$. On peut donc appliquer le critère de Cauchy. ∎

Comme application de ce résultat, nous proposons l'exercice suivant qui est un grand classique.

Exercice 18.

Montrer que l'intégrale impropre $\displaystyle\int_1^{+\infty} \frac{\sin x}{x}\,\mathrm{d}x$ est semi-convergente.

Solution. La fonction $f : [1, +\infty[\to \mathbb{R}$ définie par $f(x) = (\sin x)/x$ est continue, donc localement intégrable. Une intégration par parties montre que pour tout $c \in]1, +\infty[$:

$$\int_1^c f = \Big[-\frac{\cos x}{x}\Big]_1^c - \int_1^c \frac{\cos(x)}{x^2}\,\mathrm{d}x.$$

Notons que $\lim\limits_{c\to+\infty} (\cos c)/c = 0$. De plus, les fonctions $x \mapsto (\cos x)/x^2$ et $x \mapsto \big|(\cos x)/x^2\big|$ sont continues sur $[1, +\infty[$, $\big|(\cos x)/x^2\big| \leqslant 1/x^2$ et l'intégrale impropre $\displaystyle\int_1^{+\infty} \frac{1}{x^2}\,\mathrm{d}x$ est convergente. Par conséquent, l'intégrale $\displaystyle\int_1^c \frac{\cos x}{x^2}\,\mathrm{d}x$ a une limite finie quand c tend vers $+\infty$ et l'intégrale impropre $\displaystyle\int_1^{+\infty} \frac{\sin x}{x}\,\mathrm{d}x$ est convergente.

Si $c \geqslant n\pi$, on a la minoration suivante :

$$\begin{aligned}\int_1^c |f| \geqslant \int_\pi^{n\pi} \left|\frac{\sin x}{x}\right| \mathrm{d}x &= \sum_{k=1}^{n-1} \int_{k\pi}^{(k+1)\pi} \left|\frac{\sin x}{x}\right| \mathrm{d}x \\ &\geqslant \sum_{k=1}^{n-1} \frac{1}{(k+1)\pi} \int_{k\pi}^{(k+1)\pi} |\sin x|\,\mathrm{d}x = \sum_{k=1}^{n-1} \frac{2}{(k+1)\pi}.\end{aligned}$$

Par conséquent, $\lim\limits_{c\to+\infty} \displaystyle\int_1^c |f| = +\infty$ et l'intégrale impropre $\displaystyle\int_1^{+\infty} \frac{\sin x}{x}\,\mathrm{d}x$ n'est pas absolument convergente.

La proposition suivante permet de montrer la semi-convergence d'un certain nombre d'intégrales utiles dans les applications (transformée de Fourier, physique…). Dans certains cas on peut la remplacer par une intégration par parties (comme nous venons de le faire dans l'exercice ci-dessus), mais la règle d'Abel est plus générale.

Proposition 111 (Règle d'Abel).
Soient $a \in \mathbb{R}$, $b \in \mathbb{R} \cup \{+\infty\}$, $a < b$, et $f, g : [a, b[\to \mathbb{R}$ deux fonctions telles que :
(i) f est décroissante, et $\lim\limits_{x \to b^-} f(x) = 0$;
(ii) g est une fonction réglée sur tout sous segment de $[a, b[$ et il existe $M \geqslant 0$ tel que, pour tous $c, d \in [a, b[$, on ait $\left| \int_c^d g \right| \leqslant M$.
Alors l'intégrale impropre $\int_a^b fg$ est *convergente*.

Démonstration.
On utilise la seconde formule de la moyenne (*cf.* la proposition 100). Pour tous $a \leqslant u < v < b$, il existe $\chi \in [u, v[$ tel que $\int_u^v fg = f(u) \int_u^\chi g$, donc $\left| \int_u^v fg \right| \leqslant M f(u)$ (d'après (i) f est positive).
Soit $\varepsilon > 0$. D'après (i) il existe $a \leqslant c < b$ tel que, pour tout $x \in [c, b[$, on ait $0 \leqslant f(x) \leqslant \varepsilon / M$.
Alors, pour tous $u, v \in [c, b[$ on a $\left| \int_u^v fg \right| \leqslant \varepsilon$. On peut donc appliquer le critère de Cauchy. ∎

On peut montrer que le résultat reste vrai si l'on suppose seulement que g est localement intégrable (au lieu de réglée sur tout sous-segment) en utilisant une version plus générale de la seconde formule de la moyenne.
On vérifie facilement le résultat suivant (en utilisant des coordonnées).

Corollaire 112. La règle d'Abel reste vraie si g prend ses valeurs dans un $\mathbb{K}$-espace vectoriel normé de dimension finie en remplaçant dans (ii) la valeur absolue par la norme.

On applique souvent la règle d'Abel à des « intégrales oscillantes » de la forme $\int_a^b fg$ avec $g : x \in \mathbb{R} \mapsto e^{i\lambda x} \in \mathbb{C}$ ($\lambda \in \mathbb{R}^*$).
Dans ce cas $\left| \int_u^v g(x)\,\mathrm{d}x \right| = \left| \dfrac{e^{i\lambda}u - e^{i\lambda}u}{i\lambda} \right| \leqslant M := 2/|\lambda|$.

Exercice 19.
Montrer que, pour tout $0 < \alpha \leqslant 1$, l'intégrale impropre $\int_1^{+\infty} \dfrac{e^{ix}}{x^\alpha}\,\mathrm{d}x$ est semi-convergente.

Solution. On applique la règle d'Abel avec $f : x \mapsto \frac{1}{x^\alpha}$ et $g : x \mapsto e^{ix}$. L'intégrale n'est pas absolument convergente car l'intégrale impropre $\int_1^{+\infty} \dfrac{\mathrm{d}x}{x^\alpha}$ est divergente.
Pour $\alpha = 1$, on retrouve le résultat de l'exercice 18.
On peut aussi prouver le résultat par une intégration par parties (*cf.* ci-dessous l'exercice 22).

Exercice 20.
Montrer que la fonction $g : x \in]0,1] \mapsto 2x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2} \in \mathbb{R}$ admet une intégrale impropre convergente sur $]0,1]$. En déduire que si $f : x \in]0,1] \mapsto x^2 \sin \frac{1}{x^2} \in \mathbb{R}$, on a $\displaystyle\int_0^1 g = f(1)$.

Solution. La fonction $x \in]0,1] \mapsto 2x \sin \frac{1}{x^2} \in \mathbb{R}$ est majorée en valeur absolue par la fonction $x \in]0,1] \mapsto 2x$, elle est donc $\mathscr{R}$-intégrable sur $]0,1]$ au sens généralisé. On montre que la fonction $x \in]0,1] \mapsto \frac{2}{x} \cos \frac{1}{x^2} \in \mathbb{R}$ est $\mathscr{R}$-intégrable sur $]0,1]$ au sens généralisé en utilisant la règle d'Abel (on peut aussi prouver ce résultat par une intégration par parties). Donc g est $\mathscr{R}$-intégrable sur $]0,1]$ au sens généralisé. On a $f' = g$, donc, pour tout $0 < \varepsilon < 1$, on a $\displaystyle\int_\varepsilon^1 g = \int_\varepsilon^1 f' = f(1) - f(\varepsilon)$ et l'on en déduit $\displaystyle\int_0^1 g = f(1)$.

7.2 Changement de variables, intégration par parties

7.2.1 Changement de variables

Proposition 113. Soient $\alpha,\ \beta,\ a,\ b \in \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$, avec $\alpha < \beta$ et $a < b$. Soit $\varphi :]\alpha,\beta[\to]a,b[$ un $\mathcal{C}^1$-*difféomorphisme* croissant et $f :]a,b[\to E$ une application continue. Alors $(f \circ \varphi)\,\varphi' :]\alpha,\beta[\to E$ est une application continue et l'intégrale impropre $\displaystyle\int_a^b f$ converge si, et seulement si, l'intégrale impropre $\displaystyle\int_\alpha^\beta (f \circ \varphi)\varphi'$ converge. On a alors :

$$\int_a^b f(x)\,\mathrm{d}x = \int_\alpha^\beta f\big(\varphi(t)\big)\varphi'(t)\,\mathrm{d}t.$$

Il y a des variantes avec φ décroissante ($\alpha < \beta$ et $a > b$) ou des intervalles non ouverts. Nous laissons les détails au lecteur.

Démonstration. En appliquant la formule de changement de variables pour les segments (*cf.* le théorème 97 de la page 607), on obtient :

$$\forall \gamma, \delta \in]\alpha,\beta[\,, \quad \int_{\varphi(\gamma)}^{\varphi(\delta)} f(x)\,\mathrm{d}x = \int_\gamma^\delta f\big(\varphi(t)\big)\,\varphi'(t)\,\mathrm{d}t$$

De même, en notant ψ la bijection inverse de φ, on obtient :

$$\forall c, d \in]a,b[\,, \quad \int_{\psi(c)}^{\psi(d)} f\big(\varphi(t)\big)\,\varphi'(t)\,\mathrm{d}t = \int_c^d f(x)\,\mathrm{d}x.$$

Le résultat s'en déduit aisément. ∎

Exercice 21 (Intégrales de Fresnel).

Montrer que l'intégrale $\displaystyle\int_0^{+\infty} e^{it^2}\,\mathrm{d}t$ est convergente. en déduire que les *intégrales de Fresnel* $\displaystyle\int_0^{+\infty} \cos(t^2)\,\mathrm{d}t$ et $\displaystyle\int_0^{+\infty} \sin(t^2)\,\mathrm{d}t$ sont convergentes.

Solution. Il suffit de montrer que l'intégrale $\displaystyle\int_1^{+\infty} e^{it^2}\,\mathrm{d}t$ est convergente. On utilise le changement de variables $\varphi : u \in [1,+\infty[\mapsto t := \sqrt{u} \in [1,+\infty[$. On est ramené à l'étude de l'intégrale impropre $\displaystyle\int_1^{+\infty} \frac{e^{iu}}{2\sqrt{u}}\,\mathrm{d}u$. Nous avons vu dans l'exercice 19 de la page 621 que cette dernière était convergente.
On obtient la convergence des intégrales de Fresnel en prenant les parties réelles et imaginaires (on peut aussi la montrer en utilisant une intégration par parties). Le calcul des intégrales de Fresnel fait l'objet de l'exercice II.3.30 :

$$\int_0^{+\infty} \cos(t^2)\,\mathrm{d}t = \int_0^{+\infty} \sin(t^2)\,\mathrm{d}t = \sqrt{\frac{\pi}{8}}.$$

Un peu d'Histoire

Le physicien français Augustin Fresnel (1788-1827) a introduit les intégrales portant son nom dans son étude de la diffraction de la lumière au travers d'une fente rectangulaire. Le front d'onde correspondant est une spirale de paramétrage :

$$s \mapsto \left(a\int_0^s \cos t^2\,\mathrm{d}t, a\int_0^s \sin t^2\,\mathrm{d}t \right)$$

et les intégrales de Fresnel donnent la position de son « centre » : $a(\sqrt{\pi/8}, \sqrt{\pi/8})$. En réalité, cette spirale avait été découverte bien avant par Bernoulli (en 1694) dans l'étude des déformations d'une fine lame élastique, et par Euler. Elle est appelée spirale d'Euler ou de Cornu, ou clothoïde. Après Fresnel, la clothoïde a été utilisée par l'ingénieur américain Talbot pour des tracés de voies ferrées (1890). Aujourd'hui, elle est utilisée pour des « fonctions splines » et des dessins de police de caractères.

Exercice 22.

Retrouver le résultat de l'exercice 19 en utilisant une intégration par parties.

Solution. On applique la proposition ci-dessus, avec $a := \pi$, $b := +\infty$, $f : x \mapsto e^{ix}$ et $g : x \mapsto x^{-\alpha}$. On obtient :

$$-\alpha\int_\pi^{+\infty} \frac{e^{ix}}{x^{\alpha+1}}\,\mathrm{d}x = \left[\frac{e^{ix}}{x^\alpha}\right]_\pi^{+\infty} - i\int_\pi^{+\infty} \frac{e^{ix}}{x^\alpha}\,\mathrm{d}x.$$

L'intégrale impropre $\displaystyle\int_\pi^{+\infty} \frac{e^{ix}}{x^{\alpha+1}}\,\mathrm{d}x$ est (absolument) convergente et $\left[\frac{e^{ix}}{x^\alpha}\right]_\pi^{+\infty} = \frac{1}{\pi^\alpha}$, donc l'intégrale impropre $\displaystyle\int_\pi^{+\infty} \frac{e^{ix}}{x^\alpha}\,\mathrm{d}x$ est convergente.

7.3 Intégration des relations de comparaison

Pour compléter cette partie, nous allons examiner comment les relations de comparaison s'intègrent.

Proposition 114 (Intégration des relations de comparaison).
Soient $f : [a, b[\to \mathbb{R}$ et $g : [a, b[\to \mathbb{R}$ deux fonctions localement intégrables avec $g \geqslant 0$.

1. Si l'intégrale $\displaystyle\int_a^b g$ est divergente, alors en posant $F(x) := \int_a^x f$ et $G(x) := \int_a^x g$, on a :
$$f \underset{b}{=} \mathrm{O}(g) \Rightarrow F \underset{b}{=} \mathrm{O}(G), \quad f \underset{b}{=} \mathrm{o}(g) \Rightarrow F \underset{b}{=} \mathrm{o}(G) \quad \text{et} \quad f \underset{b}{\sim} g \Rightarrow F \underset{b}{\sim} G.$$
2. Si l'intégrale $\displaystyle\int_a^b g$ est convergente, alors en posant $F(x) := \int_x^b f$ et $G(x) := \int_x^b g$, on a :
$$f \underset{b}{=} \mathrm{O}(g) \Rightarrow F \underset{b}{=} \mathrm{O}(G), \quad f \underset{b}{=} \mathrm{o}(g) \Rightarrow F \underset{b}{=} \mathrm{o}(G) \quad \text{et} \quad f \underset{b}{\sim} g \Rightarrow F \underset{b}{\sim} G.$$

On a des énoncés similaires en remplaçant $[a, b[$ par $]a, b]$ ou par un intervalle non borné $[a, +\infty[$ ou $]-\infty, a]$.

Démonstration.

1. Comme g est positive, si $\displaystyle\int_a^b g$ diverge, alors $G(x) \to +\infty$ quand $x \to b^-$.
Si $f = \mathrm{O}(g)$, alors on peut trouver $c_0 \in]a, b[$ et une constante C telle que pour $t \in [c_0, b[$, on ait $-C\,g(t) \leqslant f(t) \leqslant C\,g(t)$. Alors :
$$-C\,\big(G(x) - G(c_0)\big) \leqslant F(x) - F(c_0) \leqslant C\,\big(G(x) - G(c_0)\big).$$
Comme $G(x) \to +\infty$ quand $x \to b^-$, cela montre que $F = \mathrm{O}(G)$.
Si $f = \mathrm{o}(g)$, alors pour tout $\varepsilon > 0$, on peut trouver $c_0 \in]a, b[$ tel que pour $t \in [c_0, b[$, on ait $-\varepsilon\,g(t) \leqslant f(t) \leqslant \varepsilon\,g(t)$. Alors, pour $x \in [c_0, b[$, on a :
$$-\varepsilon\,G(x) \leqslant -\varepsilon\big(G(x) - G(c_0)\big) \leqslant F(x) - F(c_0) \leqslant \varepsilon\big(G(x) - G(c_0)\big) \leqslant \varepsilon\,G(x).$$
Si x est suffisamment proche de b, $\big|F(c_0)\big| \leqslant \varepsilon\,G(x)$ et donc $F(x) \leqslant 2\varepsilon\,G(x)$. Cela montre que $F = \mathrm{o}(G)$.
Si $f \sim g$, alors $f - g = \mathrm{o}(g)$ et d'après ce qui précède, $F - G = \mathrm{o}(G)$. Cela montre que $F \sim G$.
2. Si $f = \mathrm{O}(g)$, alors on peut trouver $c_0 \in]a, b[$ et une constante C telle que pour $t \in [c_0, b[$, on ait $\big|f(t)\big| \leqslant C \cdot g(t)$. Alors :
$$-C\,G(x) \leqslant F(x) \leqslant C\,G(x).$$
Cela montre que $F = \mathrm{O}(G)$.
Si $f = \mathrm{o}(g)$, alors pour tout $\varepsilon > 0$, on peut trouver $c_0 \in]a, b[$ tel que pour $t \in [c_0, b[$, on ait $-\varepsilon\,g(t) \leqslant f(t) \leqslant \varepsilon\,g(t)$. Alors, pour $x \in [c_0, b[$, on a :
$$-\varepsilon\,G(x) \leqslant F(x) \leqslant \varepsilon\,G(x).$$

Cela montre que $F = \mathrm{o}(G)$.
Si $f \sim g$, alors $f - g = \mathrm{o}(g)$ et d'après ce qui précède, $F - G = \mathrm{o}(G)$. Cela montre que $F \sim G$.

Corollaire 115. Soient $f : [a,b[\to \mathbb{R}$ et $g : [a,b[\to \mathbb{R}$ deux fonctions localement intégrables avec $g \geqslant 0$. Si f et g sont équivalentes en b, alors les intégrales impropres $\displaystyle\int_a^b f$ et $\displaystyle\int_a^b g$ sont de même nature.

Corollaire 116 (Intégration des développements limités).
Soient I un intervalle réel et $f : I \to \mathbb{R}$ une application localement intégrable. On suppose que f admet, au voisinage de $a \in I$, le développement limité :

$$f(x) = \sum_{k=0}^{n} a_k (x-a)^k + \mathrm{o}\big((x-a)^n\big).$$

alors $F : x \mapsto \displaystyle\int_a^x f$ admet, au voisinage de a, le développement limité :

$$F(x) = \sum_{k=0}^{n} \frac{a_k}{k+1} (x-a)^{k+1} + \mathrm{o}\big((x-a)^{n+1}\big).$$

Démonstration. Soit $g := F(x) - \sum\limits_{k=0}^{n} \frac{a_k}{k+1}(x-a)^{k+1}$. On a $g'(x) = \mathrm{o}\big((x-a)^n\big)$.

Si a n'est pas le plus grand élément de I, il existe $b \in I$, $a < b$. La fonction g' est continue sur $[a,b]$ donc intégrable sur ce segment et, d'après le corollaire ci-dessus, quand $x \to a$, on a :

$$g(x) = \int_a^x g'(t)\,\mathrm{d}t = \mathrm{o}\Big(\int_a^x (t-a)^n\,\mathrm{d}t\Big) = \mathrm{o}\big((x-a)^{n+1}\big).$$

Exercice 23.

Soit $f : x \in [0,1] \mapsto \arcsin(1-x) \in \mathbb{R}$. Montrer que :

$$f(x) = \frac{\pi}{2} - \sqrt{2x} - \frac{\sqrt{2}}{12} x^{3/2} + \mathrm{o}(x^{3/2}).$$

Solution. La fonction f est de classe $\mathcal{C}^1$ sur $]0,1]$ et, quand $x \to 0^+$:

$$f'(x) = -\frac{1}{\sqrt{x(2-x)}} = -\frac{1}{\sqrt{2x}}\big(1 + \frac{x}{4} + \mathrm{o}(x)\big).$$

On pose, pour $x \in]0,1]$, $g(x) := f'(x) + \frac{1}{\sqrt{2x}} + \frac{\sqrt{x}}{4\sqrt{2}}$.
La fonction g se prolonge continûment à $[0,1]$ par $g(0) := 0$ et $g(t) = \mathrm{o}_0(\sqrt{t})$.
Alors $\displaystyle\int_0^x g(t)\,\mathrm{d}t = \mathrm{o}\left(\int_0^x \sqrt{t}\,\mathrm{d}t\right) = \mathrm{o}(x^{3/2})$. Le résultat s'en déduit.

Un peu d'Histoire

L'intégration des relations de comparaison permet de montrer que :

$$\mathrm{Li}(x) := \int_2^x \frac{\mathrm{d}t}{\ln t}\,\mathrm{d}t \underset{+\infty}{\sim} \frac{x}{\ln x}.$$

C'est l'objet de l'exercice II.3.23 de la page 654.
Cela est lié au problème de la *densité des nombres premiers*. Notons, pour tout $x \in \mathbb{R}$, par $\pi(x)$ le nombre de nombres premiers inférieurs à x. Gauß, alors âgé d'une quinzaine d'années, avait conjecturé vers 1792 que $\pi(x) \underset{+\infty}{\sim} \frac{x}{\ln x}$.
En 1896, De la Vallée Poussin et Hadamard ont démontré (en utilisant la fonction ζ de Riemann) que $\pi(x) \underset{+\infty}{\sim} \displaystyle\int_2^x \frac{\mathrm{d}t}{\ln t}$, ce qui justifie la conjecture de Gauß. Suivant von Koch (1906), une amélioration conjecturale (apparemment « un peu mineure » !) de ce résultat : $\pi(x) = \mathrm{Li}(x) + \mathrm{O}(\sqrt{x}\ln x)$ est équivalente à *l'hypothèse de Riemann* (sur les zéros de la fonction ζ) qui est sans doute le plus grand défi mathématique de ce siècle.

8 Théorèmes de convergence

8.1 Interversion des limites et des intégrales

Dans cette partie on étudie des suites de fonctions à valeurs *réelles*. Certains résultats s'étendent facilement au cas des applications à valeurs dans un $\mathbb{K}$-espace vectoriel normé de dimension finie (en utilisant des coordonnées et, dans le cas complexe, des parties réelles et imaginaires).
Nous ne traiterons pas le cas des applications à valeurs dans un espace de Banach.

8.1.1 Position du problème

Soit $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions $\mathscr{R}$-intégrables sur $[a,b]$ convergeant *simplement* vers une fonction $f : [a,b] \to \mathbb{R}$. Il est naturel de se demander si f est $\mathscr{R}$-intégrable, si la limite des $\displaystyle\int_a^b f_n$ quand n tend vers $+\infty$ existe et enfin si, en supposant ces deux conditions réalisées, l'on a : $\displaystyle\lim_{n\to+\infty} \int_a^b f_n = \int_a^b f$ (autrement dit si l'on peut *échanger la limite et l'intégrale*). La réponse à ces questions est en général *négative* :

- il est possible que la limite simple f ne soit pas $\mathscr{R}$-intégrable ;
- il est possible que la suite $\left(\int_a^b f_n\right)_{n\in\mathbb{N}}$ ne converge pas ;
- même si f est $\mathscr{R}$-intégrable et si la suite $\left(\int_a^b f_n\right)_{n\in\mathbb{N}}$ converge, il est possible que $\displaystyle\lim_{n\to+\infty} \int_a^b f_n \neq \int_a^b f$.

On le verra dans l'exercice et les exemples qui suivent.

Exercice 24 (La limite simple n'est pas intégrable).
Soit $f : \mathbb{R}^+ \to \mathbb{R}$ définie par $f(0) := 0$ et $f(x) := x^2 \sin(1/x^2)$ si $x \neq 0$.
Soit, pour tout $n \in \mathbb{N}^*$, la fonction $g_n : [0,1] \to \mathbb{R}$ définie par :

$$g_n(0) = 0 \qquad \text{et} \qquad g_n(x) = n\left(f(x + \frac{1}{n}) - f(x)\right) \quad \text{si} \quad x \neq 0.$$

Montrer que les fonctions g_n sont $\mathscr{R}$-intégrables sur $[0,1]$ et que la suite $(g_n)_{n\in\mathbb{N}^*}$ converge simplement vers une fonction g qui n'est pas $\mathscr{R}$-intégrable sur $[0,1]$.

Solution. Si l'on modifie la valeur de la fonction g_n en 0 en $nf(\frac{1}{n})$ on obtient une fonction continue, donc g_n est $\mathscr{R}$-integrable (*cf.* l'exercice 3).
La suite $(g_n)_{n\in\mathbb{N}^*}$ converge simplement vers la fonction $g : [0,1] \to \mathbb{R}$ définie par $g(0) := 0$ et $g(x) := f'(x) = 2x\sin(1/x^2) - \frac{2\cos(1/x^2)}{x}$ si $x \neq 0$. On vérifie que cette fonction n'est *pas bornée* sur $[0,1]$, elle n'est donc pas $\mathscr{R}$-intégrable sur cet intervalle.

Exemples.

1. *On ne peut pas permuter limite et intégration*
   Soit, pour tout $n \in \mathbb{N}$, la fonction f_n définie sur $[0,1]$ par : $f_n : x \mapsto n^2 x e^{-nx}$. La suite $(f_n)_{n\in\mathbb{N}}$ converge simplement vers 0 sur $[0,1]$ et l'on a $\displaystyle\int_0^1 f = 0$, mais

   $$\int_0^1 f_n(x)\,\mathrm{d}x = \int_0^1 nxe^{-nx}\,\mathrm{d}(nx) = \int_0^n te^{-t}\,\mathrm{d}t \to \int_0^{+\infty} te^{-t}\,\mathrm{d}t = 1$$

   quand $n \to +\infty$.

2. *On ne peut pas permuter limite et intégration*
   Soit, pour tout $n \in \mathbb{N}^*$, la fonction f_n définie sur $[0,1]$ par :

   $$\begin{cases} f_n(x) = \quad 4n^2x & \text{si } 0 \leqslant x \leqslant \frac{1}{2n} \\ f_n(x) = -4n^2x + 4n & \text{si } \frac{1}{2n} \leqslant x \leqslant \frac{1}{n} \\ f_n(x) = \quad 0 & \text{si } \frac{1}{n} \leqslant x \leqslant 1. \end{cases}$$

   Les fonctions f_n sont continues (affines par morceaux) et $\displaystyle\int_0^1 f_n = 1$. La suite de fonctions $(f_n)_{n\in\mathbb{N}^*}$ converge *simplement* vers la fonction nulle $f = 0$ et l'on a $\displaystyle\int_0^1 f = 0$.

   Ainsi l'intégrale de la limite f n'est pas égale à la limite des intégrales $\displaystyle\int_0^1 f_n$ qui est 1.

3. *La suite des intégrales ne converge pas toujours*
On modifie légèrement la suite de fonctions de l'exemple précédent.
Soit, pour tout $n \in \mathbb{N}^*$, la fonction g_n définie sur $[0,1]$ par :
$$\begin{cases} g_n(x) = \quad 4n^3x & \text{si } 0 \leqslant x \leqslant \frac{1}{2n} \\ g_n(x) = -4n^3x + 4n^2 & \text{si } \frac{1}{2n} \leqslant x \leqslant \frac{1}{n} \\ g_n(x) = \quad 0 & \text{si } \frac{1}{n} \leqslant x \leqslant 1. \end{cases}$$
Les fonctions f_n sont continues (affines par morceaux) et $\displaystyle\int_0^1 g_n = n$. La suite de fonctions $(g_n)_{n\in\mathbb{N}^*}$ converge *simplement* vers la fonction nulle $f = 0$ mais la suite $(n)_{n\in\mathbb{N}^*}$ diverge.

8.1.2 Convergence uniforme

Il est possible de donner des réponses positives aux questions ci-dessus en ajoutant des conditions de convergence uniforme, de « domination » (ou plus généralement « d'encadrement ») ou de « monotonie ».
Le cas le plus « élémentaire » est celui où la convergence de la suite $(f_n)_{n\in\mathbb{N}}$ est *uniforme* sur un segment.

Théorème 117. Si $(f_n : [a,b] \to \mathbb{R})_{n\in\mathbb{N}}$ est une suite de fonctions intégrables au sens de Riemann qui converge uniformément vers $f : [a,b] \to \mathbb{R}$, alors f est intégrable au sens de Riemann et $\displaystyle\int_a^b f = \lim_{n\to+\infty} \int_a^b f_n$.

Démonstration. Pour tout $\varepsilon > 0$ il existe un entier positif N tel que :
$$\forall n \geqslant N,\ \forall x \in [a,b],\ f_n(x) - \varepsilon \leqslant f(x) \leqslant f_n(x) + \varepsilon.$$
Par suite l'intégrale supérieure (resp. inférieure) de f est comprise entre les intégrales supérieures (resp. inférieure) de $f_n(x) - \varepsilon$ et de $f_n(x) + \varepsilon$:
$$\int_a^b f_n - \varepsilon(b-a) \leqslant \overset{*}{\int_a^b} f \leqslant \overset{*}{\int_a^b} f_n + \varepsilon(b-a) = \int_a^b f_n + \varepsilon(b-a)$$
$$\int_a^b f_n - \varepsilon(b-a) = \underset{*}{\int_a^b} f_n - \varepsilon(b-a) \leqslant \underset{*}{\int_a^b} f \leqslant \int_a^b f_n + \varepsilon(b-a).$$
D'où $\displaystyle\overset{*}{\int_a^b} f - \underset{*}{\int_a^b} f \leqslant 2\varepsilon(b-a)$, ce qui prouve la $\mathscr{R}$-intégrabilité de f.
On a, pour tout $n \in \mathbb{N}$:
$$\left| \int_a^b f - \int_a^b f_n \right| \leqslant \int_a^b |f - f_n| \leqslant (b-a)\,\|f - f_n\|_\infty \underset{n\to+\infty}{\longrightarrow} 0,$$
donc $\displaystyle\int_a^b f = \lim_{n\to+\infty} \int_a^b f_n$. ∎

Notons que la condition de convergence uniforme est *suffisante* pour permuter limite et intégration mais qu'elle n'est pas *nécessaire*.

Exemple. Pour tout $n \in \mathbb{N}$ on définit la fonction f_n sur $[0,1]$ par :

$$f_n : x \in [0,1] \mapsto x^n \in \mathbb{R}.$$

La suite de fonctions $(f_n)_{n\in\mathbb{N}}$ converge *simplement* vers la fonction f définie par :

$$f : x \in [0,1] \mapsto \begin{cases} 0 & \text{si} \quad x \in [0,1[\\ 1 & \text{si} \quad x = 1. \end{cases}$$

La limite *n'est pas uniforme* puisque f n'est pas continue en $x = 1$.

Mais $\displaystyle\int_0^1 f_n = \int_0^1 x^n \,\mathrm{d}x = \frac{1}{n+1}$ tend vers $0 = \displaystyle\int_0^1 f$ quand n tend vers $+\infty$.

Le résultat du théorème 117 ne se généralise pas sur un intervalle non borné.

Exemple. Soit $f_n : \mathbb{R} \to \mathbb{R}$ la fonction qui vaut $1/n$ sur l'intervalle $[0,n]$ et 0 en dehors. Alors, l'intégrale de f_n sur $\mathbb{R}$ vaut 1. La suite (f_n) converge uniformément vers la fonction identiquement nulle, dont l'intégrale sur $\mathbb{R}$ vaut 0. On ne peut donc pas échanger limite et intégrale.

Quand on remplace le segment $[a,b]$ par un intervalle arbitraire la condition de convergence uniforme devient trop forte pour beaucoup d'applications. On peut y remédier en utilisant la convergence uniforme sur tout sous-segment et une condition de domination.

Proposition 118 (Convergence dominée « du pauvre »).
Soit $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions *absolument intégrables* au sens de Riemann sur un intervalle réel I. On suppose :

(i) que la suite $(f_n)_{n\in\mathbb{N}}$ converge *uniformément* sur tout segment $K \subset I$ vers une fonction f ;

(ii) qu'il existe une fonction $g \geqslant 0$ définie et *intégrable au sens de Riemann* sur I telle que l'on ait $\big|f_n(x)\big| \leqslant g(x)$ pour tout $n \in \mathbb{N}$ et tout $x \in I$.

Alors f est *absolument intégrable* au sens de Riemann sur I et $\displaystyle\int_I f = \lim_{n\to+\infty} \int_I f_n$.

Démonstration. La suite $(f_n)_{n\in\mathbb{N}}$ converge *simplement* vers f sur I, donc, d'après (ii), $|f(x)| \leqslant g(x)$ pour tout $x \in I$. Soit $\varepsilon > 0$, puisque $g \geqslant 0$ est intégrable, il existe un segment $K \subset I$ tel que $\displaystyle\int_{I\setminus K} g < \varepsilon/3$. On en déduit $\displaystyle\int_{I\setminus K} |f_n| < \varepsilon/3$, pour tout $n \in \mathbb{N}$ et $\displaystyle\int_{I\setminus K} |f| < \varepsilon/3$. Par ailleurs la convergence est *uniforme* sur le segment K, il existe donc $N \in \mathbb{N}$ tel que, pour tout $n \geqslant N$, l'on ait $|f - f_n| < \frac{\varepsilon}{3 \operatorname{long} K}$.

On a donc, pour tout $n \geqslant N$:

$$\begin{aligned} \left|\int_I f - \int_I f_n\right| &\leqslant \left|\int_K f - \int_K f_n\right| + \left|\int_{I\setminus K} f - \int_{I\setminus K} f_n\right| \\ &\leqslant \int_K |f - f_n| + \int_{I\setminus K} |f| + \int_{I\setminus K} |f_n| < \varepsilon. \end{aligned}$$

∎

Nous avons emprunté la terminologie *Convergence dominée « du pauvre »* à R. Godement (*cf.* [G2], pages 110–113) : le pauvre n'a pas de très bons outils à sa disposition, il doit donc travailler plus que le riche qui dispose du théorème de convergence dominée énoncé plus loin (*cf.* le théorème 121).

8.1.3 Vers les grands théorèmes de la théorie de l'intégration

Dans bien des cas, la convergence de la suite $(f_n)_{n\in\mathbb{N}}$ n'est pas uniforme (nous en avons vu un exemple simple plus haut) même si l'on se restreint aux sous-intervalles compacts, ou bien l'uniformité est difficile à prouver… Si l'on reste dans le cadre de la théorie de Riemann, montrer que l'on peut échanger la limite et l'intégrale peut devenir très laborieux. C'est en particulier pour combler cette lacune que Lebesgue a introduit une nouvelle définition de l'intégrale au début du XX$^{\text{e}}$ siècle. Nous allons donner trois théorèmes permettant d'intervertir limites et intégrales (convergence encadrée, convergence dominée, convergence monotone). Les *énoncés* ne font intervenir que la notion d'intégrale de Riemann (éventuellement généralisée), ils sont très simples à formuler et faciles à appliquer. Malheureusement, les *preuves* de ces théorèmes sont difficiles et hors de portée des techniques dont nous disposons ici. Nous les admettrons provisoirement et nous en donnerons des démonstrations (basées sur l'intégrale de Henstock-Kurzweil) dans le module *Intégration* de [L2-L3]. On peut aussi prouver une partie de ces résultats (convergence dominée et une version restreinte de la convergence monotone) en utilisant la théorie de l'intégrale de Lebesgue. En revanche, cette théorie ne permet pas de prouver le théorème de convergence encadrée.

Les hypothèses des trois théorèmes ont une partie commune que l'on devra vérifier dans tous les cas :

(i) les fonctions f_n ($n \in \mathbb{N}$) sont *intégrables au sens de Riemann* (éventuellement généralisé) sur I et la suite $(f_n)_{n\in\mathbb{N}}$ converge *simplement* vers f sur I ;

(ii) la fonction limite f est *localement intégrable* sur I.

Dans le cadre de la théorie de Henstock-Kurzweil (resp. de Lebesgue) la condition (ii) *disparaît*. En effet, avec les théorèmes parallèles dans ce cadre, on obtient « gratuitement » que f est intégrable au sens de la théorie. Plus on est savant et moins l'on a d'hypothèses à vérifier pour intervertir limites et intégrales !

Théorème 119 (Convergence monotone). Soient I un intervalle réel et $(f_n)_{n\in\mathbb{N}}$ une suite *croissante* de fonctions *intégrables au sens de Riemann* (éventuellement généralisé) sur I qui converge *simplement* sur I vers une fonction f *localement intégrable* sur I.

La fonction f est *intégrable au sens de Riemann* (éventuellement généralisé) sur I si, et seulement si, la suite $\left(\int_I f_n\right)_{n\in\mathbb{N}}$ est *majorée*, et alors :

$$\int_I f = \lim_{n\to+\infty} \int_I f_n.$$

Remarque. On a bien entendu un résultat entièrement analogue pour les suites *décroissantes* de fonctions, obtenu en remplaçant $(f_n)_{n\in\mathbb{N}}$ par $(-f_n)_{n\in\mathbb{N}}$.

Théorème 120 (Convergence encadrée). Soient I un intervalle réel et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions *intégrables au sens de Riemann* (éventuellement généralisé) sur I. On suppose que :

(i) la suite $(f_n)_{n\in\mathbb{N}}$ converge *simplement* sur I vers une fonction f *localement intégrable* ;

(ii) il existe deux fonctions g et h définies et *intégrables au sens de Riemann* (éventuellement généralisé) sur I telle que l'on ait $g(x) \leqslant f_n(x) \leqslant h(x)$ pour tout $n \in \mathbb{N}$ et tout $x \in I$.

Alors f est *intégrable au sens de Riemann* (éventuellement généralisé) sur I et :

$$\int_I f = \lim_{n\to+\infty} \int_I f_n.$$

Si $I := [a, b]$ la condition d'encadrement (ii) est équivalente à supposer que la famille $(f_n)_{n\in\mathbb{N}}$ est *uniformément bornée* (c'est-à-dire qu'il existe $M \geqslant 0$ tel que, pour tout $n \in \mathbb{N}$, et tout $x \in [a, b]$, $\left|f_n(x)\right| \leqslant M$).

En corollaire du théorème de convergence encadrée, on a le théorème de convergence dominée.

Théorème 121 (Convergence dominée).

Soient I un intervalle réel et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions *intégrables au sens de Riemann* (éventuellement généralisé) sur I. On suppose que :

(i) la suite $(f_n)_{n\in\mathbb{N}}$ converge *simplement* sur I vers une fonction f *localement intégrable* sur I ;

(ii) il existe une fonction g définie et *absolument intégrable* sur I telle que l'on ait $\left|f_n(x)\right| \leqslant g(x)$ pour tout $n \in \mathbb{N}$ et tout $x \in I$.

Alors les f_n $(n \in \mathbb{N})$ et f sont *absolument intégrables* sur I et :

$$\int_I f = \lim_{n\to+\infty} \int_I f_n.$$

Dans l'énoncé du théorème ci-dessus, on peut remplacer l'hypothèse « $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions *intégrables au sens de Riemann* (éventuellement généralisé) sur I » par l'hypothèse (apparemment) plus faible « $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions définies et *localement intégrables* sur I ». Cela revient au même puisque $\left|f_n(x)\right| \leqslant g(x)$.

On a $\left|f_n\right| \leqslant g$ si, et seulement si, $-g \leqslant f_n \leqslant g$. Le théorème ci-dessus est donc un cas particulier du précédent. Il est « apparemment » moins général (*cf.* la remarque 2 ci-dessous) car il ne s'applique qu'aux cas où les fonctions f_n sont *absolument intégrables* sur I. C'est l'analogue pour l'intégration d'un résultat classique sur les suites de séries normalement convergentes.

Remarques.

1. Dans le cas de l'exemple 2 de la page 627 il est impossible de trouver une fonction g satisfaisant (ii). En effet une telle fonction doit être *bornée* tandisque $f_n(\frac{1}{2n}) = 2n$. Le lecteur vérifiera qu'il en est de même pour l'exemple 2.

2. En fait le théorème de convergence dominée est *équivalent* au théorème de convergence encadrée. En effet $g \leqslant f_n \leqslant h$ équivaut à $0 \leqslant f_n - g \leqslant h - g$ et cette dernière condition implique $|f_n - g| \leqslant h - g$. Les hypothèses du théorème de convergence encadrée pour la suite (f_n) et sa limite simple f entraînent donc celles du théorème de convergence dominée pour la suite de fonctions $(f_n - g)_{n\in\mathbb{N}}$ et sa limite $f - g$.

   Cela étant, il est facile de fabriquer un exemple tel que le théorème de convergence dominée ne s'applique pas *directement* tandis que le théorème de convergence encadré s'applique. On part de g intégrable mais non absolument intégrable et d'une suite de fonctions intégrables *positives* φ_n convergeant simplement vers φ et majorées par une fonction intégrable ψ. Alors la suite $(f_n := g + \varphi_n)_{n\in\mathbb{N}}$ est encadrée : $g \leqslant f_n \leqslant g + \psi$, mais les f_n ne sont pas absolument intégrables (sinon g le serait puisque $|g| \leqslant |f_n| + \varphi_n)$.

3. On vérifie facilement que le théorème de convergence dominée reste vrai pour des applications à valeurs dans un $\mathbb{K}$-espace vectoriel normé de dimension finie. Il est encore vrai pour des applications à valeurs dans un $\mathbb{K}$-espace de Banach, mais nous ne le démontrerons pas dans ce cours.

4. On trouve dans certains cours les énoncés moins généraux des théorèmes 121 et 119 obtenus en remplaçant partout « localement intégrable » par « continue par morceaux ». On peut prouver ces résultats de façon élémentaire (*cf.* par exemple le livre [DW2], Chap. 15).

Si I est un segment la condition de domination (ii) est équivalente à dire que les f_n sont *uniformément bornées*. On déduit alors de la proposition précédente le résultat suivant dû au mathématicien italien Cesare Arzelà (1885).

Corollaire 122 (Théorème d'Arzela).
Soit $(f_n)_{n\in\mathbb{N}}$ une suite uniformément bornée de fonctions de $\mathscr{R}\big([a,b]\big)$ convergeant *simplement* vers une fonction $f \in \mathscr{R}\big([a,b]\big)$, alors :

$$\int_a^b f = \lim_{n\to+\infty} \int_a^b f_n.$$

On peut donner une preuve directe « très élémentaire » du théorème d'Arzela, *cf.* l'exercice II.3.32[1].
Le résultat ci-dessus permet de prouver un résultat énoncé plus haut.

Preuve du théorème 96 de la page 607 Par continuité, il suffit de prouver le résultat pour tous c, d tels que $a < c < d < b$. Soit $x \in [a,b]$. D'après le théorème des accroissements finis, pour tout $h \neq 0$ assez petit ($|h| \leqslant \inf(c-a, b-d)$), il existe $\theta = \theta(x,h) \in [0,1]$ tel que $f_h(x) = \frac{f(x+h)-f(x)}{h} = f'(x+\theta h)$. Ainsi $\big|f_h(x)\big| \leqslant M := \sup\limits_{x\in[a,b]} \big|f'(x)\big|$. Cette borne est

[1] Cette preuve est due à Jonathan W. Lewin, *A truly Elementary Approach to the Bounded Convergence Theorem*, American Mathematical Monthly, vol. 93, num. 5, 1986, pages 395 – 397.

uniforme en x et h. En utilisant le théorème d'Arzela, on obtient :

$$\int_c^d f' = \lim_{h\to 0,\ h\neq 0} \int_c^d f_h.$$

On a :

$$\int_c^d f_h = \frac{1}{h}\int_{c+h}^{d+h} f - \frac{1}{h}\int_c^d f = \frac{1}{h}\int_d^{d+h} f - \frac{1}{h}\int_c^{c+h} f.$$

Donc, en utilisant la formule de la moyenne :

$$\lim_{h\to 0, h\neq 0} \int_c^d f_h = f(d) - f(c) \quad \text{et} \quad \int_c^d f' = f(d) - f(c).$$

Exemple. Soit $g : [0,1] \to \mathbb{R}$ une fonction $\mathscr{R}$-intégrable sur $[0,1]$. Pour tout $n \in \mathbb{N}$, on définit une fonction $f_n : x \in [0,1] \mapsto \frac{g(x)}{1+nx}$.
Ces fonctions sont $\mathscr{R}$-intégrables (produit de fonctions $\mathscr{R}$-intégrables) et convergent simplement vers la fonction $f : x \in [0,1] \mapsto \begin{cases} g(0) & \text{si} \quad x = 0 \\ 0 & \text{si} \quad x \neq 0. \end{cases}$
On a $|f_n| \leqslant |g|$, la fonction $|g|$ est $\mathscr{R}$-intégrable sur $[0,1]$, on peut donc appliquer le théorème d'Arzela et on obtient $\lim\limits_{n\to+\infty} \int_0^1 f_n = \int_0^1 f = 0$.
On notera que, si $g(0) \neq 0$, la fonction f n'est pas continue. La convergence n'est donc pas uniforme et l'on ne peut pas utiliser le théorème 117 de la page 628.
On peut toutefois éviter d'utiliser le théorème d'Arzela en utilisant un argument très simple. La fonction g est bornée : $|g| \leqslant M$, donc $|f_n| \leqslant \frac{M}{1+nx}$ et :

$$\left|\int_0^1 f_n\right| \leqslant \int_0^1 |f_n| \leqslant \int_0^1 \frac{M}{1+nx}\,\mathrm{d}x = M\frac{\ln(n+1)}{n}.$$

Puisque $\lim\limits_{n\to+\infty} M\frac{\ln(n+1)}{n} = 0$, $\lim\limits_{n\to+\infty} \int_0^1 f_n = \int_0^1 f = 0$.

Exercice 25.

Pour tout $n \in \mathbb{N}^*$, on définit la fonction $f_n : x \in [0,1] \mapsto \frac{n(x^3+x)}{nx+1}e^{-x}$.

i) Montrer que la suite (f_n) est simplement convergente sur $[0,1]$ mais non uniformément convergente.

ii) Montrer que la suite $(I_n := \int_0^1 f_n)$ est convergente et déterminer sa limite ;

Solution.

i) Pour tout $n \in \mathbb{N}^*, f_n(0) = 0$. Pour tout $x > 0$ $\lim\limits_{n\to+\infty} f_n(x) = (1+x^2)e^{-x}$. La suite (f_n) converge donc simplement vers la fonction f telle que :

$$\forall x \in]0,1],\, f(x) = (1+x^2)e^{-x}, \text{et } f(0) = 0.$$

La fonction f n'est pas continue en 0, la suite (f_n) n'est donc pas uniformément convergente sur $[0,1]$.

ii) La fonction f est continue par morceaux sur $[0,1]$ donc intégrable, de plus pour tout $n \in \mathbb{N}^*$ et tout $x \in]0,1]$, $f_n(x) \leqslant (1+x^2)e^{-x} \leqslant 2$. D'après le théorème d'Arzela (*cf.* le corollaire 122) on a :

$$\lim_{n\to+\infty} I_n = \int_0^1 f = 3 - \frac{6}{e}\cdot$$

Exercice 26 (Intégrale de Gauß).

Pour $n \in \mathbb{N}^*$, posons $f_n(x) := \begin{cases} \left(1-\frac{x^2}{n}\right)^n & \text{si } x \in [0,\sqrt{n}] \\ 0 & \text{si } x > \sqrt{n}. \end{cases}$

1. Montrer que la suite $(f_n)_{n\in\mathbb{N}^*}$ converge simplement vers $x \mapsto e^{-x^2}$ sur $[0,+\infty[$.
2. Montrer que :
$$\int_0^{+\infty} e^{-x^2}\,\mathrm{d}x = \lim_{n\to+\infty} \int_0^{\sqrt{n}} \Big(1-\frac{x^2}{n}\Big)^n \mathrm{d}x.$$
En déduire $\displaystyle\int_0^{+\infty} e^{-x^2}\,\mathrm{d}x = \frac{\sqrt{\pi}}{2}$ en utilisant la formule de Wallis (*cf.* page 890).

Solution.

1. Soit $x \geqslant 0$. Pour $n > x^2$, on a $f_n(x) = \left(1-\frac{x^2}{n}\right)^n$, donc $\lim\limits_{n\to+\infty} f_n(x) = e^{-x^2}$.
2. On a, pour tout $n \in \mathbb{N}^*$ et pour tout $x \geqslant 0$, $0 \leqslant f_n(x) \leqslant e^{-x^2}$.
C'est évident si $x \geqslant \sqrt{n}$, sinon l'on a : $f_n(x) = e^{n\ln\left(1-\frac{x^2}{n}\right)} \leqslant e^{-x^2}$.
La fonction continue positive $x \mapsto e^{-x^2}$ est intégrable au sens de Riemann généralisé sur $[0,+\infty[$. Les fonctions f_n sont continues et intégrables au sens de Riemann généralisé, on peut donc appliquer le théorème de convergence dominée, d'où :
$$\int_0^{+\infty} e^{-x^2}\,\mathrm{d}x = \lim_{n\to+\infty} \int_0^{+\infty} f_n = \lim_{n\to+\infty} J_n \quad \text{où} \quad J_n := \int_0^{\sqrt{n}} \Big(1-\frac{x^2}{n}\Big)^n \mathrm{d}x.$$
En utilisant le changement de variable $x = \sqrt{n}\,\sin t$, on obtient :
$$J_n = \sqrt{n}\int_0^{\pi/2} \cos^{2n+1}t\,\mathrm{d}t.$$
Donc, en utilisant $\displaystyle\int_0^{\pi/2} \cos^m t\,\mathrm{d}t \sim \sqrt{\frac{\pi}{2m}}$ quand $m \to +\infty$, on obtient :
$$\lim_{n\to+\infty} J_n = \frac{\sqrt{\pi}}{2} \qquad \text{d'où} \qquad \int_0^{+\infty} e^{-x^2}\,\mathrm{d}x = \frac{\sqrt{\pi}}{2}\cdot$$

Exercice 27 (Formule de Gauß).

Pour $x > 0$, posons $f_n(t) := \begin{cases} t^{x-1}(1-t/n)^n & \text{si } t \in]0,n[\\ 0 & \text{si } t \geqslant n. \end{cases}$

1. Montrer que $\displaystyle\lim_{n\to+\infty} \int_0^{+\infty} f_n(t)\,\mathrm{d}t = \Gamma(x)$ avec $\displaystyle\Gamma(x) := \int_0^{+\infty} t^{x-1}e^{-t}\,\mathrm{d}t$.

2. En déduire que pour tout $x > 0$:

$$\Gamma(x) = \lim_{n\to+\infty} n^x \int_0^1 u^{x-1}(1-u)^n \,\mathrm{d}u.$$

3. En utilisant des intégrations par parties successives, montrer que pour tout $x > 0$:

$$\Gamma(x) = \lim_{n\to+\infty} \frac{n!n^x}{x(x+1)\cdots(x+n)}.$$

4. Calculer $\Gamma(1/2)$. En déduire $\displaystyle\int_0^{+\infty} e^{-x^2}\,\mathrm{d}x = \frac{\sqrt{\pi}}{2}$ en utilisant la formule de Stirling (*cf.* le théorème 4 de la page 892).

Solution.

1. Observons d'abord que, pour tout $t > 0$, $f_n(t) \to f(t) := t^{x-1}e^{-t}$ quand $n \to +\infty$. Ensuite, pour tout $n \geqslant 1$ et tout $t \in]0, n]$, on a $\ln(1-t/n) \leqslant -t/n$ et donc :

$$\forall t > 0, \quad 0 \leqslant f_n(t) \leqslant t^{x-1}e^{n\ln(1-t/n)} \leqslant t^{x-1}e^{-t} = f(t).$$

La fonction f est intégrable sur $]0, +\infty[$ (voir exercice 16 de la page 614). Par conséquent, d'après le théorème de convergence dominée, l'on a :

$$\lim_{n\to+\infty} \int_0^{+\infty} f_n = \int_0^{+\infty} f = \Gamma(x).$$

2. En faisant le changement de variable $u = t/n$, on a :

$$\int_0^{+\infty} f_n = \int_0^n t^{x-1}(1-t/n)^n \,\mathrm{d}t = n^x \int_0^1 u^{x-1}(1-u)^n \,\mathrm{d}u.$$

3. Pour $n \geqslant 1$, on a :

$$\int_0^1 u^{x-1}(1-u)^n \,\mathrm{d}u = \underbrace{\left[\frac{u^x}{x}(1-u)^n\right]_0^1}_{=0} + n\int_0^1 \frac{u^x}{x}(1-u)^{n-1} \,\mathrm{d}u.$$

Par récurrence, on obtient :

$$\begin{aligned}\int_0^1 u^{x-1}(1-u)^n \,\mathrm{d}u &= \frac{n!}{x(x+1)\cdots(x+n-1)} \int_0^1 u^{x+n-1}(1-u)^0 \,\mathrm{d}u \\ &= \frac{n!}{x(x+1)\cdots(x+n)}.\end{aligned}$$

4. On a :

$$\Gamma(1/2) = \lim_{n\to+\infty} 2^{n+1}\frac{n!\sqrt{n}}{1(1+2)\cdots(1+2n)} = \lim_{n\to+\infty} \frac{2^{2n+1}(n!)^2\sqrt{n}}{(2n+1)!}.$$

Quand $n \to +\infty$, on a, d'après la formule de Stirling :

$$(n!)^2 \sim 2\pi n\Big(\frac{n}{e}\Big)^{2n} \quad \text{et} \quad (2n+1)! \sim \sqrt{2\pi(2n+1)}\Big(\frac{2n+1}{e}\Big)^{2n+1}.$$

D'où :

$$\frac{2^{2n+1}(n!)^2\sqrt{n}}{(2n+1)!} \sim \frac{e\sqrt{\pi}}{\left(1+\frac{1}{2n}\right)^{2n+1}} \sim \sqrt{\pi}.$$

Ainsi $\Gamma(1/2) = \sqrt{\pi}$.

Par ailleurs $\Gamma(1/2) = \displaystyle\int_0^{+\infty} e^{-t}\frac{\mathrm{d}t}{\sqrt{t}} = 2\int_0^{+\infty} e^{-u^2}\,\mathrm{d}u$, en utilisant le changement de variables $t = u^2$. On retrouve ainsi $\displaystyle\int_0^{+\infty} e^{-u^2}\,\mathrm{d}u = \sqrt{\pi}/2$.

Remarque. On peut traiter les deux exercices ci-dessus en utilisant, à la place du théorème de convergence dominée la version « du pauvre » (proposition 118), mais il faut alors travailler un peu plus...

Le lecteur pourra se reporter par exemple à [G2], exemple 1, page 111.

Exercice 28.

Soit $U \subset [0,1]$ un *ouvert*. On suppose que U n'est pas une réunion finie d'intervalles ouverts disjoints, on sait qu'il existe alors une famille dénombrable d'intervalles ouverts disjoints $(I_n)_{n\in\mathbb{N}}$ telle que $U = \bigcup_{n\in\mathbb{N}} I_n$ (*cf.* la proposition 22 de la page 563). On suppose que la fonction caractéristique 1_U est $\mathscr{R}$-intégrable sur $[0,1]$, calculer $\displaystyle\int_0^1 1_U$ en utilisant la famille $(I_n)_{n\in\mathbb{N}}$.

Solution. On ne fait d'abord aucune hypothèse sur l'ouvert U. Pour tout $k \in \mathbb{N}$, on note $f_k := \displaystyle\sum_{n=0}^{k} 1_{I_n}$. La suite de fonctions en escalier $(f_k)_{k\in\mathbb{N}}$ est évidemment croissante et majorée par 1. Elle converge *simplement* vers 1_U. On a, pour tout $k \in \mathbb{N}$, $\displaystyle\int_0^1 f_k = \sum_{n=0}^{k} \operatorname{long} I_n \leqslant \int_0^1 1 = 1$, la série à termes positifs $\displaystyle\sum_{n=0}^{+\infty} \operatorname{long} I_n$ est donc convergente.

Si 1_U est $\mathscr{R}$-intégrable, on peut appliquer le théorème de convergence monotone (théorème 119) à la suite $(f_k)_{k\in\mathbb{N}}$, donc $\displaystyle\int_0^1 1_U = \lim_{k\to+\infty}\int_0^1 f_k = \sum_{n=0}^{+\infty} \operatorname{long} I_n$. On retrouve ainsi le résultat de la proposition 26 de la page 566.

Exemple. Soit $g : [0,1] \to \mathbb{R}$ une fonction *continue*.

On se demande si la fonction $J : t \in \mathbb{R} \mapsto J(t) := \displaystyle\int_0^1 \sqrt{g^2(x)+t^2}\,\mathrm{d}x$ est dérivable à droite en $t = 0$ et, si c'est le cas, on voudrait calculer la dérivée.

On considère, pour tout $t > 0$, la fonction

$$f(t,.) : x \in [0,1] \mapsto f(t,x) := \frac{\sqrt{g^2(x)+t^2} - \big|g(x)\big|}{t} \in \mathbb{R}$$

Soit $x \in [0,1]$ fixé.

(i) Si $g(x) \neq 0$, on a $0 \leqslant f(t,x) \leqslant \frac{t}{2|g(x)|}$, donc $\lim\limits_{t\to 0^+} f(t,x) = 0$.

(ii) Si $g(x) = 0$, $f(t,x) = 1$, donc $\lim\limits_{t\to 0^+} f(t,x) = 1$.

Si g ne s'annule pas sur $[0,1]$, $\inf\limits_{x\in[0,1]} \big|g(x)\big| = m > 0$ et $0 \leqslant f(t,x) \leqslant \frac{t}{2m}$.

Donc $\displaystyle\int_0^1 f(t,x)\,\mathrm{d}x \leqslant \frac{t}{2m}$ et $\displaystyle\int_0^1 f(t,x)\,\mathrm{d}x = \frac{J(t)-J(0)}{t}$ tend vers 0 quand t tend vers 0^+. Ainsi J est dérivable à droite en 0 et sa dérivée est nulle.

Le cas où g peut s'annuler est beaucoup plus difficile. Notons $K \subset [0,1]$ l'ensemble des zéros de la fonction g. C'est un compact. Soit $(t_n)_{n\in\mathbb{N}}$ une suite de réels strictement positifs tendant vers 0. Pour tout $n \in \mathbb{N}$, on pose $f_n := x \mapsto f(t_n,x)$. La suite $(f_n)_{n\in\mathbb{N}}$ converge *simplement* vers la fonction indicatrice du compact K notée 1_K.

On a $\sqrt{g^2(x)+t^2} \leqslant \big|g(x)\big| + t$ donc $0 \leqslant f_n(x) \leqslant 1$. Si la fonction 1_K est Riemann intégrable sur $[0,1]$, on peut appliquer le théorème d'Arzela (théorème 122) et l'on obtient :

$$\lim_{n\to+\infty} \int_0^1 f_n = \int_0^1 f = \int_0^1 1_K.$$

Puisque, pour tout $t > 0$:

$$\int_0^1 f_n = \frac{J(t_n)-J(0)}{t_n},$$

on en déduit que $\lim\limits_{t\to 0^+} \frac{J(t)-J(0)}{t} = \displaystyle\int_0^1 1_K$. Ainsi J est dérivable à droite en 0 et sa dérivée est $\displaystyle\int_0^1 1_K$.

Dans la situation de l'exemple ci-dessus, c'est-à-dire $K := g^{-1}(0)$, il peut malheureusement arriver que 1_K ne soit pas Riemann intégrable. Nous l'admettrons. On peut le prouver en utilisant les deux résultats suivants :

1. Pour tout compact $K \subset [0,1]$ il existe une fonction continue g sur $[0,1]$ dont K est l'ensemble des zéros.
2. Il existe des compacts de $[0,1]$ dont la fonction caractéristique n'est pas $\mathscr{R}$-intégrable (*cf.* les exercices II.3.5 et II.3.6)

8.1.4 Interversion des sommations et des intégrales

Des théorèmes de convergence uniforme, monotone et dominée, on déduit facilement des théorèmes permettant *d'échanger série et intégrale*.

Commençons par le résultat le plus élémentaire utilisant la convergence uniforme.

Proposition 123. Soit $\sum f_k$ une série de fonctions *positives* et intégrables au sens de Riemann (éventuellement généralisé) sur un intervalle I. On suppose que cette série converge *uniformément* sur tout segment de I. Alors sa somme F est *localement intégrable sur* I et :

$$\sum_{k=0}^{+\infty}\left(\int_I f_k\right) < +\infty \quad \Longrightarrow \quad \sum_{k=0}^{+\infty}\left(\int_I f_k\right) = \int_I \left(\sum_{k=0}^{+\infty} f_k\right).$$

Démonstration.

Soit $[c,d]$ un segment de I. On applique le théorème de convergence uniforme (*cf.* le théorème 117) à la suite des fonctions $F_n := \sum_{k=0}^{n} f_k$ restreintes à $[c,d]$, ce qui prouve que la fonction F est intégrable sur $[c,d]$ et que l'on a $\sum_{k=0}^{+\infty}\left(\int_c^d f_k\right) = \int_c^d \left(\sum_{k=0}^{+\infty} f_k\right)$.

Si, de plus, $\sum_{k=0}^{+\infty}\left(\int_I f_k\right) < +\infty$, on a :

$$\int_c^d F = \int_c^d \left(\sum_{k=0}^{+\infty} f_k\right) = \sum_{k=0}^{+\infty}\left(\int_c^d f_k\right) \leqslant \sum_{k=0}^{+\infty}\left(\int_I f_k\right) < +\infty$$

donc F est intégrable sur I (*cf.* la proposition 102) et :

$$\int_c^d F = \sum_{k=0}^{+\infty}\left(\int_c^d f_k\right) \leqslant \sum_{k=0}^{+\infty}\left(\int_I f_k\right) = \int_I F,$$

d'où le résultat. ∎

Proposition 124. Soit $\sum f_k$ une série de fonctions *positives* et intégrables au sens de Riemann (éventuellement généralisé) sur un intervalle I, de somme $F : I \to \mathbb{R}$ *localement intégrable sur* I. Alors :

$$\sum_{k=0}^{+\infty}\left(\int_I f_k\right) < +\infty \quad \Longrightarrow F \text{ est intégrable sur } I \text{ et } \quad \sum_{k=0}^{+\infty}\left(\int_I f_k\right) = \int_I \left(\sum_{k=0}^{+\infty} f_k\right).$$

Démonstration. On applique le théorème de convergence monotone à la suite croissante de fonctions $F_n := \sum_{k=0}^{n} f_k$ qui converge vers $F := \sum_{k=0}^{+\infty} f_k$. ∎

Exercice 29.

i) Pour quelles valeurs du réel x, la fonction $f : t \in \,]0,+\infty[\to \frac{t^x}{e^t-1}$ est-elle intégrable sur $I := \,]0,+\infty[$?

ii) Montrer que pour tout $t > 0$, on a : $f(t) = \sum_{k=1}^{+\infty} t^x e^{-kt}$.

ii) Montrer que pour tout $k \in \mathbb{N}^*$ la fonction $f_k : t \in I \to t^x e^{-kt}$ est intégrable sur I et que la série $\sum \int_I f_k$ est convergente.

iv) Montrer que pour tout $x > 0$, $\int_I f = \Gamma(x+1)\zeta(x+1)$.

Solution.

i) Pour tout $x \in \mathbb{R}$, la fonction f est continue sur I. Au voisinage de 0, on a $f(t) \sim t^{x-1}$, au voisinage de $+\infty$, on a $f(t) = o(t^{-2})$. Par suite f est intégrable sur I si, et seulement si, $x > 0$.

ii) Pour tout $t > 0$:

$$f(t) = t^x e^{-t} \frac{1}{1-e^{-t}} = t^x e^{-t} \sum_{k=0}^{+\infty} e^{-kt} = \sum_{k=1}^{+\infty} t^x e^{-kt}.$$

iii) Pour tout $k \geqslant 1$, f_k est intégrable sur I (elle est continue sur $[0, +\infty[$ et au voisinage de $+\infty$, $f_k(t) = o(t^{-2})$). De plus :

$$\int_I f_k = \frac{1}{k^{x+1}} \int_I u^x e^{-u} \, \mathrm{d}u = \frac{1}{k^{x+1}} \Gamma(x+1).$$

iv) La série $\sum \frac{1}{k^{x+1}}$ est convergente de somme $\zeta(x+1)$. La proposition 124 permet d'affirmer :

$$\forall x > 0, \quad \int_0^{+\infty} \frac{t^x}{e^t - 1} = \Gamma(x+1)\,\zeta(x+1).$$

Proposition 125. Soit E un $\mathbb{K}$-espace vectoriel normé de dimension finie. Soit $(f_k)_{k\in\mathbb{N}}$ une suite d'applications de I dans E, localement intégrables et de normes intégrables sur un intervalle I. On suppose :

(i) que la série $\sum f_k$ converge *simplement* vers $F : I \to E$;

(ii) que la fonction F est *localement intégrable* sur I ;

(iii) que la série numérique $\sum \left(\int_I \|f_k\| \right)$ est convergente.

Alors F et $\|F\|$ sont intégrables au sens de Riemann (éventuellement généralisé) sur I et :

$$\int_I F = \int_I \left(\sum_{k=0}^{+\infty} f_k \right) = \sum_{k=0}^{+\infty} \left(\int_I f_k \right). \tag{28}$$

Démonstration. On traite d'abord le cas plus simple où la série $\sum \|f_k\|$ converge simplement vers une fonction $G : I \to \mathbb{R}$ localement intégrable. On se restreint à $E = \mathbb{R}$ (le résultat pour E de dimension finie s'en déduit en utilisant des coordonnées). On pourra en première lecture se limiter à ce cas.

D'après la proposition précédente, si $\sum_{k=0}^{+\infty}\left(\int_I |f_k|\right) < +\infty$, G est intégrable sur I. La fonction F est localement intégrable sur I, donc $|F|$ l'est aussi, de plus $|F| \leqslant G$, donc $|F|$ et F sont intégrables sur I.

Pour tout $n \geqslant 0$, posons $F_n := \sum_{k=0}^{n} f_k$. Alors, F_n est intégrable sur I comme somme de fonctions intégrables et pour tout $n \geqslant 0$, $|F_n| \leqslant G$. On déduit (28) du théorème de convergence dominée.

Revenons au cas général en supposant que E est un espace normé de dimension finie et que la série $\sum_{k=0}^{+\infty} f_k$ converge *simplement*.

1. Montrons d'abord que $\|F\|$ est intégrable sur I et que :
$$\int_I \|F\| \leqslant \sum_{k\in\mathbb{N}} \int_I \|f_k\|, \tag{29}$$
puis que, pour tout $n \in \mathbb{N}$:
$$\int_I \Big\| \sum_{k>n} f_k \Big\| \leqslant \sum_{k>n} \int_I \|f_k\|. \tag{30}$$
Notons $\Phi_n := \sum_{k=0}^{n} \|f_k\|$. Pour tout $x \in I$ et tout $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que, pour tout $n \geqslant N$, $\big\|\sum_{k=0}^{n} f_k(x) - F(x)\big\| \leqslant \varepsilon$, d'où $\big\|\sum_{k=0}^{n} f_k(x)\big\| \geqslant \big\|F(x)\big\| - \varepsilon$ et $\Phi_n \geqslant \big\|F(x)\big\| - \varepsilon$.
Notons, pour tout $n \in \mathbb{N}$, $\varphi_n := \inf\big(\Phi_n, \|F\|\big)$. On a $\|F\| - \varepsilon \leqslant \varphi_n \leqslant F$, donc la suite $(\varphi_n)_{n\in\mathbb{N}}$ converge simplement vers $\|F\|$. On a :
$$\int_I \varphi_n \leqslant \int_I \Phi_n \leqslant \sum_{k=0}^{+\infty}\left(\int_I \|f_k\|\right).$$
D'après le théorème de convergence monotone, on en déduit que $\|F\|$ est intégrable sur I et l'on a $\int_I \|F\| \leqslant \sum_{k\in\mathbb{N}} \int_I \|f_k\|$. On prouve l'inégalité (30) de façon similaire.

2. Revenons à la preuve de (28). On a :
$$\left\| \int_I F - \int_I \sum_{k=0}^{n} f_k \right\| \leqslant \left\| \int_I \sum_{k>n} f_k \right\| \leqslant \int_I \Big\| \sum_{k>n} f_k \Big\|$$
et, en utilisant (30) :
$$\left\| F - \int_I \sum_{k=0}^{n} f_k \right\| \leqslant \sum_{k>n} \int_I \|f_k\| = \sum_{k\in\mathbb{N}} \int_I \|f_k\| - \sum_{k=0}^{n} \int_I \|f_k\|.$$
Par hypothèse la série $\sum_{k\in\mathbb{N}} \int_I \|f_k\|$ converge, donc $\left\| F - \int_I \sum_{k=0}^{n} f_k \right\|$ tend vers 0 quand $n \to +\infty$. D'où le résultat. ∎

Remarque. Le théorème ci-dessus reste vrai pour des applications à valeurs dans un $\mathbb{K}$-espace de Banach. La preuve est similaire à partir de la version correspondante du théorème de convergence dominée (qui ne figure pas dans ce livre).

Exercice 30.

On considère la suite de fonctions $(f_n)_{n\in\mathbb{N}}$ sur $[0,1]$ définie par :

$$f_n : x \in [0,1] \mapsto (-1)^n \frac{x^n}{\sqrt{1-n+n^2}}.$$

1. Montrer que la série de terme général f_n converge simplement sur $[0,1]$ et que sa somme F est continue sur $[0,1[$. Calculer $\displaystyle\int_0^1 |f_n(x)|\,\mathrm{d}x$ et montrer que la série numérique $\displaystyle\sum_{n\in\mathbb{N}} \int_0^1 \left|f_n(x)\right| \mathrm{d}x$ converge.
2. Montrer que
$$\int_0^1 \sum_{n\in\mathbb{N}} (-1)^n \frac{x^n}{\sqrt{1-n+n^2}} = \sum_{n\in\mathbb{N}} \frac{(-1)^n}{(n+1)\sqrt{1-n+n^2}}. \tag{31}$$
3. Montrer que la série de terme général $\left|f_n(1)\right|$ ne converge pas. Montrer que la série de terme général $|f_n|$ converge uniformément sur tout segment de $[0,1[$. Montrer, en utilisant cette convergence uniforme, que :
$$\int_0^1 \sum_{n\in\mathbb{N}} \frac{x^n}{\sqrt{1-n+n^2}} = \sum_{n\in\mathbb{N}} \int_0^1 \frac{x^n}{\sqrt{1-n+n^2}} = \sum_{n\in\mathbb{N}} \frac{1}{(n+1)\sqrt{1-n+n^2}}.$$

Solution.

1. Si $x=1$, $f_n(1) = \frac{(-1)^n}{\sqrt{1-n+n^2}}$ et la série de terme général $\frac{(-1)^n}{\sqrt{1-n+n^2}}$ converge d'après le critère des séries alternées.
   Si $x \in [0,1[$, on a, pour tout $n \in \mathbb{N}$, $\left|f_n(x)\right| = \frac{x^n}{\sqrt{1-n+n^2}} \leqslant x^n$, donc, pour tout $0 \leqslant a < 1$, la série de terme général $f_n(x)$ est normalement convergente sur $[0,a]$. Sa somme F est donc continue sur $[0,a]$. En faisant varier a on obtient la continuité de F sur $[0,1[$.
   On a :
$$\int_0^1 \left|f_n(x)\right| \mathrm{d}x = \frac{1}{\sqrt{1-n+n^2}} \int_0^1 x^n \,\mathrm{d}x = \frac{1}{(n+1)\sqrt{1-n+n^2}}.$$
   Puisque $\dfrac{1}{(n+1)\sqrt{1-n+n^2}} \sim \dfrac{1}{n^2}$, la série de terme général $\displaystyle\int_0^1 \left|f_n(x)\right| \mathrm{d}x$ converge.
2. La série de terme général $\displaystyle\int_0^1 \left|f_n(x)\right| \mathrm{d}x$ converge et F est continue sur $[0,1[$, on peut donc appliquer la proposition 125 sur $[0,1[$ et l'on obtient :
$$\begin{aligned}\int_0^1 \sum_{n\in\mathbb{N}} (-1)^n \frac{x^n}{\sqrt{1-n+n^2}} &= \sum_{n\in\mathbb{N}} \int_0^1 (-1)^n \frac{x^n}{\sqrt{1-n+n^2}} \\ &= \sum_{n\in\mathbb{N}} \frac{(-1)^n}{(n+1)\sqrt{1-n+n^2}}.\end{aligned}$$

3. La série de terme général $\left|f_n(1)\right| = \frac{1}{\sqrt{1-n+n^2}}$ est divergente.

   Soit $0 \leqslant a < 1$. Si $x \in [0,a]$, $\left|f_n(x)\right| \leqslant a^n$, donc la série de terme général $\left|f_n(x)\right|$ converge normalement (et par suite uniformément) sur $[0,a]$.

   La série numérique de terme général $\displaystyle\int_0^1 \left|f_n(x)\right| \mathrm{d}x = \frac{1}{(n+1)\sqrt{1-n+n^2}}$ est convergente. Le résultat se déduit alors de la proposition 123.

Remarque. Dans le cas de la proposition 125, on ne *suppose* que la convergence *simple* de la série $\sum f_n$, mais en fait la série des valeurs absolues $\sum|f_n|$ converge simplement presque partout (on peut le montrer en utilisant la théorie de l'intégrale de Henstock-Kurzweil ou celle de l'intégrale de Lebesgue). L'exercice ci-dessus en est une illustration.

8.2 Continuité et dérivabilité des fonctions définies par une intégrale

Dans cette partie, X est un intervalle réel ou, plus généralement, un espace métrique et I est un intervalle réel. Nous allons étudier les propriétés de continuité et de dérivabilité de fonctions $F : X \to \mathbb{R}$ définies par une intégrale :

$$F(x) := \int_I f(x,t)\,\mathrm{d}t \quad \text{avec} \quad f : X \times I \to \mathbb{R}.$$

Les résultats s'étendent facilement au cas des applications à valeurs dans un $\mathbb{K}$-espace vectoriel normé de dimension finie (en utilisant des coordonnées et, dans le cas complexe, des parties réelles et imaginaires). Nous ne traiterons pas le cas des applications à valeurs dans un espace de Banach.

Théorème 126 (Continuité).
Soient X un espace métrique, $x_0 \in X$ et $f : X \times I \to \mathbb{R}$ une fonction vérifiant :

1. pour tout $x \in X$, la fonction $t \in I \mapsto f(x,t) \in \mathbb{R}$ est intégrable au sens de Riemann (éventuellement généralisé) sur I ;
2. pour tout $t \in I$, la fonction $x \in X \mapsto f(x,t) \in \mathbb{R}$ est continue en $x_0 \in X$;
3. il existe deux fonctions $g : I \to \mathbb{R}$ et $h : I \to \mathbb{R}$ intégrables au sens de Riemann (éventuellement généralisé) sur I telles que pour tout $x \in X$ et tout $t \in I$:

$$g(t) \leqslant f(x,t) \leqslant h(t).$$

Alors, la fonction $F : X \to \mathbb{R}$ définie par $F : x \mapsto \displaystyle\int_I f(x,t)\,\mathrm{d}t$ est continue en x_0.

Démonstration. Montrons que pour toute suite $(y_n)_{n\in\mathbb{N}}$ de X qui converge vers x_0, on a $\lim\limits_{n\to+\infty} F(y_n) = F(x_0)$. Il suffit pour cela d'appliquer le théorème de convergence encadrée à la suite de fonctions intégrables $f_{y_n} := f(y_n, \cdot)$. La condition 2 implique que la suite de fonctions $(f_{y_n})_{n\in\mathbb{N}}$ converge simplement vers f_{x_0} sur I. La condition 1 garantit que les fonctions f_{y_n} sont intégrables. La condition 3 donne l'encadrement des fonctions f_{y_n} par deux fonctions intégrables g et h. On en déduit que :

$$F(y_n) = \int_I f_{y_n} \underset{n\to+\infty}{\longrightarrow} \int_I f_{x_0} = F(x_0).$$

Point Méthode

La continuité étant une propriété locale, il suffit que l' on ait un encadrement :

$$g(t) \leqslant f(x,t) \leqslant h(t)$$

sur un voisinage convenable de $x_0 \in X$, les fonctions g et h pouvant n'être définies que sur ce voisinage. Bien souvent on majore localement le *module* de f.

Corollaire 127. Si X est un espace métrique compact, si I est un segment et si $f : X \times I \to \mathbb{R}$ est continue, alors la fonction $F : x \in X \mapsto \displaystyle\int_I f(x,t)\,\mathrm{d}t$ est continue sur X.

Démonstration. Comme la fonction f est continue sur le compact $X\times I$, elle y est bornée, disons $\big|f(x,t)\big| \leqslant M$ pour tout $(x,t) \in X \times I$. On peut alors appliquer le théorème de continuité précédent avec $g := -M$ et $h := M$.

Corollaire 128.
Si X est un espace métrique compact, si I est un intervalle et si $f : X \times I \to \mathbb{R}$ est *continue*, alors l'application $(u,v,x) \in I \times I \times X \mapsto \displaystyle\int_u^v f(x,t)\,\mathrm{d}t$ est *continue*.

Démonstration.

On a, en posant $t = u + s(v-u)$, $\displaystyle\int_u^v f(x,t)\,\mathrm{d}t = (v-u)\int_0^1 f\big(x, u+s(v-u)\big)\,\mathrm{d}s$ et l'application $\big((u,v,x),s\big) \in I\times I\times X\times[0,1] \mapsto f\big(x, s(v-u)\big) \in \mathbb{R}$ est continue (d'après le corollaire ci-dessus où l'on remplace X par $I\times I\times X$ et I par $[0,1]$).

Exercice 31.

Montrer que la fonction $\Gamma : x \mapsto \displaystyle\int_0^{+\infty} t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t$ est continue sur $\mathbb{R}_+^*$.

Solution. La fonction Γ est bien définie sur $\mathbb{R}_+^*$ (voir exercice 16 de la page 614). Il suffit de montrer que pour tous $R > r > 0$, la fonction Γ est continue sur $[r,R]$.

Posons $X := [r, R]$ et $I :=]0, +\infty[$. Soit $f : X \times I \to \mathbb{R}$ la fonction définie par $f : (x, t) \mapsto t^{x-1}\mathrm{e}^{-t}$. Cette fonction est continue sur $X \times I$. Elle est en particulier continue par rapport à x.
Pour tout $\rho > 0$, la fonction $f(\rho, .) : t \mapsto t^{\rho-1}e^{-t}$ est intégrable sur $]0, +\infty[$. En effet, elle est, d'une part, intégrable sur $]0, 1]$ car $\rho > 0$ et, d'autre part, intégrable sur $[1, +\infty[$ car il existe une constante C telle que $f(\rho, t) \leqslant C/t^2$ sur $[1, +\infty[$ ($t^{\rho+1}e^{-t} \to 0$ quand $t \to +\infty$). On en déduit que h est intégrable sur $]0, +\infty[$.
Comme f est continue par rapport à t et comme $0 \leqslant f(x, t) \leqslant h(t)$, les hypothèses du théorème de continuité d'une fonction définie par une intégrale sont vérifiées (avec g identiquement nulle). Par conséquent, la fonction Γ est continue sur $[r, R]$.

Venons-en maintenant à la dérivabilité. On se limite maintenant au cas où X est un intervalle.

Théorème 129 (Dérivabilité).
Soit $f : X \times I \to \mathbb{R}$ une fonction de deux variables réelles telle que :

1. pour tout $x \in X$, la fonction $t \mapsto f(x, t)$ est intégrable au sens de Riemann (éventuellement généralisé) sur I ;
2. pour tout $t \in I$, la fonction $x \mapsto f(x, t)$ est dérivable sur X ;
3. il existe deux fonctions $g : I \to \mathbb{R}$ et $h : I \to \mathbb{R}$ intégrables au sens de Riemann (éventuellement généralisé) sur I telles que pour tout $x \in X$ et tout $t \in I$:
$$g(t) \leqslant \frac{\partial f}{\partial x}(x, t) \leqslant h(t);$$
4. pour tout $x \in X$, la fonction $t \mapsto \dfrac{\partial f}{\partial x}(x, t)$ est localement intégrable sur I.

Alors, la fonction $F : X \to \mathbb{R}$ définie par $F(x) := \displaystyle\int_I f(x, t)\,\mathrm{d}t$ est dérivable sur X, la fonction $t \mapsto \frac{\partial f}{\partial x}(x, t)$ est intégrable sur I et :
$$F'(x) = \int_I \frac{\partial f}{\partial x}(x, t)\,\mathrm{d}t.$$

Démonstration. Soit $(y_n \in X \setminus \{x_0\})_{n\in\mathbb{N}}$ une suite qui converge vers $x_0 \in X$. La condition 1 garantit que F est définie sur X : pour tout $x \in X$, la fonction $f_x := f(x, \cdot)$ est intégrable au sens de Riemann (éventuellement généralisé) sur I et $F(x) := \displaystyle\int_I f_x$.

La condition 2 permet l'utilisation du théorème de convergence encadrée :
$$\frac{F(y_n) - F(x)}{y_n - x} = \int_I \frac{f(y_n, t) - f(x_0, t)}{y_n - x_0}\,\mathrm{d}t = \int_I \frac{\partial f}{\partial x}\big(c_n(t), t\big)\,\mathrm{d}t$$
pour un certain point $c_n(t)$ compris entre x_0 et y_n.

La condition 3 implique alors l'encadrement $g \leqslant \dfrac{f_{y_n} - f_{x_0}}{y_n - x_0} \leqslant h$. Or :

$$\lim_{n\to+\infty} \frac{f_{y_n} - f_{x_0}}{y_n - x_0} = \frac{\partial f}{\partial x}(x_0, \cdot).$$

Le résultat annoncé découle du théorème de convergence encadrée. ∎

Point Méthode

La dérivabilité étant une propriété locale, il suffit d'avoir des encadrements de la dérivée partielle valables au voisinage de chaque point de X comme dans le cas de la continuité. Bien souvent, on majore le module.
Si de plus, pour tout $t \in I$, la fonction $x \mapsto \frac{\partial f}{\partial x}(x,t)$ est continue sur X alors F est de classe $\mathcal{C}^1$.

On déduit immédiatement du théorème le corollaire suivant.

Corollaire 130. Soient $I := [a,b]$ et $f : X \times I \to \mathbb{R}$ une fonction telle que :
- pour tout $x \in X$, la fonction $t \mapsto f(x,t)$ est intégrable au sens de Riemann sur I ;
- pour tout $t \in I$, la fonction $x \mapsto f(x,t)$ est dérivable sur X et la fonction $(x,t) \mapsto \frac{\partial f}{\partial x}(x,t)$ est bornée sur $X \times I$.
- pour tout $x \in X$, la fonction $t \mapsto \dfrac{\partial f}{\partial x}(x,t)$ est intégrable au sens de Riemann sur I.

Alors, la fonction $F : X \to \mathbb{R}$ définie par $F(x) := \displaystyle\int_I f(x,t)\,\mathrm{d}t$ est dérivable sur X et :

$$F'(x) = \int_I \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t.$$

Exercice 32.

Montrer que la fonction $\Gamma : x \mapsto \displaystyle\int_0^{+\infty} t^{x-1}e^{-t}\,\mathrm{d}t$ est une fonction de classe $\mathcal{C}^\infty$ sur $\mathbb{R}_+^*$ et que pour tout $k \geqslant 0$, la dérivée $k^{ème}$ de Γ est donnée par :

$$\Gamma^{(k)}(x) = \int_0^{+\infty} t^{x-1}e^{-t}(\ln t)^k\,\mathrm{d}t.$$

Solution.
Posons $f(x,t) = t^{x-1}e^{-t}\,\mathrm{d}t$. La fonction $x \mapsto f(x,t)$ est de classe $\mathcal{C}^\infty$ sur $]0,\infty[$ et :

$$\frac{\partial^k f}{\partial x^k}(x,t) = t^{x-1}e^{-t}(\ln t)^k.$$

Pour tout $k \geqslant 0$, on a la majoration suivante valable pour $0 < a < x < b < +\infty$:

$$\left|t^{x-1}e^{-t}(\ln t)^k\right| \leqslant g_k(t) := \begin{cases} t^{a-1}e^{-t}|\ln t|^k & \text{si } 0 < t < 1 \\ t^{b-1}e^{-t}|\ln t|^k & \text{si } t \geqslant 1. \end{cases}$$

La fonction g_k est intégrable sur $]0,+\infty[$.
On en déduit le résultat demandé par récurrence sur k.

Le résultat suivant est un corollaire des deux théorèmes précédents.

Corollaire 131. Soit $f : X\times[a,b] \to \mathbb{R}$ une fonction continue qui admet en tout point une dérivée partielle $\dfrac{\partial f}{\partial x}(x,t)$ continue sur $X \times [a,b]$. Alors, la fonction $f : X \to \mathbb{R}$ définie par : $F(x) := \displaystyle\int_a^b f(x,t)\,\mathrm{d}t$ est de classe $\mathcal{C}^1$ sur X et :

$$F'(x) = \int_a^b \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t.$$

Démonstration. Les fonctions f et $\dfrac{\partial f}{\partial x}$ sont continues sur $X \times [a,b]$, donc bornées sur $[x_1,x_2]\times[a,b]$ dès que $[x_1,x_2]\subset X$. ∎

Il est utile de savoir dériver une fonction définie par une intégrale dont les bornes dépendent du paramètre.

Proposition 132. Soient $I, X \subset \mathbb{R}$ deux intervalles, $a : X \to I$ et $b : X \to I$ des fonctions de classe $\mathcal{C}^1$ et $f : X \times I \to \mathbb{R}$ une fonction continue qui admet en tout point une dérivée partielle $\dfrac{\partial f}{\partial x}(x,t)$ continue sur $X \times I$. Alors, la fonction $F : X \to \mathbb{R}$ définie par : $F : x \mapsto \displaystyle\int_{a(x)}^{b(x)} f(x,t)\,\mathrm{d}t$ est de classe $\mathcal{C}^1$ sur X et :

$$F'(x) = \int_{a(x)}^{b(x)} \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t + b'(x)f\big(x,b(x)\big) - a'(x)f\big(x,a(x)\big).$$

Démonstration. Il suffit de faire la démonstration dans le cas où X est compact.

Soit $\Phi : (u,v,x) \in I\times I\times X \to \displaystyle\int_u^v f(x,t)\,\mathrm{d}t$. Elle admet des dérivées partielles continues sur $I\times I\times X$. En effet $\frac{\partial\Phi}{\partial u}(u,v,x) = -f(x,u)$, $\frac{\partial\Phi}{\partial v}(u,v,x) = f(x,v)$ et, d'après le corollaire 130, $\frac{\partial\Phi}{\partial x}(u,v,x) = \displaystyle\int_u^v \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t$, cette dernière fonction étant continue d'après le corollaire 128.
La fonction F est la *composée* de la fonction $\Psi : x \in X \mapsto \big(u(x),v(x),x\big) \in I\times I\times X$ et de la fonction Φ. La fonction Ψ est de classe $\mathcal{C}^1$ et sa dérivée est $\Psi'(x) = \big(u'(x),v'(x),1\big)$. D'après le théorème de dérivation des fonctions composées, F est de classe $\mathcal{C}^1$ et l'on a :

$$F'(x) = \left(\frac{\partial\Phi}{\partial u}\big(a(x),b(x),x\big)\right)a'(x) + \left(\frac{\partial\Phi}{\partial v}\big(a(x),b(x),x\big)\right)b'(x) + \frac{\partial\Phi}{\partial x}\big(a(x),b(x),x\big),$$

$$F'(x) = -f\big(x,a(x)\big)a'(x) + f\big(x,b(x)\big)b'(x) + \int_{a(x)}^{b(x)} \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t. \qquad \blacksquare$$

Exercice 33.

Soit $U \subset \mathbb{C}$ un ouvert, et $f : U \times [a,b] \to \mathbb{C}$ une fonction continue. On suppose que pour tout $t \in [a,b]$, l'application $z \mapsto f(z,t)$ est holomorphe sur U et que la dérivée partielle $\dfrac{\partial f}{\partial z}$ est continue sur $U \times [a,b]$. Montrer que l'application $F : U \to \mathbb{C}$ définie par $F(z) := \displaystyle\int_a^b f(z,t)\,\mathrm{d}t$ est holomorphe sur U, que sa dérivée est continue sur U et que :

$$F'(z) = \int_a^b \frac{\partial f}{\partial z}(z,t)\,\mathrm{d}t.$$

Solution. Identifions $\mathbb{C}$ à $\mathbb{R}^2$ via $z = x + iy \mapsto (x,y)$ et écrivons :

$$f(x+iy,t) = p(x,y,t) + iq(x,y,t).$$

On a alors $F(x+iy) = P(x,y) + iQ(x,y)$ avec :

$$P(x,y) := \int_a^b p(x,y,t)\ \mathrm{d}t \quad \text{et} \quad Q(x,y) := \int_a^b q(x,y,t)\,\mathrm{d}t.$$

Par hypothèse, les fonctions p et q sont de classe $\mathcal{C}^1$ par rapport à (x,y), les dérivées partielles sont continues par rapport à (x,y,t) et vérifient les conditions de Cauchy-Riemann :

$$\frac{\partial p}{\partial x}(x,y,t) = \frac{\partial q}{\partial y}(x,y,t) \quad \text{et} \quad \frac{\partial p}{\partial y}(x,y,t) = -\frac{\partial q}{\partial x}(x,y,t).$$

Les fonctions P et Q sont donc de classe $\mathcal{C}^1$ et l'on a :

$$\frac{\partial P}{\partial x}(x,y) = \int_a^b \frac{\partial p}{\partial x}(x,y,t)\,\mathrm{d}t = \int_a^b \frac{\partial q}{\partial y}(x,y,t)\,\mathrm{d}t = \frac{\partial Q}{\partial y}(x,y)$$

et :

$$\frac{\partial P}{\partial y}(x,y) = \int_a^b \frac{\partial p}{\partial y}(x,y,t)\,\mathrm{d}t = -\int_a^b \frac{\partial q}{\partial x}(x,y,t)\,\mathrm{d}t = -\frac{\partial Q}{\partial x}(x,y).$$

Les fonctions P et Q vérifient donc les conditions de Cauchy-Riemann. La fonction F est donc holomorphe sur U et :

$$\begin{aligned} F'(z) = \frac{\partial P}{\partial x}(x,y) + i\frac{\partial Q}{\partial x}(x,y) &= \int_a^b \left(\frac{\partial p}{\partial x}(x,y,t) + i\frac{\partial q}{\partial x}(x,y,t)\right)\ \mathrm{d}t \\ &= \int_a^b \frac{\partial f}{\partial z}(z,t)\,\mathrm{d}t. \end{aligned}$$

Proposition 133 (Théorème de Fubini).
Soit $f : [a,b] \times [c,d] \to \mathbb{R}$ une fonction *continue*.
(i) Les applications :

$$x \in [a,b] \mapsto \int_c^d f(x,y)\,\mathrm{d}y \quad \text{et} \quad y \in [c,d] \mapsto \int_a^b f(x,y)\,\mathrm{d}x$$

sont continues.
(ii) On a :

$$\int_a^b \Big(\int_c^d f(x,y)\,\mathrm{d}y\Big)\,\mathrm{d}x = \int_c^d \Big(\int_a^b f(x,y)\,\mathrm{d}x\Big)\,\mathrm{d}y. \tag{32}$$

Démonstration.
(i) Les applications :

$$x \in [a,b] \mapsto \int_c^d f(x,y)\,\mathrm{d}y \quad \text{et} \quad y \in [c,d] \mapsto \int_a^b f(x,y)\,\mathrm{d}x$$

sont continues d'après le théorème 126 de la page 642.
(ii) On définit deux fonctions H et G sur $[a,b]$ par :

$$H : t \mapsto \int_a^t \Big(\int_c^d f(x,y)\,\mathrm{d}y\Big)\,\mathrm{d}x \quad \text{et} \quad G : t \mapsto \int_c^d \Big(\int_a^t f(x,y)\,\mathrm{d}x\Big)\,\mathrm{d}y.$$

D'après (i) la fonction H est dérivable et l'on a $H'(t) = \displaystyle\int_c^d f(x,y)\,\mathrm{d}y$.

La fonction $\Phi : (t,y) \in [a,b] \times [c,d] \mapsto \displaystyle\int_a^t f(x,y)\,\mathrm{d}x$ admet $f(x,y)$ pour dérivée partielle par rapport à x. D'après le corollaire 130, la fonction G est dérivable par rapport à t et l'on a $G'(t) = \displaystyle\int_c^d \frac{\partial}{\partial t}\Phi(t,y)\,\mathrm{d}y = \int_c^d f(x,y)\,\mathrm{d}y$.

Par suite $H' = G'$. Par ailleurs $H(a) = G(a) = 0$, donc $H = G$ et, en particulier, $H(b) = G(b)$, c'est-à-dire (32). ∎

Nous reviendrons sur le théorème de Fubini à propos des intégrales doubles dans le module *intégration* de [L2-L3]. Sans utiliser ce contexte elle est utile pour calculer diverses intégrales.

Exemple. Soit $0 < a < 1$. On va utiliser le théorème de Fubini pour montrer que :

$$I := \int_0^{\pi/2} \ln(1 + a\cos x)\,\frac{\mathrm{d}x}{\cos x} = \frac{\pi^2}{8} - \frac{1}{2}(\arccos a)^2.$$

La fonction $f : x \mapsto \ln(1 + a\cos x)\,\frac{1}{\cos x}$ est continue sur $]0,\pi/2[$, on la prolonge en une fonction continue sur $[0,\pi/2]$ en posant $f(0) = a$. Ainsi l'intégrale I est convergente.

La fonction $(x,y) \in [0,\pi/2] \times [0,a] \mapsto \dfrac{1}{1 + y\cos x}$ est continue ($|y\cos x| \leqslant a < 1$).

On a $\displaystyle\int_0^a \frac{\mathrm{d}y}{1 + y\cos x} = \frac{\ln(1 + a\cos x)}{\cos x}$ et $I := \displaystyle\int_0^{\pi/2} \mathrm{d}x \int_0^a \frac{\mathrm{d}y}{1 + y\cos x}$.

D'après le théorème de Fubini :

$$I = \int_0^{\pi/2} \mathrm{d}x \int_0^a \frac{\mathrm{d}y}{1 + y\cos x} = \int_0^a \mathrm{d}y \int_0^{\pi/2} \frac{\mathrm{d}x}{1 + y\cos x}.$$

Calculons $\displaystyle\int_0^a \mathrm{d}y \int_0^{\pi/2} \frac{\mathrm{d}x}{1 + y\cos x}$.

On a $\displaystyle\int_0^{\pi/2} \frac{\mathrm{d}x}{1 + y\cos x} = \frac{2}{\sqrt{1-y^2}} \arctan \sqrt{\frac{1-y}{1+y}}$. D'où :

$$I = \int_0^a \frac{2}{\sqrt{1-y^2}} \arctan \sqrt{\frac{1-y}{1+y}}\, \mathrm{d}y.$$

On définit $\varphi : y \in [0,a] \mapsto \arctan \sqrt{\frac{1-y}{1+y}}$. Cette fonction est décroissante sur $[0,a]$ et son image est $[\alpha, \pi/4]$, avec $\alpha := \arctan \sqrt{\frac{1-a}{1+a}} = \frac{1}{2}\arccos a$. On a $y = \cos 2\varphi$, en appliquant la formule de changement de variables, on obtient $I = 4\displaystyle\int_\alpha^{\pi/4} \varphi d\varphi = \frac{\pi^2}{8} - \frac{1}{2}(\arccos a)^2$.

Exercice type corrigé

II.3.0 Soient $f : (x,t) \in \mathbb{R}^+ \times \mathbb{R} \to \dfrac{e^{ixt}}{1+t^2}$ et $g : (x,t) \in \mathbb{R}^+ \times \mathbb{R} \to \dfrac{te^{ixt}}{(1+t^2)^2}$.

i) Montrer que, pour tout $x \in \mathbb{R}^+$, les intégrales $\displaystyle\int_{-\infty}^{+\infty} f(x,t)\,\mathrm{d}t$ et $\displaystyle\int_{-\infty}^{+\infty} g(x,t)\,\mathrm{d}t$ sont convergentes.

ii) Soient $F : x \in \mathbb{R}^+ \mapsto \displaystyle\int_{-\infty}^{+\infty} f(x,t)\,\mathrm{d}t$ et $G : x \in \mathbb{R}^+ \mapsto \displaystyle\int_{-\infty}^{+\infty} g(x,t)\,\mathrm{d}t$.
Montrer que pour tout $x \in]0,+\infty[$, $ixF(x) = 2G(x)$.

iii) Montrer que G est de classe $\mathcal{C}^1$ sur $\mathbb{R}^+$.

iv) Montrer que F est de classe $\mathcal{C}^2$ et que $F'' = F$ sur $]0,+\infty[$.
Montrer que F est continue sur $\mathbb{R}^+$ et que, pour tout $x \in \mathbb{R}^+$, $|F(x)| \leqslant \pi$.
En déduire que pour tout $x \in [0,+\infty[$, $F(x) = \pi e^{-x}$.

Solution.

i) Les fonctions f et g sont continues sur $\mathbb{R}^+ \times \mathbb{R}$. Pour tout $x \geqslant 0, |f(x,t)| = \frac{1}{1+t^2}$ et $|g(x,t)| = \frac{t}{(1+t^2)^2}$, f et g sont donc intégrables sur $]-\infty,+\infty[$.

ii) Soit $x \in]0,+\infty[$. Pour tout $A > 0$, en effectuant une intégration par parties, on a :

$$ix\int_{-A}^{A} f(x,t)\,\mathrm{d}t = \Big[\frac{e^{ixt}}{1+t^2}\Big]_{-A}^{+A} + 2\int_{-A}^{+A} \frac{te^{ixt}}{(1+t^2)^2}\,\mathrm{d}t.$$

Par suite, pour tout $x \in]0,+\infty[$, $ixF(x) = 2G(x)$.
Cette égalité reste vraie pour $x = 0$ ($G(0) = 0$).

iii) Pour tous $(x,t) \in \mathbb{R}^+ \times \mathbb{R}$, on a $\dfrac{\partial g}{\partial x}(x,t) = e^{ixt}\dfrac{it^2}{(1+t^2)^2}$, donc $\left|\dfrac{\partial g}{\partial x}(x,t)\right| \leqslant \dfrac{1}{1+t^2}$.

D'après le théorème 129 de la page 644, la fonction G est de classe $\mathcal{C}^1$ sur $\mathbb{R}^+$ et :

$$G'(x) = \int_{-\infty}^{+\infty} e^{ixt} \frac{it^2}{(1+t^2)^2}\,\mathrm{d}t.$$

iv) Pour tout $x > 0$, $F = -2i\frac{G}{x}$ est dérivable et :

$$i\big(F(x) + xF'(x)\big) = 2G'(x) = 2\int_{-\infty}^{+\infty} e^{ixt} \frac{it^2}{(1+t^2)^2}\,\mathrm{d}t.$$

On en déduit $xF'(x) = F(x) - 2H(x)$ avec $H(x) = \displaystyle\int_{-\infty}^{+\infty} e^{ixt}\frac{1}{(1+t^2)^2}\,\mathrm{d}t$ (en utilisant $t^2 = t^2 + 1 - 1$). On vérifie que H est dérivable sur $]0,+\infty[$ ($\left|e^{ixt}\frac{it}{(1+t^2)^2}\right| = \frac{t}{(1+t^2)^2}$). On a $H'(x) = iG(x)$ et, par suite, pour tout $x \in]0,+\infty[$, $F''(x) = F(x)$.

On a $\left|\frac{e^{itx}}{1+t^2}\right| = \frac{1}{1+t^2}$. L'intégrale $\displaystyle\int_{-\infty}^{+\infty} \frac{\mathrm{d}t}{1+t^2}$ est convergente et égale à π. On en déduit (en utilisant le théorème 126 de la page 642) que la fonction F est continue en 0 et que l'on a, pour tout $x \in \mathbb{R}^+$, $|F(x)| \leqslant \displaystyle\int_{-\infty}^{+\infty} \frac{\mathrm{d}t}{1+t^2} = \pi$.

Puisque $F''(x) = F(x)$ sur $]0,+\infty[$, il existe $\lambda, \mu \in \mathbb{R}$ tels que, pour tout $x \in]0,+\infty[$, $F(x) = \lambda e^x + \mu e^{-x}$, or F est bornée sur $\mathbb{R}^+$, on a donc $\lambda = 0$. Par ailleurs F est continue en 0, donc $\mu = F(0) = \pi$.

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 Soit $f : [a,b] \to \mathbb{R}$ une fonction continue telle que $\big|f(x)\big| \leqslant 1$ pour tout $x \in [a,b]$ et $\displaystyle\int_a^b f = b - a$. Que dire de f ?

II.3.2 À l'aide de sommes de Riemann, calculer les limites suivantes :

$$L_1 = \lim_{n\to+\infty} \sum_{k=0}^{n-1} \frac{1}{\sqrt{n^2+kn}}, \qquad L_2 = \lim_{n\to+\infty} \sum_{k=1}^{n} \frac{k^2}{n^3}, \tag{33}$$

$$L_3 = \lim_{n\to+\infty} \sum_{k=0}^{n-1} \frac{n}{n^2+k^2}, \qquad L_4 = \lim_{n\to+\infty} \sum_{k=0}^{n-1} \frac{k}{n^2+k^2}, \tag{34}$$

$$L_5 = \lim_{n\to+\infty} \sum_{k=n}^{2n-1} \frac{n}{k^2}. \tag{35}$$

II.3.3 Montrer que la relation de Chasles reste valable pour l'intégrale inférieure (resp. supérieure) d'une fonction bornée.

II.3.4 Soit $f : [a,b] \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ sur $[a,b]$ telle que $f(a) = f(b) = 0$. Montrer que si $\displaystyle\int_a^b f^2 = 1$, alors :

$$\int_a^b x f(x) f'(x)\,\mathrm{d}x = -1/2.$$

II.3.5 Soit $(r_n)_{n\in\mathbb{N}^*}$ une énumération des rationnels de l'intervalle $]0,1[$.
On définit un ouvert U de $]0,1[$ par $U :=]0,1[\cap \bigcup_{n\in\mathbb{N}^*} \left]r_n - \frac{1}{2^{n+2}}, r_n + \frac{1}{2^{n+2}}\right[$.
Montrer que $\overset{*}{\displaystyle\int_0^1} 1_U = 1$ et que $\underset{*}{\displaystyle\int_0^1} 1_U \leqslant \frac{1}{2}$.
En déduire que la fonction 1_U n'est pas $\mathscr{R}$-intégrable sur $[0,1]$.

II.3.6 $*$ (*Ensemble de Smith-Volterra-Cantor.*)
On reprend la construction de l'ensemble de Cantor avec quelques modifications.
Pour tout $n \in \mathbb{N}$ on définit par récurrence un sous-ensemble fermé F_n de $[0,1]$ formé par la réunion de 2^n segments *disjoints* de même longueur : $F_n := \bigcup_{k=1}^{2^n} I_{n,k}$. On part de $F_0 := [0,1]$ et, pour passer de F_n à F_{n+1}, on enlève à chaque segment $I_{n,k}$ l'intervalle ouvert de longueur $1/4^{n+1}$ centré au milieu de $I_{n,k}$. Ainsi $F_{n+1} \subset F_n$ et $\log(F_{n+1}) = \log(F_n) - \frac{2^n}{4^{n+1}} = \log(F_n) - \frac{1}{2^{n+2}}$.
On a : $F_0 := [0,1]$, $F_1 := [0,\frac{3}{8}] \cup [\frac{5}{8},1]$, $F_2 := [0,\frac{5}{32}] \cup [\frac{7}{32},\frac{3}{8}] \cup [\frac{5}{8},\frac{25}{32}] \cup [\frac{27}{32},1]$.
On définit l'ensemble de Smith-Volterra-Cantor par $K := \bigcap_{n\in\mathbb{N}} F_n$. On note $U := [0,1] \setminus K$.

1. Montrer que K est un compact d'intérieur vide.
2. Calculer la longueur « naïve » de U, puis celle de K. En déduire que 1_K n'est pas $\mathscr{R}$-intégrable.
3. Montrer que K est non-dénombrable.
4. Tout ouvert de $[0,1]$ de longueur strictement plus grande que $\frac{1}{2}$ rencontre K.

II.3.7 $**$ Terminer la preuve de la question 1 de l'exercice 7 de la page 566.

II.3.8
Soit $\theta : \mathbb{R} \mapsto \mathbb{C}$ une fonction localement intégrable sur $\mathbb{R}$ et périodique admettant le réel $T > 0$ pour période.

i) Soit $[a,b]$, montrer que $\lim\limits_{x\to+\infty} \int_a^b \theta(xt)\,\mathrm{d}t = \dfrac{b-a}{T}\int_0^T \theta(t)\,\mathrm{d}t$.

ii) $f : [a,b] \mapsto \mathbb{C}$ une fonction intégrable sur $[a,b]$, montrer :

$$\lim_{x\to+\infty} \int_a^b \theta(xt)f(t)\,\mathrm{d}t = \frac{1}{T}\int_0^T \theta(t)\,\mathrm{d}t \int_a^b f(t)\,\mathrm{d}t\cdot$$

.

iii) On admet que pour tout $n \in \mathbb{N}^*$, $\displaystyle\int_0^\pi t(t-2\pi)\cos nt = \frac{2\pi}{n^2}\cdot$

Montrer que $2\sum\limits_{k=1}^{n} \cos kt = -1 + \frac{\sin(n+\frac{1}{2})t}{\sin(\frac{t}{2})}\cdot$ En déduire :

$$4\pi\sum_{k=1}^{n}\frac{1}{k^2} = -\int_0^\pi t(t-2\pi)\,\mathrm{d}t + \int_0^\pi t(t-2\pi)\frac{\sin(n+\frac{1}{2})t}{\sin\frac{t}{2}}\,\mathrm{d}t$$

et $\displaystyle\sum_{k=1}^{+\infty}\frac{1}{k^2} = \frac{\pi^2}{6}\cdot$

Exercice sur la section 2

II.3.9 Cet exercice propose une autre démonstration de la proposition 35 de la page 574.
Soit $f : [a,b] \mapsto \mathbb{R}$ une fonction de classe $\mathcal{C}^1$, montrer que f est à variation bornée et calculer sa variation sur $[a,b]$.

II.3.10 On considère l'ensemble $BV([a,b],E)$ des applications à variation bornée de $[a,b]$ dans l'espace de Banach réel E. Pour tout $f \in BV([a,b],E)$ on note :

$$N(f) := V(f) + \big\|f(a)\big\| \quad \text{et} \quad \nu(f) := V(f) + \sup_{x\in[a,b]}\big\|f(x)\big\|.$$

Montrer que N et ν sont des normes équivalentes sur $BV([a,b],E)$.

II.3.11 Soit $f : [a,b] \mapsto \mathbb{R}$ une fonction de classe $\mathcal{C}^1$, par morceaux, montrer que f est à variation bornée et calculer sa variation sur $[a,b]$.

II.3.12 Soient $f : [a,b] \mapsto \mathbb{R}$ une fonction intégrable sur $[a,b]$. Pour tout $n \in \mathbb{N}^*$, on pose :

$$R_n(f) := \int_a^b f(t)\,\mathrm{d}t - \frac{b-a}{n}\sum_{k=0}^{n-1} f\big(a + k\frac{b-a}{n}\big)\cdot$$

(i) Si f est lipschitzienne de rapport K, montrer que : $|R_n(f)| \leqslant K\frac{(b-a)^2}{2n}\cdot$

(ii) Si f est de classe $\mathcal{C}^1$, montrer que $\lim\limits_{n\to+\infty} nR_n(f) = \frac{(b-a)}{2}\big(f(b)-f(a)\big)\cdot$

(ii) Si f est de classe $\mathcal{C}^2$, montrer que :

$$R_n(f) = \frac{(b-a)}{2n}\big(f(b)-f(a)\big) - \frac{(b-a)}{12n^2}\big(f'(b)-f'(a)\big) + \mathrm{o}(\frac{1}{n^2})\cdot$$

Exercices sur la section 4

II.3.13 Soit $I := [a, b]$ un segment de $\mathbb{R}$.

i) Soit $\varphi : I \mapsto \mathbb{C}$ une fonction en escalier sur I, montrer : $\lim_{x\to+\infty} \int_a^b e^{ixt}\,\varphi(t)\,\mathrm{d}t = 0$.

ii) Soit $f : I \mapsto \mathbb{C}$ une fonction intégrable sur I, montrer : $\lim_{x\to+\infty} \int_a^b e^{ixt}\,f(t)\,\mathrm{d}t = 0$.

II.3.14

On note ℓ^∞ le $\mathbb{R}$-espace vectoriel des suites réelles $x := (x_n)_{n\in\mathbb{N}}$ telles que $\sup_{n\in\mathbb{N}} |x_n| < +\infty$. On vérifie que $x \in \ell^\infty \mapsto \|x\| := \sup_{n\in\mathbb{N}} |x_n|$ est une norme sur ℓ^∞. Pour tout $n \in \mathbb{N}$, on note e_n la suite dont tous les termes sont nuls sauf le $n^{\text{ème}}$ qui est égal à un. On a $e_n \in \ell^\infty$.

1. Montrer que $(\ell^\infty, \|.\|)$ est un espace de Banach.
2. On choisit une bijection $\varphi : \mathbb{N} \mapsto \mathbb{Q} \cap [0, 1]$.
   On définit une application $f : [0, 1] \to \ell^\infty$ par :

$$f : t \in [0,1] \mapsto \begin{cases} 0 & \text{si} \quad t \notin \mathbb{Q} \\ e_n & \text{si} \quad t = \varphi(n) \end{cases}.$$

   a) Montrer que f n'est pas $\mathscr{R}$-intégrable.
   b) On considère, pour tout $n \in \mathbb{N}$ la forme linéaire $u_n : x \in \ell^\infty \mapsto x_n \in \mathbb{R}$. Montrer que la fonction $u_n \circ f : [0, 1] \to \mathbb{R}$ est $\mathscr{R}$-intégrable et calculer $\int_0^1 u_n \circ f$.
   c) Que peut-on dire de la convergence des sommes de Riemann de f ?

Exercice sur la section 5

II.3.15 Soient $f : t \in [0, 1] \mapsto t \sin(\frac{1}{t})$ et $g : \mathbb{R} \to \mathbb{R}$ définie par :

$$g(x) = 1 \quad \text{si } x > 0, \qquad g(x) = -1 \quad \text{si } x < 0 \qquad \text{et} \qquad g(0) = 0.$$

Montrer que f et g sont réglées. La fonction $g \circ f$ est-elle réglée ?

Exercices sur la section 6

II.3.16 Soit $f : [0, 1] \to \mathbb{R}$ une fonction continue et croissante. Montrer que la fonction $g :]0, 1]$ définie par :

$$g(x) := \frac{1}{x} \int_0^x f$$

est croissante. Déterminer sa limite éventuelle quand $x \to 0^+$.

Exercices sur la section 7

II.3.17 Parmi les intégrales impropres suivantes, lesquelles sont convergentes ?

1. $\displaystyle\int_0^{+\infty} \sin(x)\,\mathrm{d}x$,
2. $\displaystyle\int_0^{+\infty} \frac{1}{x(x+1)}\,\mathrm{d}x$,
3. $\displaystyle\int_1^{+\infty} \frac{g(x)}{x}\,\mathrm{d}x$ avec $g(x) = (-1)^k$ si $k \leqslant x < k+1$.
4. $\displaystyle\int_0^1 \frac{1}{\sqrt{1-t^3}}\,\mathrm{d}t$.

II.3.18 Étant donnés deux réels $a < b$, étudier la convergence, et déterminer le cas échéant la valeur de l'intégrale $\displaystyle\int_a^b \frac{1}{\sqrt{(t-a)(b-t)}}\,\mathrm{d}t$.

II.3.19

Étudier la convergence, et le cas échéant, déterminer la valeur de l'intégrale $\displaystyle\int_0^1 \frac{\ln t}{\sqrt{1-t}}\,\mathrm{d}t$.

II.3.20

Déterminer pour quelles valeurs de $\lambda \in \mathbb{R}$ et $\mu \in \mathbb{R}$ l'intégrale de Bertrand $\displaystyle\int_0^{1/e} \frac{1}{t^\lambda |\ln t|^\mu}\,\mathrm{d}t$ est convergente.

II.3.21 Étudier la convergence des intégrales $\displaystyle\int_1^{+\infty} t^{-\alpha}\cos t\,\mathrm{d}t$ et $\displaystyle\int_1^{+\infty} t^{-\alpha}\sin t\,\mathrm{d}t$ en fonction de $\alpha \in \mathbb{R}$.

II.3.22 Montrer que pour tout $x \in \mathbb{R}$, l'intégrale $\mathrm{Ai}(x) = \dfrac{1}{\pi}\displaystyle\int_0^{+\infty} \cos\left(\frac{t^3}{3} + xt\right)\,\mathrm{d}t$ est semi-convergente (Ai est appelée *fonction d'Airy*).

II.3.23 $*$

1. Soit g une fonction strictement positive et de classe $\mathcal{C}^1$ au voisinage de $+\infty$. On suppose que $\frac{g'(x)}{g(x)} = \mathrm{o}(\frac{1}{x})$.
   a) Montrer que, pour tout $\varepsilon > 0$, $x^{-\varepsilon} = \mathrm{o}(g)$ et que l'intégrale impropre $\displaystyle\int_a^{+\infty} g(x)\,\mathrm{d}x$ est divergente.
   b) Montrer que $g(x) + xg'(x) \sim g(x)$.
   c) Montrer que $\displaystyle\int_a^x \big(g(t) + tg'(t)\big)\,\mathrm{d}t = xg(x) - ag(a)$.
   En déduire que $\displaystyle\int_a^x g(t)\,\mathrm{d}t \sim xg(x)$.

2. Soit $a > 1$. Montrer que l'on a, au voisinage de $+\infty$, $\displaystyle\int_a^x \frac{\mathrm{d}t}{\ln t} \sim \frac{x}{\ln x}$.

II.3.24

1. Montrer qu'il existe une fonction positive $f : [1, +\infty[\to \mathbb{R}$, localement intégrable, ne convergeant pas vers 0 quand $x \to +\infty$, et telle que l'intégrale $\displaystyle\int_1^{+\infty} f$ converge.
2. Même question en supposant en plus que f est continue.

II.3.25 Montrer que l'intégrale $\displaystyle\int_1^{+\infty} \frac{x - E(x) - \frac{1}{2}}{x}\,\mathrm{d}x$ est convergente.

La calculer en utilisant la formule de Stirling $n! \sim \sqrt{2\pi}(\frac{n}{e})^n$.

II.3.26 Soit $f : [a, +\infty[\mapsto \mathbb{C}$ une application de classe $\mathcal{C}^1$ telle que $\displaystyle\int_a^{+\infty} |f'(t)|\,\mathrm{d}t$ converge.

Pour tout $n \in \mathbb{N}$ on pose $x_n := \displaystyle\sum_{k=0}^{n} f(a+k) - \int_a^{a+n+1} f(t)\,\mathrm{d}t$.

i) Montrer que la suite (x_n) converge.

ii) Montrer que la série $\sum f(a+n)$ et l'intégrale $\displaystyle\int_a^{+\infty} f(t)\,\mathrm{d}t$ sont de même nature.

iii) Pour quelles valeurs de $s \in \mathbb{C}$ la série $\displaystyle\sum_{n\in\mathbb{N}^*} \frac{1}{n^s}$ est-elle convergente ?

(Voir la définition de $t^s, t \in]0, +\infty[, s \in \mathbb{C}$, dans le module *Suites et séries de fonctions*, page 523).

II.3.27 Soient $f : (x,t) \in [1, +\infty[\times]0, +\infty[\to \dfrac{e^{-t} - e^{-xt}}{t}$.

i) Montrer que pour tout $x \in [1, +\infty[$, l'intégrale $\displaystyle\int_0^{+\infty} f(x,t)\,\mathrm{d}t$ converge.

ii) Soit $F : x \in [1, +\infty[\to \displaystyle\int_0^{+\infty} f(x,t)\,\mathrm{d}t$.

Montrer que F est de classe $\mathcal{C}^1$ sur $[1, +\infty[$ et que $F'(x) = \frac{1}{x}$.

iii) Soient $a, b \in]0, +\infty[^2$ avec $a < b$.

Montrer que l'intégrale $\displaystyle\int_0^{+\infty} \frac{e^{-at} - e^{-bt}}{t}\,\mathrm{d}t$ converge, la calculer.

II.3.28 Montrer que $I := \displaystyle\int_0^1 \frac{\ln t \ln(1-t)}{t}\,\mathrm{d}t$ converge et exprimer I en fonction de ζ.

II.3.29 Soient $\alpha \in \mathbb{N}^*$ et $f : x \in]0,+\infty[\mapsto \displaystyle\int_0^1 \frac{1}{(t^\alpha + x^\alpha)^{\frac{1}{\alpha}}}\,\mathrm{d}t$.

i) Montrer que $f(x) \underset{0}{\sim} -\ln(x)$.

ii) Montrer que la fonction $x \in]0,+\infty[\to f(x) + \ln(x)$ admet au voisinage de 0 un développement limité à tout ordre.
Écrire le développement limité de $f(x) + \ln x$ au voisinage de 0 à l'ordre 11 si $\alpha = 3$.

II.3.30 *Intégrales de Fresnel*
On souhaite calculer les intégrales de Fresnel et montrer que :

$$\int_0^{+\infty} \cos t^2\,\mathrm{d}t = \int_0^{+\infty} \sin t^2\,\mathrm{d}t = \frac{1}{2}\sqrt{\frac{\pi}{2}}.$$

i) Montrer que pour tout $x \in \mathbb{R}$, l'intégrale $\displaystyle\int_0^{+\infty} \frac{e^{-(t^2+i)x^2}}{t^2+i}\,\mathrm{d}t$ est convergente.

ii) Soit $F : x \in \mathbb{R} \mapsto \displaystyle\int_0^{+\infty} \frac{e^{-(t^2+i)x^2}}{t^2+i}\,\mathrm{d}t$.
Montrer que F est de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et que $F'(x) = -\sqrt{\pi}\,e^{-ix^2}$.

iii) En déduire $\displaystyle\int_0^{+\infty} e^{-ix^2}\,\mathrm{d}x = \frac{1}{2}\sqrt{\frac{\pi}{2}}(1-i)$.

Exercices sur la section 8

II.3.31 $**$
La question 2 de cet exercice généralise la question 1 de l'exercice 7 de la page 566.
Soient $a, b \in \mathbb{R}$, avec $a < b$.

1. Montrer que si X est un sous-ensemble élémentaire non vide de $[a,b]$ et si $\varepsilon > 0$, il existe un sous-ensemble élémentaire *fermé* $K \subset X$ tel que $\log K > \log X - \varepsilon$.
2. Soit $(A_n)_{n\in\mathbb{N}^*}$ une suite décroissante de parties bornées de $[a,b]$. On suppose $\displaystyle\bigcap_{n\in\mathbb{N}^*} A_n = \varnothing$.
   Pour tout $n \in \mathbb{N}^*$, on note :
   $$\alpha_n := \sup\{\log X \mid X \text{ est un sous-ensemble élémentaire de } [a,b] \text{ contenu dans } A_n\}.$$
   Montrer que α_n tend vers 0 quand n tend vers $+\infty$.

II.3.32 $*$
Le but de cet exercice est de démontrer le théorème d'Arzela (*cf.* le corollaire 122) en utilisant les résultats de l'exercice II.3.31 ci-dessus.

1. Soit $(f_n)_{n\mathbb{N}}$ une suite de fonctions *positives* $\mathscr{R}$-intégrables sur $[a,b]$ convergeant *simplement* vers 0 sur $[a,b]$. On suppose qu'il existe $K > 0$ tel que $f_n \leqslant K$ pour tout $n \in \mathbb{N}$. Montrer que $\lim\limits_{n\to+\infty} \displaystyle\int_a^b f_n = 0$ en utilisant les résultats de l'exercice II.3.31 ci-dessus.
2. Montrer le théorème d'Arzela en utilisant le résultat de la question 1.

II.3.33 On considère $\displaystyle\int_0^{+\infty} \frac{\mathrm{e}^{-x(1+t^2)}}{1+t^2}\,\mathrm{d}t$.

1. Montrer que, pour tout $x \geqslant 0$ cette intégrale impropre est convergente. Montrer que la fonction $F : x \mapsto \displaystyle\int_0^{+\infty} \frac{\mathrm{e}^{-x(1+t^2)}}{1+t^2}\,\mathrm{d}t$ est continue sur $[0,+\infty[$ et déterminer $\lim\limits_{x\to+\infty} F(x)$.
2. Montrer que F est dérivable sur $]0,+\infty[$ et déterminer sa dérivée.
3. En intégrant F' sur $]0,+\infty[$, montrer que $\displaystyle\int_0^{+\infty} e^{-t^2}\,\mathrm{d}t = \frac{\sqrt{\pi}}{2}$.

II.3.34 Étudier la continuité et la dérivabilité de la fonction $f : [0,+\infty[\to \mathbb{R}$ définie par :

$$f(x) = \int_0^{+\infty} \frac{\sin t}{t}\, e^{-tx}\,\mathrm{d}t.$$

II.3.35 Étudier la fonction $f : x \mapsto f(x) = \displaystyle\int_0^{+\infty} \frac{t^{x-1}}{1+t}\,\mathrm{d}t$.

II.3.36 On dit que l'on a défini une théorie de l'intégration $\mathscr{T}$ sur $[a,b]$ si l'on s'est donné un sous-espace vectoriel de $\mathcal{F}\big([a,b]\big)$ appelé espace des fonctions $\mathscr{T}$-intégrables et une forme linéaire positive sur cet espace dite $\mathscr{T}$-intégrale.
On dit qu'une théorie de l'intégration $\mathscr{T}$ sur $[a,b]$ est *plus forte* que l'intégration au sens de Riemann si toute fonction $f : [a,b] \mapsto \mathbb{R}$ $\mathscr{R}$-intégrable sur $[a,b]$ est aussi $\mathscr{T}$-intégrable sur $[a,b]$ et si les intégrales coïncident. Ce n'est évidemment intéressant que s'il existe une fonction $\mathscr{T}$-intégrable qui n'est pas $\mathscr{R}$-intégrable, on dit alors que $\mathscr{T}$ est *strictement* plus forte que $\mathscr{R}$. Nous verrons dans le module *Intégration* de [L2-L3] que les théories de Lebesgue et de Henstock-Kurzweil sont strictement plus fortes que $\mathscr{R}$.
Dans cet exercice, *on suppose qu'il existe* une théorie de l'intégration $\mathscr{T}$ sur $[a,b]$ telle que le résultat suivant soit vrai.
($\star$) Soit $(f_n)_{n\in\mathbb{N}}$ une suite *uniformément bornée* de fonctions $\mathscr{T}$-intégrables sur $[a,b]$ convergeant *simplement* vers f sur $[a,b]$, alors f est $\mathscr{T}$-intégrable sur $[a,b]$ et l'on a :

$$\int_a^b f = \lim_{n\to+\infty} \int_a^b f_n.$$

1. Montrer que si U est un ouvert de $[a,b]$, 1_U est $\mathscr{T}$-intégrable sur $[a,b]$ et que l'on a $\displaystyle\int_a^b 1_U = \mathrm{long}(U)$, la longueur étant définie « naïvement » comme dans 1.3.6. En déduire que, pour tout compact $K \subset [a,b]$, 1_K est $\mathscr{T}$-intégrable. Calculer $\displaystyle\int_0^1 1_K$ si K est le compact de Smith-Cantor-Volterra de l'exercice II.3.6.
2. Montrer que $\mathscr{T}$ est *strictement* plus forte que $\mathscr{R}$.
3. Donner une preuve du corollaire 122 de la page 632 (théorème d'Arzela) en utilisant la théorie $\mathscr{T}$.

Séries de Fourier

II.4

En 1809, dans son *Mémoire sur la propagation de la chaleur*, Fourier s'intéresse au problème suivant : l'évolution de la température d'un anneau circulaire métallique dont les deux moitiés sont initialement à des températures différentes. Il énonce alors un principe tout à fait surprenant : toute fonction périodique peut être représentée par une série de sinus et de cosinus. C'est Dirichlet qui en 1829 donnera des bases solides à la théorie de Fourier. Cette théorie est un outil fondamental dans l'étude des équations aux dérivées partielles comme l'équation de la chaleur ou l'équation des ondes. Nous y reviendrons en fin de module.
La première série qu'exhibe Fourier est :

$$\cos x - \frac{1}{3}\cos(3x) + \frac{1}{5}\cos(5x) - \frac{1}{7}\cos(7x) + \cdots = \begin{cases} +\pi/4 & \text{pour } |x| < \pi/2 \\ 0 & \text{pour } |x| = \pi/2 \\ -\pi/4 & \text{pour } \pi/2 < |x| < \pi \end{cases}$$

(c'est en tout cas ce qu'il prétend). Un des buts de ce module est de donner les outils qui permettent de justifier cette affirmation.

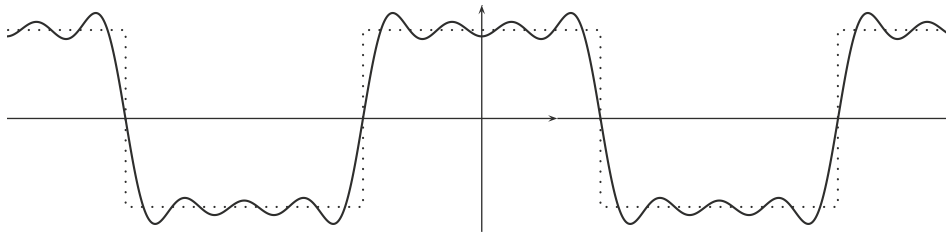


FIGURE II.4.1 – La somme des quatre premiers termes de la série $\displaystyle\sum_{n\geqslant 0} \frac{1}{2n+1}\cos(2n+1)x$ (en trait continu) et de la fonction 2π-périodique qui vaut $\pi/4$ pour $|x| < \pi/2$ et $-\pi/4$ pour $\pi/2 < |x| < \pi$ (en pointillés).

Dans tout le module, on se donne un réel $T > 0$ et on pose $\omega := \dfrac{2\pi}{T}$.

Définition 1. On appelle *polynôme trigonométrique de degré N et de période T* toute fonction de la forme :

$$x \mapsto \sum_{n=-N}^{N} c_n e^{i\omega n x}.$$

On appelle *série trigonométrique* toute série de la forme :

$$c_0 + \sum_{n=1}^{+\infty} \left(c_n e^{i\omega n x} + c_{-n} e^{-i\omega n x} \right).$$

Remarque. Un polynôme trigonométrique ou une série trigonométrique peut également s'écrire comme somme de sinus et de cosinus, et vice et versa.

Le problème qui nous intéresse dans ce module est le suivant. Quelles fonctions peuvent être approchées par des polynômes trigonométriques, en quel sens ? Quelles sont les fonctions qui sont somme d'une série trigonométrique ?

Nous commencerons par étudier le cas des fonctions continues. Nous verrons que le problème qui nous intéresse est fortement lié au module I.4 et en particulier aux espaces préhilbertiens complexes étudiés pages 186 et suivantes : l'espace vectoriel des fonctions continues T-périodiques est muni d'un produit scalaire et la famille $\left\{x \mapsto e^{i\omega n x}\right\}_{n\in\mathbb{Z}}$ est une famille orthonormée. Nous serons amenés à projeter orthogonalement des fonctions continues sur l'espace vectoriel des polynômes trigonométriques de degré N.

Nous montrerons ensuite un théorème de Weierstraß qui dit que toute fonction continue T-périodique est limite uniforme d'une suite de polynômes trigonométriques T-périodiques.

La topologie de la convergence uniforme n'est pas la même que la topologie induite par le produit scalaire.

Nous étudierons ensuite le cas plus général des fonctions réglées, qui contient en particulier le cas des fonctions continues par morceaux.

Enfin, nous donnerons quelques exemples pour illustrer la multitude d'applications de la théorie des séries de Fourier.

Nous tenons à remercier vivement John H. Hubbard qui a grandement influencé la rédaction de ce module.

1 Coefficients de Fourier

Dans toute la suite, f est une fonction définie sur $\mathbb{R}$, périodique de période T. La plupart du temps, les fonctions que nous considérerons seront des fonctions à valeurs réelles. Mais il est plus naturel de se placer dans le cadre plus général des fonctions à valeurs complexes. La quantité $2\pi/T$ interviendra fréquemment et nous la noterons $\omega := 2\pi/T$.

L'espace naturel dans lequel la théorie de Fourier prend place est celui des fonctions qui sont de carré intégrable sur une période, c'est-à-dire, telles que :

$$\int_0^T |f|^2 < +\infty.$$

Nous commencerons par nous placer dans le cadre des fonctions continues. Nous traiterons ensuite le cas des fonctions réglées, qui contient le cas des fonctions continues par morceaux.

Définition 2.
Nous noterons $\mathcal{C}_T$ l'espace des fonctions continues $f : \mathbb{R} \to \mathbb{C}$ qui sont T-périodiques.

Proposition 1. L'espace $\mathcal{C}_T$ muni du produit scalaire :

$$(f \mid g) := \frac{1}{T}\int_0^T \bar{f}(t)g(t)\,\mathrm{d}t$$

est un espace préhilbertien complexe.

Démonstration. La linéarité de l'intégrale montre que l'application $g \mapsto (f \mid g)$ est linéaire et que l'application $f \mapsto (f \mid g)$ est semi-linéaire. La symétrie hermitienne $(f \mid g) = \overline{(g \mid f)}$ est évidente. Enfin, si $f \in \mathcal{C}_T$ et si $(f \mid f) = \displaystyle\int_0^T |f|^2 = 0$, alors f est nulle sur $[0, T]$ par continuité et donc sur $\mathbb{R}$ par périodicité.
Par conséquent, si $f \neq 0$, on a $(f \mid f) > 0$. ∎

Dans la suite, nous supposerons $\mathcal{C}_T$ muni de ce produit scalaire et nous noterons $\|\cdot\|_2$ la norme associée.

1.1 La famille des exponentielles

L'observation de départ est que pour $n \in \mathbb{Z}$, les exponentielles :

$$\begin{array}{rccc} \mathrm{e}_n : & \mathbb{R} & \longrightarrow & \mathbb{C} \\ & t & \longmapsto & \mathrm{e}^{i\omega nt} \end{array}$$

forment une famille orthonormale de $\mathcal{C}_T$. En effet, pour tout $n \in \mathbb{Z}$:

$$\|\mathrm{e}_n\|_2^2 = \frac{1}{T}\int_0^T |\mathrm{e}_n|^2 = 1$$

et pour $n \neq m$:

$$(e_n \mid e_m) = \frac{1}{T}\int_0^T \mathrm{e}^{i\omega(m-n)t}\,\mathrm{d}t = \left[\frac{\mathrm{e}^{i\omega(m-n)t}}{i\omega(m-n)}\right]_0^T = \frac{\mathrm{e}^{i2\pi(m-n)} - 1}{i\omega(m-n)} = 0.$$

Remarque. On utilise la notation $\mathbb{U}$ pour désigner l'ensemble des nombres complexes u de module 1. Si l'on pose $u := e^{i\omega t}$, alors $\mathrm{e}_n(t) = u^n$. Si l'on pose $\chi(u) := u^n$, on obtient sur $\mathbb{U}$ une fonction continue telle que $\chi(uv) = \chi(u)\chi(v)$ pour tous u et v dans $\mathbb{U}$. Autrement dit, χ est un homomorphisme de groupe de $\mathbb{U}$ dans $\mathbb{C}^*$, c'est un *caractère* de $\mathbb{U}$. De plus, les applications $u \mapsto u^n$ sont les seuls caractères de $\mathbb{U}$. C'est cette remarque qui est à l'origine des généralisations contemporaines de la théorie.

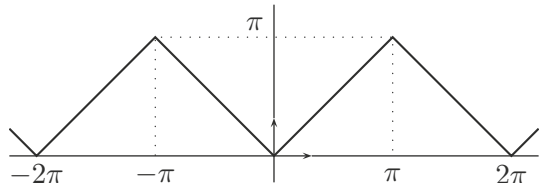
Si $f \in \mathcal{C}_T$ et si J est une partie finie de $\mathbb{Z}$, le projeté orthogonal de f sur l'espace vectoriel engendré par $\{\mathrm{e}_j\}_{j\in J}$ est donc donné par :

$$\sum_{j\in J} c_j(f)\cdot \mathrm{e}_j \quad \text{avec} \quad c_j(f) := (\mathrm{e}_j \mid f).$$

Définition 3. Si $f \in \mathcal{C}_T$, les *coefficients de Fourier* de f sont :

$$c_n(f) := (\,\mathrm{e}_n \mid f\,) = \frac{1}{T}\int_0^T f(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t.$$

Exemple. Considérons la fonction $f \in \mathcal{C}_{2\pi}$ définie par $f(x) = |x|$ si $x \in [-\pi, \pi]$ et prolongée à $\mathbb{R}$ par périodicité.



Alors :

$$c_0(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |t|\,\mathrm{d}t = \frac{1}{\pi}\int_0^{\pi} t\,\mathrm{d}t = \frac{\pi}{2}$$

et pour $n \neq 0$:

$$\begin{aligned}
c_n(f) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |t|\mathrm{e}^{-int}\,\mathrm{d}t \\
&= \frac{1}{2\pi}\int_{-\pi}^{0} -t\mathrm{e}^{-int}\,\mathrm{d}t + \frac{1}{2\pi}\int_0^{\pi} t\mathrm{e}^{-int}\,\mathrm{d}t \\
&= \frac{1}{2\pi}\int_0^{\pi} u\mathrm{e}^{inu}\,\mathrm{d}u + \frac{1}{2\pi}\int_0^{\pi} t\mathrm{e}^{-int}\,\mathrm{d}t \\
&= \frac{1}{2\pi}\left[t\frac{\mathrm{e}^{int}-\mathrm{e}^{-int}}{in}\right]_0^{\pi} - \frac{1}{2\pi}\int_0^{\pi}\frac{\mathrm{e}^{int}-\mathrm{e}^{-int}}{in}\,\mathrm{d}t \\
&= \left[-\frac{\mathrm{e}^{int}+\mathrm{e}^{-int}}{(in)^2}\right]_0^{\pi} = \frac{(-1)^n-1}{\pi n^2}\cdot
\end{aligned}$$

Autrement dit, si $n \neq 0$ est pair, alors $c_n(f) = 0$ et si n est impair, $c_n(f) = -\dfrac{2}{\pi n^2}\cdot$

Définition 4. Si $f \in \mathcal{C}_T$, la série de Fourier de f est la série :

$$c_0(f) + \sum_{n\geqslant 1}\big(c_n(f)\cdot \mathrm{e}_n + c_{-n}(f)\cdot \mathrm{e}_{-n}\big).$$

La $N^{\text{ème}}$ *somme partielle* de la série de Fourier de f est :

$$S_N(f) : t \mapsto \sum_{n=-N}^{N} c_n(f)\mathrm{e}^{i\omega nt}.$$

Pour simplifier l'écriture, nous utiliserons la notation $\displaystyle\sum_{n\in\mathbb{Z}} c_n(f)\cdot \mathrm{e}_n$ pour désigner la série de Fourier de la fonction f.

Rappelons (théorème 14 de la page 152 et corollaire 15) que si E est un espace préhilbertien complexe et si $\{v_i\}_{i\in I}$ est une famille (finie ou infinie) orthonormale dans V, alors pour tout vecteur $w \in E$ et toute partie finie $J \subset I$, le vecteur :

$$w_J := \sum_{j\in J} (\, v_j \mid w\,) v_j$$

est l'unique élément du sous-espace vectoriel engendré par $\{v_j\}_{j\in J}$ qui minimise l'application $v \mapsto \|w - v\|_2$. De plus, on a l'inégalité de Bessel :

$$\|w_J\|_2^2 = \sum_{j\in J} \big|(\, v_j \mid w\,)\big|^2 \leqslant \|w\|_2^2.$$

Rappelons que si $(c_n)_{n\in\mathbb{Z}}$ est une famille de nombre complexes, et s'il existe une constante M telle que pour toute partie finie $J \subset \mathbb{Z}$, on ait $\sum\limits_{j\in J} |c_j| \leqslant M$, on dit que la série $\sum\limits_{n\in\mathbb{Z}} c_n$ est absolument convergente. On écrit $\sum\limits_{n\in\mathbb{Z}} |c_n| < +\infty$. Si les c_n sont des nombres réels et positifs, on définit :

$$\sum_{n\in\mathbb{Z}} c_n := \sup_{\substack{J\subset\mathbb{Z}\\ J\text{fini}}} \sum_{j\in J} u_j.$$

Si c_n est un nombre réel de signe quelconque, on peut écrire $c_n = c_n^+ - c_n^-$ avec :

$$c_n^+ = \max(c_n, 0) \quad \text{et} \quad c_n^- = \max(-c_n, 0).$$

Si $\sum\limits_{n\in\mathbb{Z}} |c_n| < +\infty$, on définit :

$$\sum_{n\in\mathbb{Z}} c_n := \sum_{n\in\mathbb{Z}} c_n^+ - \sum_{n\in\mathbb{Z}} c_n^-.$$

Enfin, si $c_n = a_n + ib_n$ avec a_n et b_n des nombres réels et si $\sum\limits_{n\in\mathbb{Z}} |c_n| < +\infty$, on pose :

$$\sum_{n\in\mathbb{Z}} c_n := \sum_{n\in\mathbb{Z}} a_n + i \sum_{n\in\mathbb{Z}} b_n.$$

Définition 5. On note $\ell^2(\mathbb{Z})$ l'ensemble des familles de nombre complexes $(c_n)_{n\in\mathbb{Z}}$ indexées par $\mathbb{Z}$ telles que $\sum\limits_{n\in\mathbb{Z}} |c_n|^2 < +\infty$.

Notons que si $(c_n)_{n\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$, alors $c_n \underset{n\to\pm\infty}{\longrightarrow} 0$. De plus, l'inégalité de Cauchy-Schwarz montre que si $c := (c_n)_{n\in\mathbb{Z}}$ et $d := (d_n)_{n\in\mathbb{Z}}$ sont deux éléments de $\ell^2(\mathbb{Z})$, alors $\sum\limits_{n\in\mathbb{Z}} |c_n d_n| < +\infty$, ce qui permet de définir un produit scalaire sur $\ell^2(\mathbb{Z})$:

$$(\, c \mid d\,) := \sum_{n\in\mathbb{Z}} \overline{c_n} \cdot d_n.$$

Théorème 2. Si $f \in \mathcal{C}_T$, alors pour tout entier $N \geqslant 0$, la somme partielle $S_N(f)$ réalise la meilleure approximation de f parmi les polynômes trigonométriques de degré N, en ce sens que :

$$\int_0^T \big|S_N(f) - f\big|^2 \leqslant \int_0^T |P - f|^2$$

pour tout polynôme trigonométrique $P = \sum_{n=-N}^{N} d_n \cdot \mathrm{e}_n$, avec égalité si, et seulement si, $P = S_N(f)$. De plus, $\big(c_n(f)\big)_{n\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$ et on a l'inégalité suivante, dite *inégalité de Bessel* :

$$\sum_{n\in\mathbb{Z}} \big|c_n(f)\big|^2 \leqslant \frac{1}{T}\int_0^T |f|^2.$$

Corollaire 3. Si $f \in \mathcal{C}_T$, alors $c_n(f) \underset{n\to\pm\infty}{\longrightarrow} 0$.

1.2 La famille des sinus et cosinus

Lorsqu'une fonction $f \in \mathcal{C}_T$ est à valeurs réelles, il est fréquent d'utiliser une autre écriture des sommes partielles de Fourier. En effet, on peut toujours écrire :

$$\mathrm{e}_n(t) = \cos(\omega n t) + i\sin(\omega n t).$$

Par conséquent :

$$\sum_{n=-N}^{N} c_n \mathrm{e}^{i\omega n t} = \frac{a_0}{2} + \sum_{n=1}^{N} a_n \cos(\omega n t) + \sum_{n=1}^{N} b_n \sin(\omega n t)$$

avec, pour $n \geqslant 0$:

$$a_n = c_n + c_{-n} \quad \text{et} \quad b_n = ic_n - ic_{-n}.$$

Cela peut également s'écrire sous la forme :

$$c_n = \frac{1}{2}(a_n - ib_n) \quad \text{et} \quad c_{-n} = \frac{1}{2}(a_n + ib_n).$$

Notons que pour $n \geqslant 0$:

$$a_n = \frac{1}{T}\int_0^T f(t)(\mathrm{e}^{i\omega n t} + \mathrm{e}^{-i\omega n t})\,\mathrm{d}t = \frac{2}{T}\int_0^T f(t)\cos(\omega n t)\,\mathrm{d}t$$

et

$$b_n = \frac{1}{T}\int_0^T f(t)(i\mathrm{e}^{i\omega n t} - i\mathrm{e}^{-i\omega n t})\,\mathrm{d}t = \frac{2}{T}\int_0^T f(t)\sin(\omega n t)\,\mathrm{d}t.$$

Cela conduit à la définition suivante.

Définition 6. Si $f \in \mathcal{C}_T$, on pose, pour $n \geqslant 0$:

$$a_n(f) = \frac{2}{T}\int_0^T f(t)\cos(\omega n t)\,\mathrm{d}t \qquad \text{et} \qquad b_n(f) = \frac{2}{T}\int_0^T f(t)\sin(\omega n t)\,\mathrm{d}t.$$

La $N^{ème}$ somme partielle de la série de Fourier d'une fonction $f \in \mathcal{C}_T$ est donc :

$$S_N(f) : t \mapsto \frac{a_0(f)}{2} + \sum_{n=1}^{N} a_n(f)\cos(\omega n t) + \sum_{n=1}^{N} b_n(f)\sin(\omega n t).$$

Cette écriture de la somme partielle d'une série de Fourier présente au moins deux intérêts. Si $f \in \mathcal{C}_T$ est une fonction à valeurs réelles, les coefficients $a_n(f)$ et $b_n(f)$ sont réels (ce qui n'est pas nécessairement le cas des coefficients $c_n(f)$). De plus, la fonction f est une fonction paire si, et seulement si, les coefficients $b_n(f)$ sont tous nuls. La fonction f est une fonction impaire si, et seulement si, les coefficients $a_n(f)$ sont tous nuls, comme nous le verrons à la suite du corollaire 16 de la page 674.

Exemple. Si l'on reprend l'exemple de la fonction $f \in \mathcal{C}_{2\pi}$ qui coïncide avec $x \mapsto |x|$ sur $[-\pi, \pi]$, on peut écrire les sommes partielles de la série de Fourier sous la forme :

$$S_N(f)(t) = \frac{\pi}{2} + \frac{4}{\pi} \sum_{p=0}^{\lfloor \frac{N-1}{2} \rfloor} \frac{1}{(2p+1)^2} \cos\big((2p+1)t\big).$$

La fonction f est paire. La série de Fourier ne fait intervenir que des cosinus.

Dans toute la suite, les résultats que nous énoncerons concerneront les coefficients $c_n(f)$. Le lecteur intéressé pourra interpréter ces résultats en terme des coefficients $a_n(f)$ et $b_n(f)$.

1.3 Propriétés des coefficients de Fourier

Proposition 4. L'application $\mathcal{F} : \mathcal{C}_T \to \ell^2(\mathbb{Z})$ qui à une fonction $f \in \mathcal{C}_T$ associe la fonction $\hat{f} : \mathbb{Z} \to \mathbb{C}$ définie par $\hat{f}(n) := c_n(f)$ est une application linéaire. On a :

$$\|\hat{f}\|_\infty := \sup_{n\in\mathbb{Z}} \big|c_n(f)\big| \leqslant \|f\|_1 := \frac{1}{T}\int_0^T |f|.$$

Démonstration. Ces propriétés découlent directement de la définition des coefficients de Fourier et leurs démonstrations sont laissées aux soins du lecteur. ∎

Proposition 5. On a les relations suivantes entre coefficients de Fourier :

1. Coefficients de $\overline{f}$.
   On a $c_n(\overline{f}) = \overline{c_{-n}(f)}$ et si f est une fonction à valeurs réelles, $c_n(f) = \overline{c_{-n}(f)}$.
2. Coefficients de $g : t \mapsto f(-t)$. On a $c_n(g) = c_{-n}(f)$.
   Si f est paire, alors $c_{-n}(f) = c_n(f)$ et si f est impaire $c_{-n}(f) = -c_n(f)$.
3. Coefficients de $g : t \mapsto f(t+a)$. On a $c_n(g) = \mathrm{e}^{i\omega n a} c_n(f)$.

Démonstration.

1. $c_n(\overline{f}) = \dfrac{1}{T}\displaystyle\int_0^T \overline{f}(t)\mathrm{e}^{-i\omega n t}\,\mathrm{d}t = \frac{1}{T}\int_0^T \overline{f(t)\mathrm{e}^{i\omega n t}}\,\mathrm{d}t = \overline{c_{-n}(f)}.$

2. Le changement de variable $u = -t$ donne :

$$c_n(g) = \frac{1}{T}\int_0^T f(-t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t = \frac{1}{T}\int_{-T}^0 f(u)\mathrm{e}^{i\omega nu}\,\mathrm{d}u = c_{-n}(f).$$

3. Le changement de variable $u = t + a$ donne :

$$\begin{aligned} c_n(g) &= \frac{1}{T}\int_0^T f(t+a)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t \\ &= \frac{1}{T}\int_a^{T+a} f(u)\mathrm{e}^{-i\omega n(u-a)}\,\mathrm{d}u \\ &= \mathrm{e}^{i\omega na}\frac{1}{T}\int_a^{T+a} f(u)\mathrm{e}^{-i\omega nu}\,\mathrm{d}u = \mathrm{e}^{i\omega na}c_n(f). \end{aligned}$$

Proposition 6. Si $f \in \mathcal{C}_T$ est une fonction de classe $\mathcal{C}^1$, alors :

$$c_n(f') = i\omega n c_n(f).$$

Démonstration. Les coefficients de f' sont donnés par :

$$c_n(f') = \frac{1}{T}\int_0^T f'(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t.$$

Une intégration par parties donne, en utilisant la périodicité de f :

$$c_n(f') = \underbrace{\big[f(t)\mathrm{e}^{-i\omega nt}\big]_0^T}_{=0} + i\omega n\int_0^T f(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t = i\omega n c_n(f).$$

Corollaire 7. Si $f \in \mathcal{C}_T$ est une fonction de classe $\mathcal{C}^\infty$, alors pour tout $k \geqslant 1$:

$$c_n(f^{(k)}) = (i\omega n)^k c_n(f).$$

1.4 Taille des coefficients de Fourier

Nous avons vu précédemment que si $f \in \mathcal{C}_T$, alors $c_n(f)$ tend vers 0 quand $n \to \pm\infty$. Si l'on fait des hypothèses de régularité plus fortes sur f, alors la « décroissance » des coefficients $c_n(f)$ est plus rapide.

Théorème 8. Si $f \in \mathcal{C}_T$ est une fonction de classe $\mathcal{C}^k$, alors :

$$\lim_{n\to\pm\infty}\big|n^k c_n(f)\big| = 0 \quad \text{et} \quad \sum_{n\in\mathbb{Z}}\big|n^k c_n(f)\big|^2 < +\infty.$$

Démonstration. La dérivée $f^{(k)}$ appartient à $\mathcal{C}_T$ et ses coefficients sont :

$$(i\omega n)^k c_n(f) = c_n(f^{(k)}) \underset{n\to\pm\infty}{\longrightarrow} 0.$$

L'inégalité de Bessel pour $f^{(k)}$ implique donc que :

$$\omega^{2k}\sum_{n\in\mathbb{Z}}\left|n^k c_n(f)\right|^2 = \sum_{n\in\mathbb{Z}}\left|c_n(f^{(k)})\right|^2 < +\infty.$$

Corollaire 9. Si $f \in \mathcal{C}_T$ est une fonction de classe $\mathcal{C}^\infty$, alors pour tout $\alpha \geqslant 0$, on a :

$$c_n(f) = \mathrm{o}\left(\frac{1}{n^\alpha}\right) \qquad \text{quand } n \to \pm\infty.$$

Remarque. On dit alors que les coefficients $c_n(f)$ sont à décroissance rapide.

On peut aller plus loin. À toute fonction $f \in \mathcal{C}_T$, on peut associer une fonction $F : \mathbb{U} \to \mathbb{C}$ définie par $F(\mathrm{e}^{i\omega t}) = f(t)$. Cette relation définit bien la fonction F de manière univoque car lorsque $\mathrm{e}^{i\omega t_1} = \mathrm{e}^{i\omega t_2}$, alors t_2 et t_1 diffèrent d'un multiple de T et $f(t_1) = f(t_2)$.

Exercice 1.

Supposons que $F(z) = \sum_{k\geqslant 0} a_k z^k$ soit une fonction développable en série entière avec un rayon de convergence $R > 1$ et que $f(t) = F(\mathrm{e}^{i\omega t})$.

1. Montrer que :
$$c_n(f) = \begin{cases} a_n & \text{si } n \geqslant 0 \\ 0 & \text{si } n < 0 \end{cases}.$$
2. Montrer que pour tout $r < R$:
$$\lim_{n\to+\infty}\left|r^n c_n(f)\right| = 0 \quad \text{et} \quad \sum_{n\geqslant 0}\left|r^n c_n(f)\right| < +\infty.$$

Solution.

1. Soit $r < R$. La série entière $\sum a_k z^k$ converge normalement, donc uniformément, sur le disque $\overline{D}_r := \left\{z \in \mathbb{C} \;\middle|\; |z| \leqslant r\right\}$, en particulier sur $\mathbb{U}$. Si pour $N \geqslant 0$ on pose :
$$F_N(z) = \sum_{k=0}^{N} a_k z^k \quad \text{et} \quad f_N(t) = F_N(\mathrm{e}^{i\omega t}),$$
alors la suite (f_N) converge uniformément sur $[0,T]$ vers f. On a :
$$f_N = \sum_{k=0}^{N} a_k \mathrm{e}_k.$$
Pour tout $n \in \mathbb{Z}$, on peut donc échanger limite et intégrale :
$$c_n(f) = \frac{1}{T}\int_0^T \lim_{N\to+\infty} f_N(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t = \lim_{N\to+\infty}\frac{1}{T}\int_0^T f_N(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t.$$

Or :

$$\frac{1}{T}\int_0^T f_N(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t = \left(\mathrm{e}_n \,\middle|\, \sum_{k=0}^{N} a_k \mathrm{e}_k\right) = \begin{cases} a_n & \text{si } n \in [\![0, N]\!] \\ 0 & \text{sinon} \end{cases}.$$

Par conséquent :

$$c_n(f) = \begin{cases} a_n & \text{si } n \geqslant 0, \\ 0 & \text{si } n < 0. \end{cases}$$

2. La série entière $\sum a_k z^k$ est normalement convergente sur $\overline{D}_r$ pour tout $r < R$.

1.5 Problème inverse

Supposons maintenant que $\{c_n\}_{n\in\mathbb{Z}}$ soit une famille de nombre complexe. La suite de fonctions $S_N := \sum_{n=-N}^{N} c_n \cdot \mathrm{e}_n$ est-elle convergente ? Si oui, les c_n sont-ils les coefficients de Fourier de la limite ?
Nous avons vu que les coefficients de Fourier d'une fonction continue $f \in \mathcal{C}_T$ vérifient :

$$\sum_{n\in\mathbb{Z}} \big|c_n(f)\big|^2 < +\infty.$$

Un exemple simple permet de se convaincre qu'une telle condition ne garantit la convergence de la série $\sum c_n \cdot \mathrm{e}_n$.

Exemple. Considérons la série de fonctions $\sum_{p=1}^{+\infty} \frac{1}{p}\cos(2\pi 10^p x)$, c'est-à-dire, considérons le cas où $T = 1$ et :

$$c_n = \begin{cases} \dfrac{1}{2|p|} & \text{si } n = \pm 10^p \text{ avec } p \geqslant 1 \\ 0 & \text{sinon.} \end{cases}$$

Alors :

$$\sum |c_n|^2 = \sum_{n=1}^{+\infty} \frac{1}{n^2} < +\infty.$$

Choisissons $x \in [0, 1[$ et considérons l'écriture décimale $x = \sum_{p=1}^{+\infty} \frac{x_p}{10^p}$ avec $x_p \in [\![0, 9]\!]$.

Si $x_{p+1} \in \{0, 1, 8, 9\}$, alors :

$$\cos(2\pi 10^p x) \geqslant \cos\frac{2\pi}{5} > 0.$$

En effet, le premier chiffre après la virgule de $10^p x$ est x_{p+1} et modulo 2π, l'angle $2\pi 10^p x$ appartient à $[0, 2\pi/5] \cup [4\pi/5, 2\pi]$. Par conséquent si le développement décimal de x ne

contient que les chiffres $0, 1, 8$ ou 9, alors :

$$\sum_{p=1}^{+\infty} \frac{1}{p} \cos(2\pi 10^p x) \geqslant \sum_{p=1}^{+\infty} \frac{1}{p} \cos \frac{2\pi}{5} = +\infty.$$

La série diverge. On montre de même que si le développement décimal de x ne contient que les chiffres 3, 4, 5 ou 6, alors :

$$\sum_{p=1}^{+\infty} \frac{1}{p} \cos(2\pi 10^p x) \leqslant \sum_{p=1}^{+\infty} \frac{1}{p} \cos \frac{3\pi}{5} = -\infty.$$

On dispose quand même d'une version faible qui permet de conclure à la convergence de la suite S_N dans certains cas.

Proposition 10. Si (c_n) est une suite de nombre complexes tels que $\sum |c_n| < +\infty$, alors la suite :

$$S_N := \sum_{n=-N}^{N} c_n \cdot \mathrm{e}_n$$

converge uniformément vers une limite $f \in \mathcal{C}_T$.

Démonstration. Comme la série $\sum |c_n|$ est convergente, la série $\sum c_n \cdot \mathrm{e}_n$ est normalement convergente sur $\mathbb{R}$.

Exercice 2.

Montrer que si $k \geqslant 0$ est un entier et si (c_n) est une suite de nombre complexes tels que $\sum |n^k c_n| < +\infty$, alors la suite $S_N := \sum_{n=-N}^{N} c_n \cdot \mathrm{e}_n$ converge uniformément sur $\mathbb{R}$ vers une limite $f \in \mathcal{C}_T$ qui est de classe $\mathcal{C}^k$ avec : $f^{(k)} = \sum_{n\in\mathbb{Z}} (i\omega n)^k c_n \cdot \mathrm{e}_n$.

Solution. Pour tout $\ell \leqslant k$, on a :

$$\sum_{n\neq 0} |n^\ell c_n| \leqslant \sum_{n\neq 0} |n^k c_n| < +\infty.$$

La dérivée $\ell^{ème}$ de S_N est :

$$S_N^{(\ell)} = \sum_{n=-N}^{N} (i\omega n)^\ell c_n \cdot \mathrm{e}_n.$$

Pour tout $\ell \leqslant k$, la série $\sum (i\omega n)^\ell c_n \cdot \mathrm{e}_n$ converge normalement (donc uniformément) sur $\mathbb{R}$. Le résultat suit par récurrence descendante sur ℓ en utilisant le théorème 14 de la page 483.

---

2 Convergence de la série de Fourier d'une fonction continue

Dans cette partie, nous allons étudier si une fonction $f \in \mathcal{C}_T$ est somme de sa série de Fourier. Nous verrons que ce n'est pas toujours le cas mais que si l'on impose que f est une fonction de classe $\mathcal{C}^1$, alors f est somme de sa série de Fourier.

2.1 Produit de convolution

Commençons par manipuler les sommes partielles de la série de Fourier pour les écrire sous une forme plus agréable à manipuler :

$$\begin{aligned} S_N(f)(x) &= \sum_{n=-N}^{N} c_n(f)\mathrm{e}_n(x) \\ &= \sum_{n=-N}^{N} \left(\frac{1}{T}\int_0^T f(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t\right)\mathrm{e}^{i\omega nx} \\ &= \frac{1}{T}\int_0^T f(t)\cdot\left(\sum_{n=-N}^{N}\mathrm{e}^{i\omega n(x-t)}\right)\mathrm{d}t. \end{aligned}$$

On a donc :

$$S_N(f)(x) = \frac{1}{T}\int_0^T f(t)\cdot D_N(x-t)\,\mathrm{d}t \quad \text{avec} \quad D_N(t) = \sum_{n=-N}^{N}\mathrm{e}^{i\omega nt}.$$

Une notion s'impose alors naturellement, la notion de produit de convolution.

Définition 7. Si $f \in \mathcal{C}_T$ et $g \in \mathcal{C}_T$, le produit de convolution de f par g est la fonction :

$$f * g : x \mapsto \frac{1}{T}\int_0^T f(t)\cdot g(x-t)\,\mathrm{d}t.$$

Nous allons dans un premier temps étudier quelques propriétés de ce produit de convolution.

Proposition 11. Soient f, g des fonctions de $\mathcal{C}_T$. Alors :

1. $f * g \in \mathcal{C}_T$.
2. $f * g = g * f$.
3. L'application $h \mapsto f * h$ est une application linéaire de $\mathcal{C}_T$ dans $\mathcal{C}_T$.
4. Pour tout $n \in \mathbb{Z}$, $f * \mathrm{e}_n = c_n(f)\cdot \mathrm{e}_n$.

Démonstration.

1. La périodicité de $f * g$ découle de la périodicité de g : $g(x+T-t) = g(x-t)$. La continuité de $f * g$ découle du théorème de continuité d'une fonction définie par une intégrale (voir le corollaire 127 de la page 643) : la fonction $(x,t) \mapsto f(t)g(x-t)$ est continue sur $[0,T]\times[0,T]$ et donc, $x \mapsto \displaystyle\int_0^T f(t)g(x-t)\,\mathrm{d}t$ est continue sur $[0,T]$ (donc sur $\mathbb{R}$ par périodicité).

2. L'égalité $f * g = g * f$ s'obtient en opérant le changement de variable $u = x - t$ et en utilisant la périodicité de f et g :

$$\begin{aligned} f * g(x) &= \frac{1}{T}\int_0^T f(t)g(x-t)\,\mathrm{d}t \\ &= -\frac{1}{T}\int_x^{x-T} f(x-u)g(u)\,\mathrm{d}u \\ &= \frac{1}{T}\int_0^T g(u)f(x-u)\,\mathrm{d}u = g * f(x). \end{aligned}$$

3. La linéarité de l'application $h \mapsto f * h$ est une conséquence immédiate de la linéarité de l'intégrale.
4. Si $n \in \mathbb{Z}$:

$$f * \mathrm{e}_n(x) = \frac{1}{T}\int_0^T f(t)\mathrm{e}^{i\omega n(x-t)}\,\mathrm{d}t = \left(\frac{1}{T}\int_0^T f(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t\right)\mathrm{e}^{i\omega nx} = c_n(f)\cdot \mathrm{e}_n(x).$$

Corollaire 12. Si $P := \sum_{n=-N}^{N} a_n \mathrm{e}_n$ est un polynôme trigonométrique, alors $f * P$ est également un polynôme trigonométrique.

Un résultat fondamental est celui du produit de convolution avec une unité approchée de $\mathcal{C}_T$.

Définition 8. Une unité approchée de $\mathcal{C}_T$ est une suite (h_n) de fonctions positives de $\mathcal{C}_T$ telle que :

- (h_n) converge uniformément vers 0 sur tout intervalle $[\delta, T-\delta]$, $0 < \delta < T/2$ et
- $\dfrac{1}{T}\displaystyle\int_0^T h_n = 1$ pour tout n.

Notons que si (h_n) est une unité approchée de $\mathcal{C}_T$, toute la masse de l'intégrale de h_n se concentre autour des multiples de T : pour tout $\delta \in]0, T/2[$, on a :

$$\lim_{n\to+\infty}\frac{1}{T}\int_\delta^{T-\delta} h_n = 0 \quad \text{et} \quad \lim_{n\to+\infty}\frac{1}{T}\int_{-\delta}^{\delta} h_n = 1.$$

Proposition 13. Si $f \in \mathcal{C}_T$ et si (h_n) est une unité approchée de $\mathcal{C}_T$, alors la suite $(f * h_n)$ converge uniformément vers f sur $\mathbb{R}$.

Démonstration. Si f est identiquement nulle, $f * h_n$ est identiquement nulle pour tout n, et il n'y a rien à montrer. Supposons donc que f n'est pas identiquement nulle, ce qui implique $2c_0(|f|) > 0$.

Fixons $\varepsilon > 0$. La fonction f est continue et périodique sur $\mathbb{R}$. Elle est donc uniformément continue. Par conséquent, il existe $\delta \in]0, T/2[$ tel que :

$$\big|f(x) - f(y)\big| \leqslant \varepsilon \quad \text{dès que} \quad |x-y| \leqslant 2\delta.$$

Choisissons n_0 suffisamment grand pour que :

$$h_n(t) \leqslant \frac{\varepsilon}{2c_0\big(|f|\big)} \quad \text{dès que} \quad n \geqslant n_0 \quad \text{et} \quad t \in [\delta, T-\delta].$$

Alors, pour $n \geqslant n_0$:

$$\begin{aligned}
\big|f * h_n(x) - f(x)\big| &= \left|\frac{1}{T}\int_0^T f(x-t)h_n(t)\,\mathrm{d}t - \frac{1}{T}\int_0^T f(x)h_n(t)\,\mathrm{d}t\right| \\
&\leqslant \frac{1}{T}\int_0^T \big|f(x-t)-f(t)\big|\cdot h_n(t)\,\mathrm{d}t \\
&= \frac{1}{T}\int_{-\delta}^{\delta} \underbrace{\big|f(x-t)-f(t)\big|}_{\leqslant\varepsilon}\cdot h_n(t)\,\mathrm{d}t \\
&\quad + \frac{1}{T}\int_{\delta}^{T-\delta} \big|f(x-t)-f(t)\big|\cdot \underbrace{h_n(t)}_{\leqslant\varepsilon/2c_0\left(|f|\right)}\,\mathrm{d}t \\
&\leqslant \varepsilon\underbrace{\frac{1}{T}\int_{-\delta}^{\delta} h_n(t)\,\mathrm{d}t}_{\leqslant 1} + \frac{\varepsilon}{2c_0\big(|f|\big)}\underbrace{\frac{1}{T}\int_{\delta}^{T-\delta}\big|f(x-t)\big| + \big|f(t)\big|\,\mathrm{d}t}_{\leqslant 2c_0\left(|f|\right)} \leqslant 2\varepsilon.
\end{aligned}$$

2.2 Formule de Parseval

Avant de régler le cas de la convergence de la série de Fourier, nous allons d'ores et déjà utiliser le résultat que nous venons d'établir pour montrer un théorème de Weierstraß.

Théorème 14 (Weierstraß). Toute fonction continue $f \in \mathcal{C}_T$ est limite uniforme sur $\mathbb{R}$ de polynômes trigonométriques.
En particulier, la famille $\{\mathrm{e}_n\}_{n\in\mathbb{Z}}$ est dense dans $\mathcal{C}_T$.

Démonstration. Pour $N \geqslant 0$, posons :

$$h_N(t) := \frac{1}{A_N}\cos^{2N}\left(\frac{\pi t}{T}\right) \quad \text{avec} \quad A_N := \frac{1}{T}\int_0^T \cos^{2N}\left(\frac{\pi t}{T}\right)\mathrm{d}t.$$

La formule du binôme de Newton permet de voir que h_N est un polynôme trigonométrique :

$$\begin{aligned}
\cos^{2N}\left(\frac{\pi t}{T}\right) &= \frac{1}{4^N}\left(\mathrm{e}^{i\pi t/T} + \mathrm{e}^{-i\pi t/T}\right)^{2N} \\
&= \frac{1}{4^N}\sum_{k=0}^{2N}\binom{2N}{k}\mathrm{e}^{i\pi kt/T}\mathrm{e}^{-i\pi(2n-k)t/T} \\
&= \frac{1}{4^N}\sum_{k=0}^{2N}\binom{2N}{k}\mathrm{e}^{i\omega(k-N)t} = \frac{1}{4^N}\sum_{n=-N}^{N}\binom{2N}{N+n}\mathrm{e}^{i\omega nt}.
\end{aligned}$$

De plus, le fait de diviser par A_N implique que :

$$\frac{1}{T}\int_0^T h_N(t)\,\mathrm{d}t = 1.$$

Enfin, fixons $\delta \in]0, T/2[$ et posons :

$$m := \sup_{t\in[\delta,T-\delta]} \left|\cos\frac{\pi t}{T}\right| \quad \text{et} \quad M := \inf_{t\in[-\delta/2,\delta/2]} \left|\cos\frac{\pi t}{T}\right|.$$

Alors, $m < M$ et :

$$A_N \geqslant \int_{-\delta/2}^{\delta/2} M^N = \delta \cdot M^N.$$

Par conséquent, sur $[\delta, T-\delta]$, on a :

$$h_N(t) \leqslant \frac{m^N}{\delta \cdot M^N} = \frac{1}{\delta}\left(\frac{m}{M}\right)^N \underset{N\to+\infty}{\longrightarrow} 0.$$

On voit donc que (h_N) est une unité approchée de $\mathcal{C}_T$.
La suite de polynôme trigonométriques $(f * h_N)$ converge donc uniformément vers f sur $\mathbb{R}$. Cela implique en particulier que :

$$d\big(f, \mathrm{Vect}(\mathrm{e}_{-N}, \dots, \mathrm{e}_0, \dots, \mathrm{e}_N)\big) \leqslant \|f * h_N - f\|_2 = \frac{1}{T}\int_0^T |f * h_N - f|^2 \underset{N\to+\infty}{\longrightarrow} 0.$$

L'espace vectoriel engendré par la famille $\{\mathrm{e}_n\}_{n\in\mathbb{Z}}$ est donc dense (pour la norme $\|\cdot\|_2$ induite par le produit hermitien) dans $\mathcal{C}_T$. Cela montre que la famille orthonormée $\{\mathrm{e}_n\}_{n\in\mathbb{Z}}$ est une base hermitienne de $\mathcal{C}_T$.

Puisque la famille $\{\mathrm{e}_n\}_{n\in\mathbb{Z}}$ est une base hermitienne de $\mathcal{C}_T$, on a la formule de Parseval (qui est l'analogue du théorème de Pythagore en dimension infinie).

Théorème 15 (Formule de Parseval). L'application $\mathcal{F} : \mathcal{C}_T \to \ell^2(\mathbb{Z})$ qui à f associe $\big(c_n(f)\big)_{n\in\mathbb{Z}}$ est une injection isométrique : pour toute fonction $f \in \mathcal{C}_T$, on a :

$$\|f\|_2^2 = \frac{1}{T}\int_0^T |f|^2 = \sum_{n\in\mathbb{Z}} \big|(\,\mathrm{e}_n \mid f\,)\big|^2 = \sum_{n\in\mathbb{Z}} \big|c_n(f)\big|^2,$$

qui donne, par polarisation :

$$(\,f \mid g\,) = \frac{1}{T}\int_0^T \overline{f}\cdot g = \sum_{n\in\mathbb{Z}} \overline{c_n(f)}\cdot c_n(g).$$

Exemple. On a vu que dans le cas de la fonction $f \in \mathcal{C}_{2\pi}$ qui coïncide avec $x \mapsto |x|$ sur $[-\pi, \pi]$, on a $c_0(f) = \dfrac{\pi}{2}$ et pour $p \geqslant 1$, $c_{\pm 2p}(f) = 0$ et $c_{\pm(2p-1)}(f) = \dfrac{2}{\pi(2p-1)^2}$.
On a donc :

$$\frac{\pi^2}{4} + 2\sum_{p=0}^{+\infty} \frac{4}{\pi^2(2p+1)^4} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |t|^2\,\mathrm{d}t = \frac{\pi^2}{3}.$$

On en déduit que :

$$\frac{8}{\pi^2}\sum_{p=0}^{+\infty}\frac{1}{(2p+1)^4}=\frac{\pi^2}{3}-\frac{\pi^2}{4}=\frac{\pi^2}{12}\quad\text{et}\quad\sum_{p=0}^{+\infty}\frac{1}{(2p+1)^4}=\frac{\pi^4}{96}.$$

De plus :

$$\zeta(4):=\sum_{n=1}^{+\infty}\frac{1}{n^4}=\sum_{p=1}^{+\infty}\frac{1}{(2p)^4}+\sum_{p=0}^{+\infty}\frac{1}{(2p+1)^4}=\frac{1}{16}\zeta(4)+\frac{\pi^4}{96}.$$

On en déduit que :

$$\zeta(4)=\frac{1}{1-1/16}\cdot\frac{\pi^4}{96}=\frac{\pi^4}{90}.$$

Corollaire 16.
Si $f\in\mathcal{C}_T$ et $g\in\mathcal{C}_T$ ont les mêmes coefficients de Fourier, alors $f=g$.

Démonstration. Les coefficients de $f-g$ sont nuls, donc $\|f-g\|_2=0$. ∎

Exemple. Si les coefficients $b_n(f)$ du développement en série de sinus et cosinus d'une fonction $f\in\mathcal{C}_T$ sont tous nuls, alors f est une fonction paire. En effet, les coefficients de la série de Fourier de la fonction $t\mapsto f(t)-f(-t)$ sont alors tous nuls, ce qui montre que $f(t)=f(-t)$ pour tout $t\in\mathbb{R}$.
De même, si les coefficients $a_n(f)$ sont tous nuls, alors la fonction $t\mapsto f(t)+f(-t)$ est identiquement nulle et f est impaire.

2.3 Théorème de Dirichlet

Revenons maintenant au problème de convergence de la série de Fourier, ou pour être plus précis, de la suite des sommes partielles $\big(S_N(f)\big)_{N\geqslant 0}$. Nous venons de voir que toute fonction de $\mathcal{C}_T$ est limite uniforme d'une suite de polynômes trigonométriques.
La densité du sous–espace vectoriel engendré par la famille $\{\mathrm{e}_n\}_{n\in\mathbb{Z}}$ implique que la suite des sommes partielles $S_N(f)$ de la série de Fourier d'une fonction $f\in\mathcal{C}_T$ converge vers f pour la norme induite par le produit scalaire.

Théorème 17 (Convergence en moyenne quadratique).
Pour tout $f\in\mathcal{C}_T$, on a :

$$\lim_{N\to+\infty}\frac{1}{T}\int_0^T\big|f-S_N(f)\big|^2=0.$$

Parmi les polynômes trigonométriques de degré N, celui qui approche le mieux f est la somme partielle $S_N(f)$ de la série de Fourier de f. Nous pourrions hâtivement en déduire que la suite des sommes partielles $S_N(f)$ converge uniformément vers f. Cela est faux. En effet, il faut faire attention que $S_N(f)$ est la meilleure approximation de f pour la norme induite par le produit hermitien, par pour la norme de la convergence uniforme.

Nous proposons d'ailleurs un exercice montrant qu'il existe des fonctions continues qui ne sont pas la somme de leur série de Fourier (exercice 4 de la page 679). Il se peut donc que la suite des sommes partielles $S_N(f)$ ne converge même pas simplement vers f.
Nous avons mentionné plus haut que la somme partielle $S_N(f)$ peut s'écrire :

$$S_N(f) = f * D_N \quad \text{avec} \quad D_N = \sum_{n=-N}^{N} \mathrm{e}_n.$$

Les fonctions D_N s'appellent les noyaux de Dirichlet. Il est possible de donner une valeur explicite des noyaux de Dirichlet car il s'agit en fait d'une série géométrique :

$$\begin{aligned} D_N(t) &= \sum_{n=-N}^{N} \mathrm{e}^{i\omega nt} = \mathrm{e}^{-i\omega Nt}\frac{1-\mathrm{e}^{i\omega(2N+1)t}}{1-\mathrm{e}^{i\omega t}} \\ &= \frac{\mathrm{e}^{i\omega(N+1/2)t}-\mathrm{e}^{-i\omega(N+1/2)t}}{\mathrm{e}^{i\omega t/2}-\mathrm{e}^{-i\omega t/2}} = \frac{\sin\bigl(\omega(N+1/2)t\bigr)}{\sin(\omega t/2)}. \end{aligned}$$

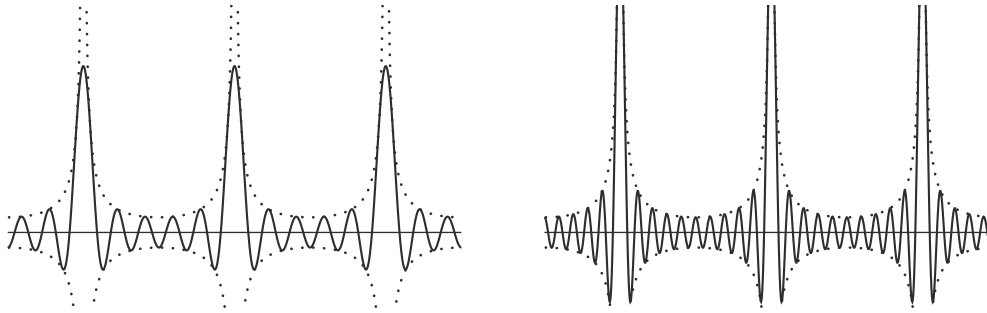


FIGURE II.4.2 – Les noyaux de Dirichlet pour $N = 5$ et $N = 10$.

La raison pour laquelle une fonction $f \in \mathcal{C}_T$ n'est pas nécessairement la somme de sa série de Fourier est que la suite (D_N) n'est pas une unité approchée. Les oscillations du noyau de Dirichlet D_N ne deviennent pas petites quand $N \to +\infty$. Bien au contraire, la fonction D_N oscille entre les graphes des fonctions $1/\sin(\omega t/2)$ et $-1/\sin(\omega t/2)$.
Nous allons maintenant démontrer une version faible d'un théorème de Dirichlet, que nous compléterons plus loin en traitant le cas de fonctions $\mathcal{C}^1$ par morceaux.

Théorème 18 (Dirichlet). Soit $f : \mathbb{R} \to \mathbb{C}$ une fonction T-périodique de classe $\mathcal{C}^1$. Alors, f est la somme de sa série de Fourier :

$$\forall x \in \mathbb{R}, \quad \lim_{N\to+\infty} S_N(f)(x) = f(x).$$

Démonstration. On écrit :

$$S_N(f)(x) - f(x) = f * D_N(x) - f(x) = \frac{1}{T}\int_0^T f(x-t)D_N(t)\,\mathrm{d}t - \frac{1}{T}\int_0^T f(x)D_N(t)\,\mathrm{d}t.$$

En effet, puisque la famille $\{\mathrm{e}_n\}_{n\in\mathbb{Z}}$ est orthonormée :

$$\frac{1}{T}\int_0^T D_N(t)\,\mathrm{d}t = (\,\mathrm{e}_0 \mid D_N\,) = \left(\,\mathrm{e}_0 \,\middle|\, \sum_{n=-N}^{N} \mathrm{e}_n\right) = 1.$$

On a donc :

$$S_N(f)(x) - f(x) = \frac{1}{T}\int_0^T \big(f(x-t)-f(x)\big)\cdot D_N(t)\,\mathrm{d}t.$$

Nous souhaitons montrer que cette quantité tend vers 0 quand N tend vers $+\infty$.
Remplaçons $D_N(t)$ par la valeur explicite calculée plus haut. On a :

$$\frac{1}{T}\int_0^T \big(f(x-t)-f(x)\big)\cdot D_N(t)\,\mathrm{d}t = \frac{1}{T}\int_0^T \frac{f(x-t)-f(x)}{\sin(\omega t/2)}\cdot\sin\left(\omega\left(N+\frac{1}{2}\right)t\right)\mathrm{d}t.$$

La fonction $t \mapsto \dfrac{f(x-t)-f(x)}{\sin(\omega t/2)}$ est périodique de période T, continue en dehors des multiples de T (là où le dénominateur s'annule). Comme f est $\mathcal{C}^1$, elle se prolonge par continuité en $t=0$ (et donc aux multiples de T par périodicité). En effet :

$$\frac{f(x-t)-f(x)}{\sin(\omega t/2)} = \frac{f(x-t)-f(x)}{t}\cdot\frac{t}{\sin(\omega t/2)} \xrightarrow[t\to 0]{} \frac{2f'(x)}{\omega}.$$

Notons $g:\mathbb{R}\to\mathbb{C}$ le prolongement par continuité.
On a alors :

$$S_N(f)(x)-f(x) = \frac{1}{T}\int_0^T g(t)\cdot\frac{\mathrm{e}^{i\omega(N+1/2)t}-\mathrm{e}^{-i\omega(N+1/2)t}}{2i}\,\mathrm{d}t = (\,\mathrm{e}_{-N}\mid g_1\,)-(\,\mathrm{e}_N\mid g_2\,)$$

avec $g_1:\mathbb{R}\to\mathbb{C}$ et $g_2:\mathbb{R}\to\mathbb{C}$ les fonctions de $\mathcal{C}_T$ définies par :

$$g_1(t) = \frac{1}{2i}g(t)\mathrm{e}^{i\omega t/2} \quad\text{et}\quad g_2(t) = \frac{1}{2i}g(t)\mathrm{e}^{-i\omega t/2}.$$

Comme $g_1\in\mathcal{C}_T$ et $g_2\in\mathcal{C}_T$, les coefficients de Fourier $c_{-N}(g_1)$ et $c_N(g_2)$ tendent vers 0 quand N tend vers l'infini. Cela montre bien que :

$$\lim_{N\to+\infty} S_N(f)(x)-f(x) = \lim_{N\to+\infty} c_{-N}(g_1)-c_N(g_2) = 0.$$

2.4 Théorème de Féjer

Nous venons de voir que les noyaux de Dirichlet ne forment pas une unité périodique approchée. À la fin du XIX$^{\mathrm{e}}$ siècle, l'italien Cesàro a l'idée de rendre convergentes des suites divergentes (u_n) en considérant les moyennes arithmétiques :

$$v_n = \frac{u_1+\cdots+u_n}{n}.$$

Par exemple, si l'on applique ce procédé à la suite (u_n) définie par $u_n := (-1)^n$, la suite (v_n) converge vers 0. Si la suite (u_n) converge, il est rassurant de constater que la suite v_n converge vers la même limite (voir exercice II.4.6 de la page 701).

Si l'on applique le procédé de Cesàro à la suite des sommes partielles $\big(S_N(f)\big)$ de la série de Fourier de f, on obtient une nouvelle suite :

$$T_N(f) := \frac{1}{N}\big(S_0(f) + S_1(f) + \ldots + S_{N-1}(f)\big) = \sum_{n=-N}^{N} \left(1 - \frac{|n|}{N}\right) c_n(f) \cdot \mathrm{e}_n.$$

Comme on l'a fait pour les sommes partielles de la série de Fourier, on peut écrire $T_N(f)$ à l'aide d'un produit de convolution :

$$T_N(f) = \frac{1}{N}\big(f * D_0 + f * D_1 + \ldots + f * D_{N-1}\big) = f * F_N$$

où les fonctions

$$F_N = \frac{1}{N}(D_0 + D_1 + \ldots + D_{N-1})$$

portent le nom de noyaux de Féjer.

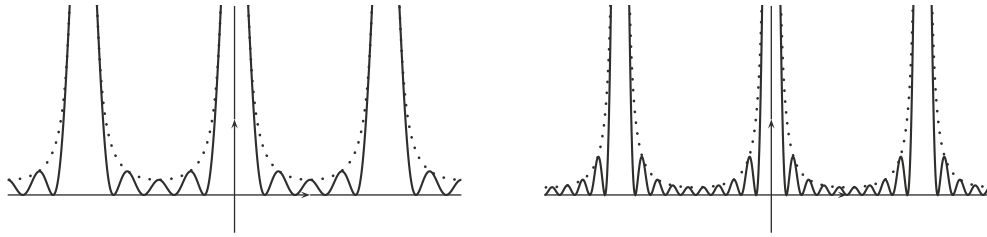


FIGURE II.4.3 – Les noyaux de Féjer pour $N = 5$ et $N = 10$.

Proposition 19.
La suite (F_N) des noyaux de Féjer est une unité périodique approchée.

Démonstration. Le noyau de Féjer se calcule explicitement :

$$\begin{aligned} F_N(t) &= \frac{\sin(\omega t/2) + \sin(3\omega t/2) + \ldots + \sin\big((2N+1)\omega t/2\big)}{N \sin(\omega t/2)} \\ &= \mathrm{Im}\left(\frac{\mathrm{e}^{i\omega t/2}(1 + \mathrm{e}^{i\omega t} + \ldots + \mathrm{e}^{i\omega(N-1)t})}{N \sin(\omega t/2)}\right) \\ &= \frac{1}{N \sin(\omega t/2)} \mathrm{Im}\left(\mathrm{e}^{i\omega t/2} \frac{1 - \mathrm{e}^{i\omega N t}}{1 - \mathrm{e}^{i\omega t}}\right). \end{aligned}$$

On a donc :

$$F_N(t) == \frac{1 - \cos(\omega N t)}{2N \sin^2(\omega t/2)} = \frac{1}{N}\left(\frac{\sin(\omega N t/2)}{\sin(\omega t/2)}\right)^2.$$

Notons que F_N est positive, que :

$$\frac{1}{T}\int_0^T F_N = (\, e_0 \mid F_N \,) = \frac{1}{N}\big((\, e_0 \mid D_0 \,) + \ldots + (\, e_0 \mid D_{N-1} \,)\big) = 1$$

et que pour tout $\delta \in]0, T/2[$, F_N converge uniformément vers 0 sur $[\delta, T-\delta]$ puisque :

$$0 \leqslant F_N(t) \leqslant \frac{1}{N \sin^2(\omega t/2)}.$$

Corollaire 20 (Théorème de Féjer).
Si $f \in \mathcal{C}_T$, les sommes de Féjer $T_n(f) := \dfrac{1}{N}\big(S_0(f) + \ldots + S_{N-1}(f)\big)$ convergent uniformément vers f.

Cela fournit une seconde démonstration que les polynômes trigonométriques sont denses dans $\mathcal{C}_T$.

Exercice 3.
Le procédé de Féjer consiste à pondérer le coefficient $c_n(f)$ avec un poids $(1 - |n|/N)$ dans la somme partielle :

$$T_n(f) := \sum_{n=-N}^{N} \left(1 - \frac{|n|}{N}\right) c_n(f) \cdot \mathrm{e}_n.$$

Montrer que considérer le produit de convolution :

$$f * B_N \quad \text{avec} \quad B_N(t) := \frac{\cos^{2N}(\omega t/2)}{\dfrac{1}{T}\displaystyle\int_0^T \cos^{2N}(\omega t/2)\,\mathrm{d}t}$$

revient à pondérer le coefficient $c_n(f)$ avec un poids $\binom{2N}{N+n}/\binom{2N}{N}$ dans la somme partielle :

$$f * B_N := \sum_{n=-N}^{N} \frac{\binom{2N}{N+n}}{\binom{2N}{N}} c_n(f) \cdot \mathrm{e}_n.$$

Solution. On a vu que $\cos^{2N}\left(\dfrac{\omega t}{2}\right) = \dfrac{1}{4^N} \displaystyle\sum_{n=-N}^{N} \binom{2N}{N+n} \cdot \mathrm{e}_n$. Comme la base e_n est orthonormée, on en déduit que $\dfrac{1}{T}\displaystyle\int_0^T \cos^{2N}\left(\frac{\omega t}{2}\right) \mathrm{d}t = \dfrac{\binom{2N}{N}}{4^N}$ et donc que :

$$B_n(t) = \sum_{n=-N}^{N} \frac{\binom{2N}{N+n}}{\binom{2N}{N}} \cdot \mathrm{e}_n.$$

Le reste suit du fait que $f * \mathrm{e}_n = c_n(f) \cdot \mathrm{e}_n$.

Il y a une leçon à tirer de cela. Lorsque l'on somme la série de Fourier en donnant le même poids à tous les coefficients $c_n(f)$ pour $n \in [\![-N, N]\!]$, la suite des sommes partielles obtenues converge mal vers f. On a intérêt à pondérer les coefficients $c_n(f)$ avec un poids qui diminue doucement vers le bord, c'est-à-dire quand $|n|$ se rapproche de N.

2.5 Une fonction continue qui n'est pas la somme de sa série de Fourier

Nous nous proposons de donner un exemple de fonction continue dont la série de Fourier diverge en 0.

Exercice 4.

1. Pour $N \geqslant 1$, posons $f_N(x) := \sum_{n=1}^{N} \frac{\sin(nx)}{n}$ (il s'agit de la $N^{ème}$ somme partielle de la série de Fourier d'une fonction en dents de scie). Étant donné $x \in]0, \pi]$, on note m_x la partie entière de $1/x$.
   (a) Montrer que $\sum_{n=1}^{m_x} \frac{\sin(nx)}{n} \leqslant 1$.
   (b) Pour $n \geqslant m_x + 1$, on pose $A_n(x) := \sum_{k=m_x+1}^{n} \sin(kx)$. Calculer $A_n(x)$ et montrer que $\left|A_n(x)\right| \leqslant \frac{1}{\left|\sin(x/2)\right|} \leqslant \frac{\pi}{x}$.
   (c) Montrer que pour $N \geqslant m_x + 1$, on a :
   $$\sum_{n=m_x+1}^{N} \frac{\sin(nx)}{n} = \frac{A_N(x)}{N} + \sum_{n=m_x+1}^{N-1} A_n(x) \left(\frac{1}{n} - \frac{1}{n+1}\right).$$
   En déduire que $\sum_{n=m_x+1}^{N} \frac{\sin(nx)}{n} \leqslant \pi$.
   (d) Montrer que pour tout $x \in \mathbb{R}$ et tout $N \geqslant 1$, on a $\left|f_N(x)\right| \leqslant 1 + \pi$.
2. On écrit $f_N = g_N + h_N$ avec :
   $$g_N(x) := \frac{1}{2i} \sum_{n=-N}^{-1} \frac{e^{inx}}{n} \quad \text{et} \quad h_N(x) := \frac{1}{2i} \sum_{n=1}^{N} \frac{e^{inx}}{n}.$$
   Montrer que pour tout $N \geqslant 1$, $\left|g_N(0)\right| \geqslant \frac{1}{2} \ln N$.
3. Pour $k \geqslant 1$, on pose $N_k := 3^{k^3}$ et :
   $$F(x) := \sum_{k=1}^{+\infty} \frac{f_{N_k}(x) e^{i2N_k x}}{k^2}.$$
   (a) Montrer que la série est normalement convergente et en déduire que F est continue.
   (b) Montrer que $\left|S_{2N_m}(F)(0)\right| \underset{m \to +\infty}{\longrightarrow} +\infty$.

Solution.

1. (a) Il suffit d'utiliser l'inégalité $\sin(nx) \leqslant nx$ valable pour $x \geqslant 0$. On a alors $\displaystyle\sum_{n=1}^{m_x} \frac{\sin(nx)}{n} \leqslant m_x \cdot x \leqslant 1$.

   (b) $\displaystyle A_n(x) = \frac{\sin\dfrac{(n+m_x)x}{2}\sin\dfrac{(n-m_x)x}{2}}{\sin\dfrac{x}{2}}$.

   Si $x \in]0,\pi]$, $\sin(x/2) \geqslant \frac{1}{\pi}x$, ce qui donne le résultat.

   (c) L'égalité demandée est donnée par la transformation d'Abel. On a donc :

$$\sum_{n=m_x+1}^{N} \frac{\sin(nx)}{n} \leqslant \frac{\pi}{x}\left(\frac{1}{N} + \sum_{n=m_x+1}^{N-1}\left(\frac{1}{n} - \frac{1}{n+1}\right)\right) = \frac{\pi}{x(m_x+1)} \leqslant \pi.$$

   (d) Puisque f_N est impaire et 2π-périodique, il suffit de le montrer pour $x \in [0,\pi]$. La fonction s'annule en 0 et les résultats précédents montrent que pour $x \in]0,\pi]$, $\big|f_N(x)\big| \leqslant 1+\pi$.

2. On a : $\big|g_N(0)\big| = \dfrac{1}{2}\displaystyle\sum_{n=1}^{N}\frac{1}{n} \geqslant \frac{1}{2}\sum_{n=1}^{N}\int_n^{n+1}\frac{\mathrm{d}t}{t} = \frac{1}{2}\ln(N+1) > \frac{1}{2}\ln N.$

3. (a) Le terme général de la série est borné par $(1+\pi)/k^2$ qui est le terme d'une série convergente. La somme F est donc une limite uniforme de fonctions continues. Elle est continue.

   (b) Notons que le $k^{\text{ème}}$ terme de la série définissant F est un polynôme trigonométrique, combinaison linéaires de e^{inx} avec $N_k \leqslant n \leqslant 3N_k$. Notons également que $3N_k < N_{k+1}$. Chaque terme de la série définissant F ne contribue donc qu'à un seul coefficient de la série de Fourier de F. On a :

$$S_{2N_m}(F) = \underbrace{\frac{f_{N_1}\cdot \mathrm{e}_{2N_1}}{1^2}}_{\substack{\text{contribue aux } c_n(F)\\ \text{pour } N_1\leqslant n\leqslant 3N_1}} + \cdots + \underbrace{\frac{f_{N_{m-1}}\cdot \mathrm{e}_{2N_{m-1}}}{(m-1)^2}}_{\substack{\text{contribue aux } c_n(F)\\ \text{pour } N_{m-1}\leqslant n\leqslant 3N_{m-1}}} + \underbrace{\frac{g_{N_m}\cdot \mathrm{e}_{2N_m}}{m^2}}_{\substack{\text{contribue aux } c_n(F)\\ \text{pour } N_m\leqslant n\leqslant 2N_m}}.$$

   Comme pour tout $N \geqslant 1$, $f_N(0) = 0$, on a alors :

$$\big|S_{2N_m}(F)(0)\big| = \left|\frac{g_{N_m}(0)}{m^2}\right| \geqslant \frac{\ln N_m}{2m^2} = \frac{m\ln 3}{2} \underset{m\to+\infty}{\longrightarrow} +\infty.$$

3 Série de Fourier d'une fonction réglée

Ce que nous avons raconté jusqu'à présent reste valide dans un cadre plus général que celui des fonctions continues. L'espace naturel dans lequel la théorie de Fourier prend place est celui des fonctions qui sont de carré intégrable sur une période, c'est-à-dire, telles que :

$$\frac{1}{T}\int_0^T |f|^2 < +\infty.$$

Nous n'avons pas encore développé suffisamment les outils d'intégration pour pouvoir travailler dans ce cadre (qui sera abordé dans le module *Intégration* de [L2-L3]). Nous nous contenterons de généraliser les résultats au cas des fonctions réglées.
Rappelons (*cf.* le module II.3, page 590) qu'une application $f : I \to \mathbb{R}$ est *réglée* dans un intervalle $I \subset \mathbb{R}$ d'intérieur non vide si f possède des valeurs limites à gauche et à droite en tout point de I. On note alors $f(x^+)$ la limite à droite et $f(x^-)$ la limite à gauche.
Si $f : [a, b] \to \mathbb{R}$ est réglée, alors

- f peut être approchée uniformément par des applications en escaliers,
- l'ensemble des points de discontinuité de f est dénombrable et
- f est intégrable au sens de Riemann.

3.1 Coefficients de Fourier, formule de Parseval

Définition 9. Nous noterons $\mathcal{R}_T$ l'ensemble des fonctions $f : \mathbb{R} \to \mathbb{C}$ qui sont réglées dans $\mathbb{R}$ et T-périodiques.

Notons que $\mathcal{R}_T$ est un sous-espace vectoriel de l'espace des fonctions de $\mathbb{R}$ dans $\mathbb{C}$: la fonction identiquement nulle est réglée, la somme de deux fonctions réglée est réglée et le produit d'une fonction réglée par un scalaire est une fonction réglée. Notons également que le produit de deux fonctions réglées est une fonction réglée et que le conjugué complexe d'une fonction réglée est lui-même réglé.
Les définitions des coefficients de Fourier, de la série de Fourier, de sa $N^{ème}$ somme partielle sont inchangées.

Définition 10.
Si $f \in \mathcal{R}_T$ les coefficients de Fourier de f sont les nombres complexes :

$$c_n(f) := \frac{1}{T}\int_0^T f(t)\mathrm{e}^{-i\omega nt}\,\mathrm{d}t.$$

La $N^{ème}$ somme partielle de la série de Fourier est :

$$S_N(f) = \sum_{n=-N}^{N} c_n(f)\cdot \mathrm{e}_n.$$

Exercice 5.
Calculer les coefficients de Fourier de la fonction créneau $f \in \mathcal{R}_{2\pi}$ définie par :

$$f(x) = \begin{cases} +\pi/4 & \text{pour } |x| < \pi/2 \\ 0 & \text{pour } |x| = \pi/2 \\ -\pi/4 & \text{pour } \pi/2 < |x| < \pi \end{cases}$$

et déterminer la $N^{ème}$ somme partielle de la série de Fourier.

Solution. Si $n = 0$, on a :

$$c_0(f) = \frac{1}{2\pi}\left(\int_{-\pi}^{-\pi/2} -\frac{\pi}{4} + \int_{-\pi/2}^{\pi/2} \frac{\pi}{4} + \int_{\pi/2}^{\pi} -\frac{\pi}{4}\right) = 0.$$

Sinon :

$$\begin{aligned} c_n(f) &= \frac{1}{2\pi}\left(\int_{-\pi}^{-\pi/2} -\frac{\pi}{4}\mathrm{e}^{-int}\,\mathrm{d}t + \int_{-\pi/2}^{\pi/2} \frac{\pi}{4}\mathrm{e}^{-int}\,\mathrm{d}t + \int_{\pi/2}^{\pi} -\frac{\pi}{4}\mathrm{e}^{-int}\,\mathrm{d}t\right) \\ &= \frac{1}{8}\left(\left[-\frac{\mathrm{e}^{-int}}{-in}\right]_{-\pi}^{-\pi/2} + \left[\frac{\mathrm{e}^{-int}}{-i}\right]_{-\pi/2}^{\pi/2} + \left[-\frac{\mathrm{e}^{-int}}{-in}\right]_{\pi/2}^{\pi}\right) \\ &= \frac{1}{8in}\left(\mathrm{e}^{in\pi/2} - \mathrm{e}^{in\pi} - \mathrm{e}^{-in\pi/2} + \mathrm{e}^{in\pi/2} + \mathrm{e}^{-in\pi} - \mathrm{e}^{-in\pi/2}\right) \\ &= \frac{1}{4in}\left(\mathrm{e}^{in\pi/2} - \mathrm{e}^{-in\pi/2}\right) = \frac{\sin(n\pi/2)}{2n}. \end{aligned}$$

Autrement dit :

$$c_{2p} = 0 \quad \text{et} \quad c_{2p+1} = \frac{(-1)^p}{2(2p+1)}.$$

On a alors :

$$\begin{aligned} S_N(f)(x) &= \sum_{n=-N}^{N} c_n(f)\cdot \mathrm{e}^{inx} = \sum_{p=1}^{\lfloor\frac{N-1}{2}\rfloor} \frac{(-1)^p}{2p+1}\cdot\frac{\mathrm{e}^{i(2p+1)x} + \mathrm{e}^{-i(2p+1)x}}{2} \\ &= \sum_{p=1}^{\lfloor\frac{N-1}{2}\rfloor} \frac{(-1)^p}{2p+1}\cos\bigl((2p+1)x\bigr). \end{aligned}$$

On retrouve la série de Fourier annoncée dans l'introduction :

$$\cos x - \frac{1}{3}\cos(3x) + \frac{1}{5}\cos(5x) - \frac{1}{7}\cos(7x) + \cdots$$

Si pour $f \in \mathcal{R}_T$ et $g \in \mathcal{R}_T$, on pose :

$$(\,f \mid g\,) := \frac{1}{T}\int_0^T \overline{f}\cdot g,$$

alors $c_n(f) = (\,\mathrm{e}_n \mid f\,)$. On définit ainsi une forme sesquilinéaire hermitienne positive. Cependant, cette forme n'est pas un produit scalaire car elle n'est pas définie positive. Par exemple, la fonction qui vaut 0 sur $\mathbb{R}\setminus\mathbb{Z}$ et 1 sur $\mathbb{Z}$ appartient à $\mathcal{R}_1$ et $(\,f \mid f\,) = 0$ bien que f ne soit pas la fonction nulle.
Cela a pour conséquence que deux fonctions de $\mathcal{R}_T$ peuvent avoir les mêmes coefficients de Fourier sans être égales. C'est un problème mineur qu'il est possible de contourner, par exemple en travaillant dans un espace quotient pour la relation d'équivalence :

$$f \in \mathcal{R}_T \sim g \in \mathcal{R}_T \quad\Longleftrightarrow\quad (\,f-g \mid f-g\,) = 0.$$

On peut aussi, et c'est ce que nous ferons ici, restreindre l'étude au sous-espace vectoriel $\underline{\mathcal{R}}_T \subset \mathcal{R}_T$ des fonctions telles qu'en tout point $x \in \mathbb{R}$, on a :

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

Cela revient à choisir une fonction préférentielle dans chaque classe d'équivalence. En effet si deux fonctions réglées sont égales sauf sur un ensemble dénombrable, leurs intégrales sont égales (*cf.* la proposition 75 de la page 595), et l'on dispose du résultat suivant.

Proposition 21. Si $f \in \underline{\mathcal{R}}_T$, alors $(\, f \mid f \,) = 0$ si, et seulement si, $f \equiv 0$.

Démonstration. Il est clair que si $f \equiv 0$, alors $(\, f \mid f \,) = 0$.

Supposons donc que $(\, f \mid f \,) = 0$. Notons que si $f \in \underline{\mathcal{R}}_T$ et $g \in \underline{\mathcal{R}}_T$, alors $\overline{f} \in \underline{\mathcal{R}}_T$ et $\overline{f}g \in \underline{\mathcal{R}}_T$; en particulier $|f|^2 \in \underline{\mathcal{R}}_T$.

Supposons par l'absurde qu'il existe un point $x_0 \in \mathbb{R}$ tel que $f(x_0) \neq 0$. Alors :

$$\big|f(x_0^-)\big|^2 + \big|f(x_0^+)\big|^2 = 2\big|f(x_0)\big|^2 > 0.$$

Donc, $\big|f(x_0^-)\big|^2 > 0$ ou $\big|f(x_0^+)\big|^2 > 0$. Supposons pour simplifier que $\big|f(x_0^-)\big|^2 > 0$. On peut alors trouver $\delta \in \,]0, T[$ tel que pour tout $x \in \,]x_0 - \delta, x_0[$, on ait $\big|f(x)\big|^2 \geqslant \frac{1}{2}\big|f(x_0^-)\big|^2$. Dans ce cas :

$$\frac{1}{T}\int_0^T |f|^2 \geqslant \frac{1}{T}\int_{x_0-\delta}^{x_0} |f|^2 \geqslant \frac{\delta}{2T}\big|f(x_0^-)\big|^2 > 0.$$

On a donc une contradiction. ∎

L'espace $\underline{\mathcal{R}}_T$ muni du produit scalaire $(\cdot \mid \cdot)$ est donc un espace préhilbertien complexe. Nous noterons $\|\cdot\|_2$ la norme associée au produit scalaire $(\cdot \mid \cdot)$, c'est-à-dire $\|f\|_2^2 := (\, f \mid f \,)$.

Proposition 22.
Le sous-espace vectoriel engendré par la famille $\{e_n\}_{n \in /Z}$ est dense dans $\underline{\mathcal{R}}_T$.

Démonstration. Il s'agit de montrer que l'espace vectoriel des polynômes trigonométriques est dense dans $\underline{\mathcal{R}}_T$ pour la norme $\|\cdot\|_2$. Nous savons qu'il est dense dans $\mathcal{C}_T$. Toute fonction de $f \in \underline{\mathcal{R}}_T$ peut être uniformément approximée par une suite (h_n) de fonctions en escalier. La suite $\big(|f - h_n|^2\big)$ converge alors uniformément vers 0 et donc, $\|f - h_n\|_2 \to 0$. Il suffit donc de montrer que pour toute fonction en escalier $h : \mathbb{R} \to \mathbb{R}$ qui est T-périodique, il existe une suite (g_n) de fonctions continues T-périodiques telle que :

$$\frac{1}{T}\int_0^T |h - g_n|^2 \underset{n \to +\infty}{\longrightarrow} 0.$$

Comme les fonctions g_n sont continues mais que h n'est pas continue, la suite (g_n) ne peut pas converger uniformément vers h.

On se restreint à un segment $[a,b]$ de longueur T en sorte que a et b ne soient pas des points de discontinuité de h. On appelle $a_1 < a_2 < \ldots < a_N$ les points de discontinuité de h sur $[a,b]$. Si n est suffisamment grand, les segments $I_k := [a_k - 1/n, a_k + 1/n]$ sont disjoints et contenus dans $[a,b]$. On note g_n la fonction continue et T-périodique qui sur $[a,b]$ coïncide avec h en dehors des segments I_k et qui est affine sur chaque segment I_k.

Si $M := \max\big|h(a_k^+) - h(a_k^-)\big|$ est la hauteur du plus grand saut de h, alors $|h - g_n|$ est une fonction nulle sur $[a,b]$ en dehors des segments I_k et majoré par M sur $I_k \setminus \{a_k\}$. Puisque la longueur de chacun des N segments I_k est égale à $2/n$, on en déduit que :

$$\|h - g_n\|_2^2 := \frac{1}{T}\int_a^b |h - g_n|^2 \leqslant \frac{2NM^2}{n} \underset{n\to+\infty}{\longrightarrow} 0.$$

On voit donc que si $f \in \underline{\mathcal{R}}_T$, alors on a la formule de Parseval :

$$\sum_{n\in\mathbb{Z}} \big|c_n(f)\big|^2 = \frac{1}{T}\int_0^T |f|^2$$

et, par polarisation :

$$(\, f \mid g \,) = \sum_{n\in\mathbb{Z}} \overline{c_n(f)} \cdot c_n(g).$$

Proposition et définition 23. Si $f \in \mathcal{R}_T$, il existe une unique fonction $\underline{f} \in \underline{\mathcal{R}}_T$ qui coïncide avec f en dehors des points de discontinuité de f. On dit que $\underline{f}$ est la régularisée de f.

Démonstration. La fonction $\underline{f}$ est nécessairement la fonction définie par :

$$\underline{f}(x) = \begin{cases} f(x) & \text{si } x \text{ est est un point de continuité de } f \\ \dfrac{f(x^+) + f(x^-)}{2} & \text{si } x \text{ est un point de discontinuité de } f. \end{cases}$$

Il faut et il suffit donc de montrer que cette fonction appartient à $\underline{\mathcal{R}}_T$.

Pour tout $x_0 \in \mathbb{R}$ et tout $\varepsilon > 0$, il existe $\delta > 0$ tel que $f(x) \in \big]f(x_0^+) - \varepsilon, f(x_0^+) + \varepsilon\big[$ si $x \in]x_0, x_0 + \delta[$ et $f(x) \in \big]f(x_0^-) - \varepsilon, f(x_0^-) + \varepsilon\big[$ si $x \in]x_0 - \delta, x_0[$. On en déduit que $f(x^\pm) \in \big]f(x_0^+) - \varepsilon, f(x_0^+) + \varepsilon\big[$ si $x \in]x_0, x_0 + \delta[$ et $f(x^\pm) \in \big]f(x_0^-) - \varepsilon, f(x_0^-) + \varepsilon\big[$ si $x \in]x_0 - \delta, x_0[$. On voit alors que la fonction $\underline{f}$ définie par la formule précédente vérifie : $\underline{f}(x) \in \big]f(x_0^+) - \varepsilon, f(x_0^+) + \varepsilon\big[$ si $x \in]x_0, x_0 + \delta[$ et $\underline{f}(x) \in \big]f(x_0^-) - \varepsilon, f(x_0^-) + \varepsilon\big[$ si $x \in]x_0 - \delta, x_0[$. Cela montre que $\underline{f}$ possède des limites à gauche et à droite en x_0 qui sont égales à $f(x_0^+)$ et $f(x_0^-)$. La fonction $\underline{f}$ appartient donc à $\underline{\mathcal{R}}_T$ comme requis.

Étant donné que f et $\underline{f}$ coïncident en dehors d'un ensemble dénombrable, leurs coefficients de Fourier sont égaux et $(\, f \mid f \,) = (\, \underline{f} \mid \underline{f} \,)$. On en déduit que la formule de Parseval est valide pour les fonctions réglées.

Théorème 24 (Formule de Parseval). Si $f \in \mathcal{R}_T$, on a la formule de Parseval :

$$\sum_{n\in\mathbb{Z}} \big|c_n(f)\big|^2 = \frac{1}{T}\int_0^T |f|^2,$$

qui donne, par polarisation :

$$(\, f \mid g\,) = \sum_{n\in\mathbb{Z}} \overline{c_n(f)}\cdot c_n(g).$$

Démonstration. On a :

$$\sum_{n\in\mathbb{Z}} \big|c_n(f)\big|^2 = \sum_{n\in\mathbb{Z}} \big|c_n(\underline{f})\big|^2 = (\,\underline{f} \mid \underline{f}\,) = (\, f \mid f\,).$$

Exemple. En appliquant ce résultat à la fonction créneau, on obtient :

$$2\sum_{p=1}^{+\infty} \frac{1}{4(2p+1)^2} = \frac{1}{2\pi}\int_0^{2\pi} \frac{\pi^2}{16} = \frac{\pi^2}{16}.$$

Autrement dit :

$$\sum_{p=1}^{+\infty} \frac{1}{(2p+1)^2} = \frac{\pi^2}{8}.$$

On a alors :

$$\begin{aligned}\zeta(2) := \sum_{n=1}^{+\infty} \frac{1}{n^2} &= \sum_{p=1}^{+\infty} \frac{1}{(2p)^2} + \sum_{p=1}^{+\infty} \frac{1}{(2p+1)^2} \\ &= \frac{1}{4}\zeta(2) + \frac{\pi^2}{8}.\end{aligned}$$

D'où :

$$\zeta(2) = \frac{1}{1-1/4}\cdot\frac{\pi^2}{8} = \frac{\pi^2}{6}.$$

Corollaire 25. Si $f \in \mathcal{R}_T$, alors $c_n(f) \underset{n\to\pm\infty}{\longrightarrow} 0$.

3.2 Théorèmes de Dirichlet et de Féjer

Nous allons maintenant montrer que les théorèmes de Dirichlet et de Féjer s'étendent au cas des fonctions réglées. La limite (quand elle existe) des sommes partielles de Fourier et de Féjer d'une fonction $f \in \mathcal{R}_T$ est la régularisée $\underline{f} \in \underline{\mathcal{R}_T}$ de f. Cela tient au fait que les noyaux D_N de Dirichlet et F_N de Féjer sont des fonctions paires. En particulier :

$$\frac{1}{T}\int_0^{T/2} D_N = \frac{1}{T}\int_{-T/2}^{0} D_N = \frac{1}{2} \quad \text{et} \quad \frac{1}{T}\int_0^{T/2} F_N = \frac{1}{T}\int_{-T/2}^{0} F_N = \frac{1}{2}.$$

Théorème 26 (Dirichlet). Soit f une fonction de $\mathcal{R}_T$ et soit $x \in \mathbb{R}$ un point tel que les deux limites suivantes existent :

$$\lim_{y\to x^+} \frac{f(y)-f(x^+)}{y-x} \quad \text{et} \quad \lim_{y\to x^-} \frac{f(y)-f(x^-)}{y-x}.$$

Alors :

$$\lim_{N\to+\infty} S_N(x) = \frac{1}{2}\big(f(x^-)+f(x^+)\big).$$

Démonstration. On reprend la démonstration du théorème de Dirichlet pour les fonctions continues, en remplaçant l'intégration sur $[0,T]$ par l'intégration sur $[-T/2,0]\cup[0,T/2]$. Cela donne :

$$\begin{aligned} S_N(f)(x) - \frac{1}{2}\big(f(x^-)+f(x^+)\big) =& \frac{1}{T}\int_{-T/2}^{0} \big(f(x-t)-f(x^+)\big)\cdot D_N(t)\,\mathrm{d}t \\ &+ \frac{1}{T}\int_{0}^{T/2} \big(f(x-t)-f(x^-)\big)\cdot D_N(t)\,\mathrm{d}t. \end{aligned}$$

On appelle g la fonction définie par :

$$g(t) = \begin{cases} \dfrac{f(x-t)-f(x^+)}{\sin(\omega t/2)} & \text{si } t\in\,]-T/2,0[\\ \dfrac{f(x-t)-f(x^-)}{\sin(\omega t/2)} & \text{si } t\in\,]0,T/2] \\ 0 & \text{si } t=0 \end{cases}$$

et prolongée à $\mathbb{R}$ par périodicité. Les hypothèses faites sur la fonction f garantissent que g est une fonction réglée. On conclut alors comme dans la démonstration du théorème de Dirichlet pour les fonctions continues que :

$$S_N(f)(x) - \frac{1}{2}\big(f(x^-)+f(x^+)\big) = c_{-N}(g_1) - c_N(g_2)$$

avec g_1 et g_2 les fonctions réglées définie par :

$$g_1(t) = \frac{1}{2i}g(t)\mathrm{e}^{i\omega t/2} \quad \text{et} \quad g_2(t) = \frac{1}{2i}g(t)\mathrm{e}^{-i\omega t/2}.$$

Cela permet de conclure puisque $c_{-N}(g_1) \underset{N\to+\infty}{\longrightarrow} 0$ et $c_N(g_2) \underset{N\to+\infty}{\longrightarrow} 0$. ∎

Corollaire 27. Soit $f:\mathbb{R}\to\mathbb{C}$ une fonction périodique de classe $\mathcal{C}^1$ par morceaux. Alors :

$$\lim_{N\to+\infty} S_N(f)(x) = \frac{1}{2}\big(f(x^-)+f(x^+)\big).$$

Exemple. Considérons la fonction $f\in\underline{\mathcal{R}}_1$ définie par $f(x)=x$ si $x\in\,]-1/2,1/2[$ et $f(1/2)=0$.

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

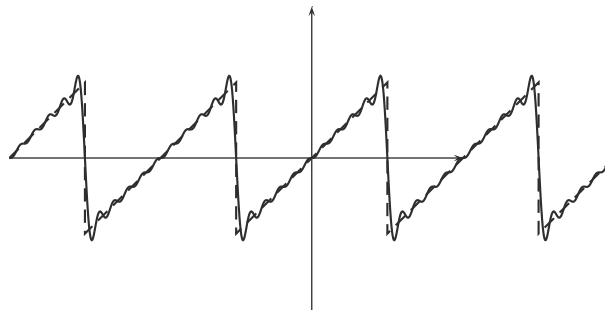


FIGURE II.4.4 – La somme partielle $S_{10}(f)$ d'une fonction en dents de scie.

Alors $c_0(f) = 0$ (car f est impaire) et une intégration par parties donne :

$$c_n(f) = \int_{-1/2}^{1/2} t\mathrm{e}^{-i2\pi nt}\,\mathrm{d}t = \left[t\frac{\mathrm{e}^{-i2\pi nt}}{-i2\pi n}\right]_{-1/2}^{1/2} + \frac{1}{i2\pi}\int_{-1/2}^{1/2}\mathrm{e}^{i2\pi nt}\,\mathrm{d}t = \frac{(-1)^n}{i2\pi n}\cdot$$

On a donc :

$$S_N(f)(x) = \sum_{n=1}^{N}\frac{(-1)^n}{i2\pi n}\big(\mathrm{e}^{i2\pi nx} - \mathrm{e}^{-i2\pi nx}\big) = \sum_{n=1}^{N}\frac{(-1)^n}{\pi n}\sin(2\pi nx).$$

Comme f est sa propre régularisée et qu'en tout point, elle admet une dérivée à droite et une dérivée à gauche, on a :

$$\lim_{N\to+\infty} S_N(f)(x) = f(x).$$

Passons maintenant au théorème de Féjer.

Théorème 28 (Féjer). Si $f \in \mathcal{R}_T$, alors la suite des sommes partielles de Féjer $T_N(f) := f * F_N$ converge simplement vers la régularisée $\underline{f}$ de f. Si f est continue sur un intervalle ouvert I, la convergence est uniforme sur tout segment $[a,b] \subset I$.

Démonstration. Si f est identiquement nulle, le résultat est évident. On peut donc supposer que f n'est pas identiquement nulle et donc, que $c_0\big(|f|\big) \neq 0$.

Nous avons vu que la suite (F_N) des noyaux de Féjer est une unité approchée. De plus, les noyaux de Féjer sont des fonctions paires. On a donc :

$$\frac{1}{T}\int_0^{T/2} F_N = \frac{1}{2} \quad \text{et} \quad \frac{1}{T}\int_{-T/2}^{0} F_N = \frac{1}{2}\cdot$$

On reprend la démonstration de la proposition 13 de la page 671 en la modifiant légèrement. Étant donné $x \in \mathbb{R}$, on choisit $\delta \in \,]0, T/2[$ en sorte que :

$$\big|f(x^-) - f(y)\big| \text{ si } y \in \,]x-\delta, x[\quad \text{et} \quad \big|f(x^+) - f(y)\big| \text{ si } y \in \,]x, x+\delta[\,.$$

Notons que si f est continue sur un intervalle ouvert I et si $[a,b] \subset I$, un tel δ peut être choisi indépendamment de $x \in [a,b]$. On utilise pour cela la continuité uniforme de f sur $[a',b'] \subset I$ avec $a' < a$ et $b' > b$ et l'on choisit $\delta < \min(a-a', b'-b)$ tel que $\big|f(x)-f(y)\big| < \delta$ dès que $x \in [a',b']$ et $y \in [a',b']$ vérifient $|x-y| < \delta$.
Choisissons N_0 suffisamment grand pour que :

$$F_N(t) \leqslant \frac{\varepsilon}{2c_0\big(|f|\big)} \quad \text{dès que} \quad N \geqslant N_0 \quad \text{et} \quad t \in [\delta, T-\delta].$$

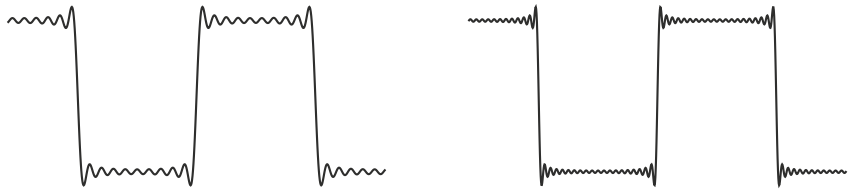
Alors, pour $N \geqslant N_0$:

$$\begin{aligned}\left|f * F_N(x) - \frac{1}{2}\big(f(x^-)+f(x^+)\big)\right| &= \left|\frac{1}{T}\int_{-T/2}^{0}\big(f(x-t)-f(x^+)\big)\cdot F_N(t)\,\mathrm{d}t\right. \\ &\qquad \left. + \frac{1}{T}\int_{0}^{T/2}\big(f(x-t)-f(x^-)\big)\cdot F_N(t)\,\mathrm{d}t\right| \\ &\leqslant 2\varepsilon.\end{aligned}$$

La majoration est la même que dans la démonstration de la proposition 13 de la page 671. Nous laissons au lecteur le soin de le vérifier.

3.3 Phénomène de Gibbs

Nous allons maintenant étudier un phénomène curieux, dû au fait que le noyau de Dirichlet n'est pas une unité approchée. Si l'on observe les sommes partielles $S_N(f)$ d'une fonction créneau f, on constate que, près de la discontinuité, le graphe de $S_N(f)$ dépasse du graphe de f d'une quantité appréciable. En ajoutant plus de termes de la série, le dépassement ne disparaît pas et ne diminue pas.



$N = 20$ $N = 40$

Nous nous proposons d'étudier ce phénomène, appelé le phénomène de Gibbs, dans l'exercice suivant.

Exercice 6.
Considérons la fonction $f \in \mathcal{R}_1$ qui vaut 1 si $0 < x < 1/2$ et -1 si $1/2 < x < 1$.

1. Montrer que la série de Fourier est :

$$\sum_{n>0, n \text{ impair}} \frac{4}{\pi}\frac{\sin(2\pi n x)}{n}.$$

2. Calculer le premier maximum M_p de la $2p^{\text{ème}}$ somme partielle sur $[0,1]$.

3. Montrer que
$$\lim_{p\to+\infty} M_p = \frac{2}{\pi}\int_0^1 \frac{\sin(\pi x)}{x}\,\mathrm{d}x > 1.$$

Solution.

1. Comme la fonction f est périodique de période 1 et impaire, il est possible d'écrire la série de Fourier comme somme de sinus :
$$\sum_{n=1}^{+\infty} b_n(f)\sin(nt) \quad \text{avec} \quad b_n(f) := 2\int_0^1 f(t)\sin(2\pi nt)\,\mathrm{d}t.$$
On a donc :
$$\begin{aligned} b_n(f) &= 2\int_0^{1/2}\sin(2\pi nt)\,\mathrm{d}t - 2\int_{1/2}^1 \sin(2\pi nt)\,\mathrm{d}t \\ &= \left[\frac{-\cos(2\pi nt)}{\pi n}\right]_0^{1/2} - \left[\frac{-\cos(2\pi nt)}{\pi n}\right]_{1/2}^1 \\ &= \frac{\big(-(-1)^n+1\big)-\big(1-(-1)^n\big)}{\pi n} \\ &= 2\frac{1-(-1)^n}{\pi n} = \begin{cases} 0 & \text{si } n \text{ est pair} \\ \dfrac{4}{\pi n} & \text{si } n \text{ est impair.}\end{cases}\end{aligned}$$

2. La $2p^{\,ème}$ somme partielle de la série de Fourier est donnée par :
$$S_{2p}(f)(x) = \sum_{k=0}^{p-1}\frac{4}{\pi}\frac{\sin\big(2\pi(2k+1)x\big)}{2k+1}\cdot$$
La dérivée de $S_{2p}(f)$ est égale à :
$$\begin{aligned} S_{2p}(f)'(x) &= 8\sum_{k=0}^{p-1}\cos\big(2\pi(2k+1)x\big) \\ &= \mathrm{Re}\left(8\sum_{k=0}^{p-1}\mathrm{e}^{i2\pi(2k+1)x}\right) = \mathrm{Re}\left(8\mathrm{e}^{i2\pi x}\frac{1-\mathrm{e}^{i4\pi px}}{1-\mathrm{e}^{i4\pi x}}\right) \\ &= \mathrm{Re}\left(8\mathrm{e}^{i2\pi px}\frac{\sin(\pi px)}{\sin(\pi x)}\right) = 8\cos\big(2\pi px\big)\frac{\sin(\pi px)}{\sin(\pi x)}\cdot\end{aligned}$$
La première fois que $S_{2p}(f)'$ s'annule sur $[0,1]$, c'est lorsque $x = \dfrac{1}{4p}\cdot$ La valeur de $S_{2p}(f)$ en ce point est égale à :
$$M_p = \frac{4}{\pi}\sum_{k=0}^{p-1}\frac{1}{2k+1}\sin\left(\pi\frac{(2k+1)}{2p}\right)\cdot$$

3. Si l'on pose $a_k := \dfrac{k}{p}$ pour $k \in [\![0, p+1]\!]$ et $x_k := \dfrac{2k+1}{2p} \in [a_k, a_{k+1}]$ pour $k \in [\![0, p]\!]$, on peut écrire :

$$M_p = \frac{2}{\pi} \cdot \left(\sum_{k=0}^{p-1} \frac{\sin(\pi x_k)}{x_k} \cdot (a_{k+1} - a_k) \right).$$

On reconnaît entre parenthèse une somme de Riemann de la fonction $x \mapsto \frac{\sin(\pi x)}{x}$ sur le segment $[0, 1]$. Cette fonction est continue sur le segment $[0, 1]$, donc intégrable au sens de Riemann. On a donc convergence des sommes de Riemann vers l'intégrale :

$$\lim_{p \to +\infty} M_p = \frac{2}{\pi} \int_0^1 \frac{\sin(\pi x)}{x} \, \mathrm{d}x.$$

La valeur de $\dfrac{2}{\pi} \displaystyle\int_0^1 \frac{\sin(\pi x)}{x} \, \mathrm{d}x$ n'est pas calculable à l'aide des fonctions élémentaires. Un calcul sur ordinateur donne une valeur approchée de $1.1789797976\ldots$.

---

4 Quelques applications

Nous allons maintenant illustrer la multitude d'applications des séries de Fourier sur quelques exemples. La liste n'est pas exhaustive. Nous allons présenter la plupart des applications sous forme d'exercices.

4.1 Équation de la chaleur

Comme nous l'avons signalé dans l'introduction, la motivation de Fourier pour développer des fonctions en séries de sinus et de cosinus est intimement liée à la résolution de l'équation de la chaleur. Nous nous proposons d'expliciter ce lien dans l'exercice suivant.

Exercice 7.

Nous allons étudier la propagation de la chaleur dans une barre de longueur L avec la contrainte que les extrémités sont gardées en permanence à température nulle. La barre est modélisée par le segment $[0, L]$. Nous noterons $u(x, t)$ la température de la tige en l'abscisse x au temps t. L'équation aux dérivées partielles associée à ce modèle unidimensionnel est l'équation de la chaleur :

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} \quad \text{avec} \quad u(0, t) = u(L, t) = 0 \quad \text{avec} \quad a \neq 0.$$

Nous allons résoudre ce problème dans le cas où la température est soumise à la condition initiale $u(x, 0) = \varphi(x)$, avec φ est une fonction de classe $\mathcal{C}^2$ sur $[0, L]$.

1. On commence par rechercher les solutions $u : \mathbb{R}^2 \to \mathbb{R}$ de la forme :

$$u(x, t) = f(x) \cdot g(t)$$

avec f une fonction de classe $\mathcal{C}^2$ sur $[0, L]$ et g une fonction de classe $\mathcal{C}^1$ sur $[0, +\infty[$. On suppose de plus que u n'est pas identiquement nulle sur $\mathbb{R}^2$.

(a) Montrer que :

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}$$

si, et seulement s'il existe une constante $C \in \mathbb{R}$ telle que f et g soient solutions des équations différentielles :

$$a^2 f'' = Cf \quad \text{et} \quad g' = Cg.$$

(b) Résoudre ces équations différentielles.

(c) Montrer que si $f(0) = f(L) = 0$, il existe un entier $n \geqslant 1$ et un réel b tels que :

$$u(x,t) = b \sin\left(\frac{\pi n x}{L}\right) \mathrm{e}^{-\left(\frac{2\pi a n}{L}\right)^2 t}.$$

2. Soit φ une fonction de classe $\mathcal{C}^2$ sur $[0, L]$ telle que $\varphi(0) = \varphi(L) = 0$ et $\varphi'(0) = \varphi'(L) = 0$. et $\tilde{\varphi}$ l'unique fonction impaire $2L$ périodique qui coïncide avec φ sur $[0, L]$.

(a) Montrer que la série de Fourier de $\tilde{\varphi}$ peut s'exprimer comme somme de sinus et exprimer les coefficients $b_n := b_n(\tilde{\varphi})$ à l'aide de φ.

(b) Montrer que la série $\displaystyle\sum_{n=1}^{+\infty} b_n \sin\left(\frac{\pi n x}{L}\right) \mathrm{e}^{-\left(\frac{2\pi a n}{L}\right)^2 t}$ définit une fonction $u(x,t)$ continue sur $[0, L] \times [0, +\infty[$.

(c) Montrer que $u(x, 0) = \varphi(x)$ pour tout $x \in [0, L]$ et que $u(0, t) = u(L, t) = 0$ pour tout $t \geqslant 0$.

(d) Montrer que u est $\mathcal{C}^\infty$ sur $[0, L] \times]0, +\infty[$ et que :

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}.$$

Solution.

1. (a) Choisissons un point (x_0, t_0) tel que $u(x_0, t_0) \neq 0$. Alors, $f(x_0) \neq 0$ et $g(t_0) \neq 0$. On a :

$$f(x) \cdot g'(t) = \frac{\partial u}{\partial t}(x,t) = a^2 \frac{\partial^2 u}{\partial x^2}(x,t) = a^2 f''(x) \cdot g(t).$$

En particulier, en (x_0, t_0) on obtient :

$$\frac{g'(t_0)}{g(t_0)} = a^2 \frac{f''(x_0)}{f(x_0)}.$$

De plus, en (x, t_0) et (x_0, t), on obtient :

$$f(x) \cdot g'(t_0) = a^2 f''(x) \cdot g(t_0) \quad \text{et} \quad f(x) \cdot g'(t_0) = a^2 f''(x) \cdot g(t_0).$$

On obtient le résultat en posant :

$$C := \frac{g'(t_0)}{g(t_0)} = a^2 \frac{f''(x_0)}{f(x_0)}.$$

(b) Les solutions de l'équation différentielle $g' = Cg$ sont de la forme :

$$g(t) = g(0) \cdot \mathrm{e}^{Ct}.$$

- Si $C/a^2 = \omega^2 > 0$ avec $\omega > 0$, les solutions de l'équation différentielle $a^2 f'' = Cf$ sont de la forme :

$$f(t) = A\mathrm{e}^{\omega t} + B\mathrm{e}^{-\omega t} \quad \text{avec} \quad (A, B) \in \mathbb{R}^2.$$

- Si $C/a^2 = 0$, les solutions de l'équation différentielle $a^2 f'' = Cf$ sont de la forme :

$$f(t) = At + B \quad \text{avec} \quad (A, B) \in \mathbb{R}^2.$$

- Si $C/a^2 = -\omega^2 < 0$ avec $\omega > 0$, les solutions de l'équation différentielle $a^2 f'' = Cf$ sont de la forme :

$$f(t) = A\sin(\omega t) + B\cos(\omega t) \quad \text{avec} \quad (A, B) \in \mathbb{R}^2.$$

(c) Pour que $f(0) = f(L) = 0$, il faut que $C < 0$. En effet, si $C > 0$ on doit avoir :

$$\begin{cases} A + B = 0 \\ \mathrm{e}^{\omega L} A + \mathrm{e}^{-\omega L} B = 0. \end{cases}$$

Le déterminant du système est $\mathrm{e}^{-\omega L} - \mathrm{e}^{\omega L} = -2\sinh(\omega L) \neq 0$. Par conséquent, la seule solution est la solution triviale $A = B = 0$, ce qui ne conduit pas à une solution f non identiquement nulle. Si $C = 0$, alors on doit avoir :

$$\begin{cases} B = 0 \\ AL + B = 0. \end{cases}$$

Cela conduit de nouveau à la solution triviale $A = B = 0$.
On a donc $C < 0$ et :

$$\begin{cases} B = 0 \\ A\sin(\omega L) + B = 0. \end{cases}$$

On voit donc que $\sin(\omega L) = 0$, ce qui restreint les valeurs possibles pour ω, qui doit être égal à $n\dfrac{\pi}{L}$ pour un entier $n \geqslant 1$.
On voit alors que g est solution de l'équation différentielle :

$$g' = -a^2\omega^2 g$$

et donc

$$g(t) = g(0) \cdot \mathrm{e}^{-a^2\omega^2 t}.$$

Au final :

$$u(x,t) = f(x) \cdot g(t) = b\sin\left(\frac{\pi n x}{L}\right) \mathrm{e}^{-\left(\frac{2\pi a n}{L}\right)^2 t}$$

avec b un réel et $n \geqslant 1$ un entier.

2. (a) La fonction $\tilde{\varphi}$ est continue (car φ s'annule en 0 et L). Elle admet donc un développement en série de Fourier. Étant donné que la fonction $\tilde{\varphi}$ est impaire, sa série de Fourier peut s'exprimer comme somme de sinus. On a alors :
$$b_n(\tilde{\varphi}) = \frac{2}{2L}\int_{-L}^{L} \tilde{\varphi}(t)\sin\left(\frac{2\pi}{2L}nt\right)\mathrm{d}t = \frac{2}{L}\int_0^L \varphi(t)\sin\left(\frac{\pi nt}{L}\right)\mathrm{d}t.$$

   (b) La fonction φ est de classe $\mathcal{C}^2$ sur $[0,L]$. Étant donné que $\varphi(0)=\varphi(L)=0$ et $\varphi'(0)=\varphi'(L)=0$, la fonction $\tilde{\varphi}$ est $\mathcal{C}^1$, sa dérivée $\tilde{\varphi}'$ est $\mathcal{C}^1$ par morceaux et $\tilde{\varphi}''$ est continue sur $[0,L]$. Deux intégrations par parties successives montrent que :
$$b_n := b_n(\tilde{\varphi}) = -\left(\frac{L}{\pi n}\right)^2 b_n(\tilde{\varphi}'') = \mathrm{o}\left(\frac{1}{n^2}\right).$$
   Pour tout $x \in [0,L]$ et tout $t \geqslant 0$, la série $\displaystyle\sum_{n=1}^{+\infty} b_n \sin\left(\frac{\pi n x}{L}\right)\mathrm{e}^{-\left(\frac{2\pi a n}{L}\right)^2 t}$ est donc normalement convergente sur $[0,L]\times[0,+\infty[$ (le terme général est borné par $|b_n|$). Cela montre que la somme $u(x,t)$ est définie et continue sur $[0,L]\times[0,+\infty[$.

   (c) On a trivialement $u(x,0) = \tilde{\varphi}(x) = \varphi(x)$ pour $x \in [0,L]$ puisque la série de Fourier de $\tilde{\varphi}$, qui est $\mathcal{C}^1$, converge vers $\tilde{\varphi}$. On voit également que pour tout $t \geqslant 0$, $u(0,t) = u(L,t) = 0$ puisque tous les termes de la série s'annulent quand $x = 0$ ou $x = L$.

   (d) Si $t > 0$, on a :
$$b_n \mathrm{e}^{-\left(\frac{2\pi a n}{L}\right)^2 t} = \mathrm{o}\left(\frac{1}{n^\alpha}\right)$$
   pour tout $\alpha \geqslant 0$. On vérifie facilement que si l'on dérive terme à terme k fois par rapport à x ou à t, le coefficient est multiplié par une puissance de n. La série des dérivées $k^{ème}$ converge donc normalement sur $[0,L]\times[t_0,+\infty[$ pour tout $t_0 > 0$. Cela montre que u est de classe $\mathcal{C}^\infty$ sur $[0,L]\times]0,+\infty[$ et qu'on a le droit de dériver terme à terme les séries. Comme chaque terme est solution de l'équation de la chaleur, il en est donc de même pour la somme de la série.

4.2 Théorème ergodique

On dénote par $R_\alpha : \mathbb{U} \to \mathbb{U}$ la rotation d'angle $2\pi\alpha$:
$$R_\alpha(z) = \mathrm{e}^{2i\pi\alpha} z.$$
L'orbite d'un point $z \in \mathbb{U}$ sous itération de R_α est la suite :
$$\big(z_0 = z, z_1 = R_\alpha(z_0), z_2 = R_\alpha(z_1), \ldots\big).$$
Le théorème qui suit affirme dans une certaine mesure que l'orbite de tout point $z \in \mathbb{U}$ s'équidistribue bien dans $\mathbb{U}$ au cours du temps, en sorte que si $\varphi : \mathbb{U} \to \mathbb{C}$ est une application continue, alors moyenner les valeurs de φ le long de l'orbite d'un point $z \in \mathbb{U}$ revient à moyenner les valeurs de φ sur le cercle.

Théorème 29 (Théorème ergodique). Soit $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Pour toute fonction continue $\varphi : \mathbb{U} \to \mathbb{C}$ et tout point $z \in \mathbb{U}$, on a l'égalité :

$$\lim_{m\to+\infty} \frac{1}{m} \sum_{k=0}^{m-1} \varphi(z_k) = \int_0^1 \varphi(\mathrm{e}^{2i\pi t})\,\mathrm{d}t \quad \text{avec} \quad z_k := R_\alpha^{\circ k}(z).$$

Démonstration. En écrivant $z = \mathrm{e}^{2i\pi x}$ et $f(t) = \varphi(\mathrm{e}^{2i\pi t})$, on voit qu'il est équivalent de démontrer que pour tout fonction $f \in \mathcal{C}_1$, et tout réel x, on a :

$$\lim_{m\to+\infty} \frac{1}{m} \sum_{k=0}^{m-1} f(x+k\alpha) = \int_0^1 f.$$

Fixons $x \in \mathbb{R}$ et posons :

$$T_m(f) := \frac{1}{m} \sum_{k=0}^{m-1} f(x+k\alpha).$$

Le théorème 14 de la page 672 (de Weierstraß) affirme que toute fonction $f \in \mathcal{C}_1$ peut être approximée uniformément par des polynômes trigonométriques : pour tout $\varepsilon > 0$, on peut trouver un polynôme trigonométrique $P : t \mapsto \sum_{n=-N}^{N} c_n \mathrm{e}^{i2\pi nt}$ tel que $|P - f| \leqslant \varepsilon$ sur $\mathbb{R}$.

On a alors :

$$\big|T_m(f) - T_m(P)\big| \leqslant \varepsilon \quad \text{et} \quad \left|\int_0^1 f - \int_0^1 P\right| \leqslant \varepsilon.$$

On va montrer que si m est suffisamment grand :

$$\left|T_m(P) - \int_0^1 P\right| \leqslant \varepsilon.$$

On en déduit immédiatement que si n est suffisamment grand,

$$\left|T_m(f) - \int_0^1 f\right| \leqslant 3\varepsilon,$$

ce qu'il fallait démontrer.

Nous restreignons donc dorénavant notre étude au cas où $f = P$ est un polynôme trigonométrique. Comme le problème est linéaire, il suffit de ne considérer que le cas où $P := \mathrm{e}_n : t \mapsto \mathrm{e}^{i2\pi nt}$ avec $n \in \mathbb{Z}$. Si $n \neq 0$, alors :

$$T_m(\mathrm{e}_n) = \frac{1}{m} \sum_{k=0}^{m-1} \mathrm{e}^{i2\pi n(x+k\alpha)} = \frac{\mathrm{e}^{i2\pi nx}}{m} \cdot \frac{1-\mathrm{e}^{i2\pi mn\alpha}}{1-\mathrm{e}^{i2\pi n\alpha}} \underset{m\to+\infty}{\longrightarrow} 0 = \int_0^1 \mathrm{e}_n.$$

Observons que nous utilisons là le fait que $n \neq 0$ et que α est irrationnel, car sinon le dénominateur pourrait s'annuler pour certaines valeurs de $n \neq 0$.

Quand $n = 0$, $\mathrm{e}_0 \equiv 1$ et :

$$T_m(\mathrm{e}_0) = 1 \underset{m\to+\infty}{\longrightarrow} 1 = \int_0^1 \mathrm{e}_0.$$

∎

4.3 Inégalité isopérimétrique

L'inégalité isopérimétrique affirme que parmi tous les domaines du plan d'aire A fixée, ceux qui ont le plus petit périmètre sont les disques de rayon $\sqrt{A/\pi}$. L'aire est alors reliée au périmètre L par la relation : $A \leqslant \dfrac{L^2}{4\pi}$. Nous allons montrer ce résultat sous l'hypothèse que le bord du domaine est une courbe paramétrée :

$$\begin{array}{rccc} \gamma : & [0,L] & \longrightarrow & \mathbb{R}^2 \\ & t & \longmapsto & \big(x(t),y(t)\big) \end{array}$$

de classe $\mathcal{C}^1$ n'ayant que des points réguliers. Dans ce cas, l'aire de la région délimitée par γ est donnée par la formule de Green-Riemann :

$$A = \int_0^L x(t)\cdot y'(t)\,\mathrm{d}t.$$

Notons que, quitte à reparamétrer la courbe, on peut supposer que le paramétrage est normal, c'est-à-dire que $\big(x'(t)\big)^2 + \big(y'(t)\big)^2 = 1$ pour tout $t \in [0,L]$. Notons enfin que γ se prolonge par périodicité à $\mathbb{R}$ tout entier et que le prolongement est encore une courbe continue de classe $\mathcal{C}^1$ par morceaux.

Exercice 8.

Soit :

$$\begin{array}{rccc} \gamma : & [0,L] & \longrightarrow & \mathbb{R}^2 \\ & t & \longmapsto & \big(x(t),y(t)\big) \end{array}$$

une courbe paramétrée de classe $\mathcal{C}^1$. Supposons que γ soit L-périodique et que :

$$\big(x'(t)\big)^2 + \big(y'(t)\big)^2 = 1$$

pour tout $t \in \mathbb{R}$. Les fonctions x et y sont L-périodiques. On veut montrer que :

$$A := \int_0^L x(t)\cdot y'(t)\,\mathrm{d}t \leqslant \frac{L^2}{4\pi}\cdot$$

1. Montrer que l'on peut se ramener au cas où $\displaystyle\int_0^L x(t)\,\mathrm{d}t = \int_0^L y(t)\,\mathrm{d}t = 0$.
2. Montrer que $\displaystyle\int_0^L \big(x(t)\big)^2\,\mathrm{d}t \leqslant \frac{L^2}{4\pi^2}\int_0^L \big(x'(t)\big)^2\,\mathrm{d}t$.
3. Montrer que $A^2 \leqslant \dfrac{L^2}{4\pi^2}\sup\limits_{u\in\mathbb{R}} u\cdot(L-u)$.
4. Conclure.

Démonstration.

1. Posons :

$$x_0 := \frac{1}{L}\int_0^L x(t)\,\mathrm{d}t \quad \text{et} \quad y_0 := \frac{1}{L}\int_0^L y(t)\,\mathrm{d}t.$$

Si l'on pose $X := x - x_0$ et $Y := y - y_0$, alors $\displaystyle\int_0^L X(t)\,\mathrm{d}t = \int_0^L Y(t)\,\mathrm{d}t = 0$. De plus :

$$\int_0^L X(t)\cdot Y'(t)\,\mathrm{d}t = \int_0^L x(t)\cdot y'(t)\,\mathrm{d}t - \int_0^L x_0\cdot y'(t)\,\mathrm{d}t = A$$

car $\displaystyle\int_0^L x_0\cdot y'(t) = x_0\cdot\big(y(L)-y(0)\big) = 0$.

2. Comme x est une fonction L-périodique continue de classe $\mathcal{C}^1$ par morceaux, elle est développable en série de Fourier et sa dérivée aussi. On a alors :

$$c_0(x') = c_0(x) = 0 \quad \text{et} \quad c_n(x') = i\frac{2\pi}{L}nc_n(x) \text{ pour } n\neq 0.$$

En particulier, on a :

$$\big|c_n(x)\big| \leqslant \frac{L}{2\pi}\big|c_n(x')\big|.$$

La formule de Parseval nous dit alors que :

$$\int_0^L \big(x(t)\big)^2\,\mathrm{d}t = L\cdot\sum_{n\in\mathbb{Z}}\big|c_n(x)\big|^2 \leqslant L\cdot\frac{L^2}{4\pi^2}\sum_{n\in\mathbb{Z}}\big|c_n(x')\big|^2 = \frac{L^2}{4\pi^2}\int_0^L\big(x'(t)\big)^2\,\mathrm{d}t.$$

3. D'après l'inégalité de Cauchy-Schwarz, on a donc :

$$\begin{aligned}
A^2 &= \left(\int_0^L x(t)\cdot y'(t)\,\mathrm{d}t\right)^2\\
&\leqslant \left(\int_0^L \big(x(t)\big)^2\,\mathrm{d}t\right)\cdot\left(\int_0^L\big(y'(t)\big)^2\,\mathrm{d}t\right)\\
&= \left(\int_0^L \big(x(t)\big)^2\,\mathrm{d}t\right)\cdot\left(\int_0^L 1-\big(x'(t)\big)^2\,\mathrm{d}t\right)\\
&\leqslant \left(\frac{L^2}{4\pi^2}\int_0^L\big(x'(t)\big)^2\,\mathrm{d}t\right)\cdot\left(L-\int_0^L\big(x'(t)\big)^2\,\mathrm{d}t\right).
\end{aligned}$$

on a donc bien :

$$A^2 \leqslant \frac{L^2}{4\pi^2}\sup_{u\in\mathbb{R}} u\cdot(L-u).$$

4. Le polynôme $P(u) := u\cdot(L-u)$ atteint son maximum en $L/2$. La valeur de ce maximum est $L^2/4$. Par conséquent :

$$A^2 \leqslant \frac{L^2}{4\pi^2}\cdot\frac{L^2}{4} = \left(\frac{L^2}{4\pi}\right)^2.$$

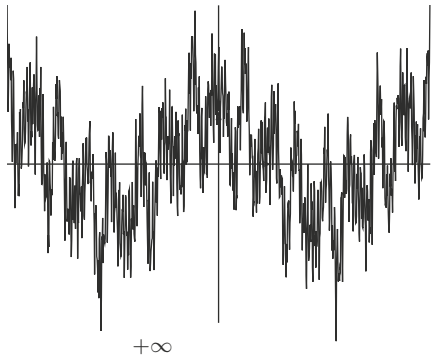
4.4 Séries lacunaires

Pour conclure ce module, nous souhaitons montrer comment on peut utiliser les séries de Fourier pour construire un exemple de fonction continue sur $\mathbb{R}$ qui n'est dérivable nulle

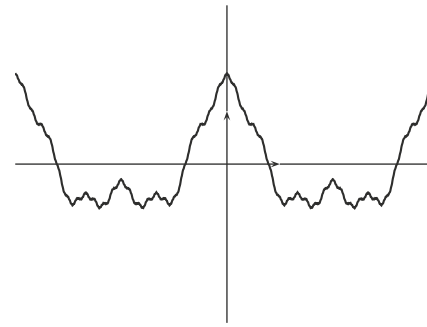
part. Le premier exemple a été construit par Weierstraß : si $0 < a < 1$ et $ab > 1 + \dfrac{3\pi}{2}$, la fonction définie par la série normalement convergente :

$$\sum_{n=0}^{+\infty} a^n \cos(b^n x)$$

est continue mais nulle part dérivable.



$$x \mapsto \sum_{n=0}^{+\infty} (3/4)^n \cos(8^n x)$$



$$x \mapsto \sum_{n=1}^{+\infty} \frac{\cos(n!x)}{n!}$$

FIGURE II.4.5 – Fonctions continues mais nulle part dérivable

Nous proposons dans l'exercice suivant d'étudier le cas de la série $\displaystyle\sum_{n=1}^{+\infty} \frac{\cos(n!x)}{n!}$.

Exercice 9.

On considère la fonction f définie par $f(x) := \displaystyle\sum_{n\geqslant 1} \frac{\cos(n!x)}{n!}$.

1. Montrer que la fonction f est définie et continue sur $\mathbb{R}$.

Étant donné $x_0 \in \mathbb{R}$, on pose $u = x - x_0$ et

$$g(u) := \frac{f(x) - f(x_0)}{x - x_0} = \sum_{n\geqslant 1} g_n(u) \quad \text{avec} \quad g_n(u) := \frac{\cos(n!x_0 + n!u) - \cos(n!x_0)}{n!u}.$$

On appelle *oscillation* d'une fonction h sur un intervalle I la quantité :

$$\sup_I h - \inf_I h.$$

2. Fixons $k \geqslant 1$ et posons $I_k := [x_0, x_0 + 4\pi/k!]$.
   (a) Montrer que pour tout $n < k$ l'oscillation de g_n sur I_k est inférieure à $4\pi n!/k!$.
   (b) Montrer que pour $n > k$ l'oscillation de g_n sur I_k est inférieure à $2k!/(\pi n!)$.
   (c) Montrer que l'oscillation de g_k sur I_k est supérieure à $1/(2\pi)$.

3. (a) Montrer que pour tout $k \geqslant 1$:
$$\sum_{1\leqslant n<k} \frac{n!}{k!} \leqslant \frac{2}{k} \quad \text{et} \quad \sum_{n>k} \frac{k!}{n!} \leqslant \frac{1}{k}\cdot$$
(b) En déduire que l'oscillation de g sur I_k ne tend pas vers 0 quand $k \to +\infty$.
(c) En déduire que la fonction f n'est pas dérivable en x_0.

Solution.

1. La série est normalement convergente car le terme général est majoré par $1/n!$. La somme est donc définie et continue.
2. (a) La fonction g_n est continue sur le segment I_k. Elle atteint donc son maximum en un point $a \in I_k$ et son minimum en un point $b \in I_k$. Notons que $g_n(u)$ est le taux d'accroissement de la fonction cosinus entre $n!x_0$ et $n!x_0+n!u$. D'après le théorème des accroissements finis, $g_n(a) = -\sin(a')$ avec a' compris entre $n!x_0$ et $n!x_0 + n!a$. De même, $g_n(b) = -\sin(b')$ avec b' compris entre $n!x_0$ et $n!x_0 + n!a$. L'oscillation de g_n est égale à $g_n(a) - g_n(b) = \sin(b') - \sin(a')$. On utilise une seconde fois le théorème des accroissements finis pour voir que l'oscillation de g_n sur I_k est égale à $(b'-a')\cos c$ avec c compris entre a' et b'. En particulier, l'oscillation de g_n est majorée par $|b'-a'| \leqslant 4\pi n!/k!$ puisque a' et b' se trouve tous deux dans le segment $[n!x_0, n!x_0 + 4\pi n!/k!]$.
(b) Pour tout $u \in I_k$ on a l'encadrement :
$$\frac{-2k!}{2\pi n!} \leqslant \frac{-2}{n!u} \leqslant g_n(u) \leqslant \frac{2}{n!u} \leqslant \frac{2k!}{2\pi n!}\cdot$$
Par conséquent, l'oscillation de g_n sur I_k est majorée par $2k!/(\pi n!)$.
(c) Quand u parcourt le segment I_k, $n!x_0 + k!u$ parcourt un segment de longueur 2π et $\cos(n!x_0 + k!u)$ prend toutes les valeurs entre -1 et 1. Il y a donc un point $a \in I_k$ tel que :
$$g_n(a) = \frac{1-\cos(k!x_0)}{k!a} \geqslant \frac{1-\cos(k!x_0)}{4\pi}\cdot$$
Il y a également un point $b \in I_k$ tel que :
$$g_n(b) = \frac{-1-\cos(k!x_0)}{k!a} \leqslant \frac{-1-\cos(k!x_0)}{4\pi}\cdot$$
L'oscillation de g_n sur I_k est donc supérieure à :
$$\frac{1}{4\pi}\big(1-\cos(k!x_0)+1+\cos(k!x_0)\big) = \frac{1}{2\pi}\cdot$$
3. (a) On fait des majorations brutales :
$$\sum_{1\leqslant n<k} \frac{n!}{k!} = \frac{1}{k} + \underbrace{\frac{1}{k(k-1)} + \frac{1}{k(k-1)(k-2)} + \cdots + \frac{1}{k!}}_{<\,(k-2)\cdot\frac{1}{k(k-1)}} < \frac{2}{k}$$

et

$$\sum_{n>k} \frac{k!}{n!} = \frac{1}{k+1} + \frac{1}{(k+1)(k+2)} + \frac{1}{(k+1)(k+2)(k+3)} + \cdots$$
$$< \frac{1}{k+1} + \frac{1}{(k+1)^2} + \frac{1}{(k+1)^3} \cdots = \frac{1}{k}.$$

(b) L'oscillation de g sur I_k est donc minorée par :

$$\frac{1}{2\pi} - \frac{3}{k}$$

qui ne tend pas vers 0 quand k tend vers l'infini.

(c) Pour tout $x_0 \in \mathbb{R}$ et tout $\varepsilon > 0$, le taux d'accroissement

$$g(u) := \frac{f(x) - f(x_0)}{x - x_0}$$

a donc une oscillation supérieure $1/(2\pi) - \varepsilon$ dans tout voisinage de x_0. par conséquent, le taux d'accroissement ne peut pas avoir de limite quand x tend vers x_0. La fonction f n'est donc pas dérivable en x_0.

Exercice type corrigé

II.4.0 Soit f une fonction $\mathcal{C}^\infty$, périodique de période 1. On s'intéresse à la vitesse de convergence des sommes de Riemann régulières à gauche vers l'intégrale comme le pas de la subdivision tend vers 0. Montrer que pour tout $\alpha \geqslant 0$, on a

$$\frac{1}{m} \sum_{k=0}^{m-1} f\left(\frac{k}{m}\right) = \int_0^1 f + \mathrm{o}\left(\frac{1}{m^\alpha}\right).$$

Solution. Quand α croît, $1/m^\alpha$ décroît. Par conséquent, il suffit de montrer le résultat pour les grandes valeurs de α.
Fixons $\alpha \geqslant 2$. Comme f est une fonction de classe $\mathcal{C}^\infty$, elle est somme de sa série de Fourier :

$$f(x) = \lim_{N\to+\infty} S_N(f) \quad \text{avec} \quad S_N(f) := \sum_{n=-N}^{N} c_n(f) \cdot \mathrm{e}_n.$$

De plus, les coefficients $c_n(f)$ sont à décroissance rapide. En particulier :

$$c_n(f) = \mathrm{o}\left(\frac{1}{n^\alpha}\right).$$

Observons que :

$$\frac{1}{m} \sum_{k=0}^{m-1} f\left(\frac{k}{m}\right) = \frac{1}{m} \sum_{k=0}^{m-1} \lim_{N\to+\infty} \sum_{n=-N}^{N} c_n(f) \cdot \mathrm{e}_n\left(\frac{k}{m}\right)$$
$$= \lim_{N\to+\infty} \sum_{n=-N}^{N} c_n(f) \cdot \left(\frac{1}{m} \sum_{k=0}^{m-1} \mathrm{e}^{i2\pi n \frac{k}{m}}\right).$$

Observons également que :

$$\frac{1}{m}\sum_{k=0}^{m-1} \mathrm{e}^{i2\pi n\frac{k}{m}} = \begin{cases} 1 & \text{si } nk/m \in \mathbb{Z} \\ 0 & \text{sinon.} \end{cases}$$

Par conséquent :

$$\frac{1}{m}\sum_{k=0}^{m-1} f\left(\frac{k}{m}\right) = \sum_{j\in\mathbb{Z}} c_{mj}(f).$$

Finalement, $c_0(f) = \displaystyle\int_0^1 f$.

De plus, pour tout ε, si m est assez grand et si $j \neq 0$, $\left|c_{mj}(f)\right| \leqslant \varepsilon/(mj)^\alpha$ et donc :

$$\sum_{\substack{j\in\mathbb{Z}\\ j\neq 0}} c_{mj}(f) \leqslant \sum_{\substack{j\in\mathbb{Z}\\ j\neq 0}} \frac{\varepsilon}{(mj)^\alpha} \leqslant \frac{\varepsilon}{m^\alpha} = \sum_{\substack{j\in\mathbb{Z}\\ j\neq 0}} \frac{1}{j^2} = \frac{\varepsilon\pi^2/3}{m^\alpha}\cdot$$

Autrement dit, $\displaystyle\sum_{\substack{j\in\mathbb{Z}\\ j\neq 0}} c_{mj}(f) = \mathrm{o}(1/m^\alpha)$. Donc :

$$\frac{1}{m}\sum_{k=0}^{m-1} f\left(\frac{k}{m}\right) = \int_0^1 f + \mathrm{o}\left(\frac{1}{m^\alpha}\right)\cdot$$

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

Exercices sur la section 1

II.4.1 Déterminer les coefficients $c_n(f)$, $a_n(f)$ et $b_n(f)$ pour les fonctions suivantes :

1. $f \in \mathcal{C}_{2\pi}$ définie par $f(t) = \cosh(t)$ pour $t \in [-\pi, \pi]$,
2. $f \in \mathcal{C}_{2\pi}$ définie par $f(t) = t^2$ pour $t \in [-\pi, \pi]$,
3. $f \in \mathcal{C}_2$ impaire, définie par $f(t) = t(1-t)$ pour $t \in [0, 1]$,
4. $f(t) = \left|\sin(\pi t)\right|$,
5. $f := \max(\cos, 0)$.

II.4.2 Supposons que $f \in \mathcal{C}_T$.

1. Montrer que si $f(x + T/2) = -f(x)$ pour tout $x \in \mathbb{R}$, alors $a_{2p}(f) = 0$ pour tout $p \geqslant 0$ et $b_{2p}(f) = 0$ pour tout $p \geqslant 1$.
2. Montrer que si $f(T/2 - x) = f(x)$ pour tout $x \in \mathbb{R}$, alors $a_{2p+1}(f) = b_{2p}(f) = 0$ pour tout $p \geqslant 0$.

II.4.3 Développer $x \mapsto \sin^3(x)$ en série de Fourier.

II.4.4 On considère la série $f(x) = \sum_{n \geqslant 1} \frac{\mathrm{e}^{inx}}{n^3}$ avec $x \in \mathbb{R}$.

1. Étudier la convergence de la série.
2. Montrer que f est dérivable. Est-elle de classe $\mathcal{C}^1$?
3. Déterminer $c_n(f)$.

II.4.5 Soit $r \in]-1, 1[$ et $f : t \mapsto \dfrac{1}{1 - 2r\cos t + r^2}$.

1. Montrer que la série de Fourier de f converge normalement.
2. Montrer que $b_n(f) = 0$ pour tout $n \geqslant 1$ et que pour $n > 0$, les coefficients $a_n := a_n(f)$ vérifient la relation de récurrence :

$$ra_{n+1} - (1 + r^2)a_n + ra_{n-1} = 0.$$

   En déduire qu'il existe α et β tels que pour tout n on ait $a_n = \alpha r^n + \dfrac{\beta}{r^n}$.
3. Montrer que $\beta = 0$, calculer a_0 et en déduire une expression de a_n.
4. En déduire que pour tout $t \in \mathbb{R}$ on a :

$$\frac{1 - r^2}{1 - 2r\cos t + r^2} = 1 + 2\sum_{n=1}^{+\infty} r^n \cos nt.$$

5. Retrouver ce résultat en calculant directement la somme de la série $\sum_{n=1}^{+\infty} r^n \cos nt$.

Exercices sur la section 2

II.4.6 Supposons que la suite (u_n) converge vers un réel ℓ et posons :

$$v_n = \frac{u_1 + \ldots + u_n}{n}.$$

1. Soit $n > 0$, et p un entier tel que $1 \leqslant p \leqslant n$. Montrer que :

$$|v_n - \ell| \leqslant \frac{1}{n}\sum_{k=1}^{p} |u_k - \ell| + \max_{p < k \leqslant n} |u_k - \ell|.$$

2. Montrer que (v_n) converge vers ℓ.

II.4.7 Vérifier que pour tout $t \in [-\pi, \pi]$, $t^2 = \dfrac{\pi^3}{3} + 4\sum_{n=1}^{+\infty}(-1)^n \dfrac{\cos(nt)}{n^2}$.

En déduire : $\sum_{n=1}^{+\infty} \dfrac{1}{n^4}$ et $\sum_{n=0}^{+\infty} \dfrac{1}{(2n+1)^4}$.

II.4.8 Supposons que $f, g \in \mathcal{C}_T$ et posons $h := f \cdot g$.

1. Montrer que $c_n(h) = (\,\overline{g} \cdot \mathrm{e}_n \mid f\,)$.
2. En déduire que $c_n(h) = \sum_{k=-\infty}^{+\infty} c_k(f) c_{n-k}(g)$.

Exercices sur la section 3

II.4.9 Soit $f \in \mathcal{R}_1$ la fonction définie par $f(x) = x$ si $x \in [0, 1[$.

1. Déterminer les coefficients de Fourier $c_n(f)$.
2. Montrer que pour tout $x \in]0, 1[$, $f(x) = \dfrac{1}{2} - \dfrac{1}{\pi}\sum_{n=1}^{+\infty} \dfrac{\sin 2\pi n x}{n}$.
3. En déduire : $\dfrac{\pi}{4} = 1 - \dfrac{1}{3} + \dfrac{1}{5} - \dfrac{1}{7} + \cdots$
4. En utilisant la formule de Parseval, montrer que $\sum_{n=1}^{+\infty} \dfrac{1}{n^2} = \dfrac{\pi^2}{6}$.

II.4.10 Soit $a \in \mathbb{R} \setminus \mathbb{Z}$. Soit $f \in \mathcal{R}_{2\pi}$ la fonction définie par $f(t) = \mathrm{e}^{iat}$ sur $]-\pi, \pi]$.

1. Montrer que $c_n(f) = \dfrac{(-1)^n \sin(a\pi)}{\pi(a-n)}$.
2. En déduire que $\sum_{n\in\mathbb{Z}} \dfrac{1}{(a-n)^2} = \dfrac{\pi^2}{\sin^2(\pi a)}$.

II.4.11 Soit $a > 0$.

1. Montrer qu'il existe une fonction $f : [-\pi, \pi] \to \mathbb{R}$ continue telle que pour tout $t \in [-\pi, \pi]$, $f(t) = \sum_{n=1}^{+\infty}(-1)^n \dfrac{\cos(nt)}{n^2 + a^2}$.
2. Déterminer le développement en série de Fourier de la fonction 2π périodique qui est égale à e^{at} sur $]-\pi, \pi]$.
3. En déduire que f est de classe $\mathcal{C}^1$ sur $]-\pi, \pi[$ et former une équation différentielle vérifiée par f.
4. En déduire une expression de f sur $[-\pi, \pi]$.
5. En déduire la valeur de $\sum_{n=1}^{+\infty} \dfrac{(-1)^n}{n^2 + a^2}$.

Fonctions de plusieurs variables

II.5

Dans ce module, nous allons généraliser certaines propriétés des fonctions d'une variable réelle au cas des fonctions de plusieurs variables réelles. Nous commencerons par étudier le problème d'inversion : étant donnée une application $f : U \subset \mathbb{R}^n \to \mathbb{R}^n$ définie sur un ouvert U de $\mathbb{R}^n$, est-il possible de trouver une fonction réciproque f^{-1} ? À une variable, la réponse est oui si la fonction f est monotone sur un intervalle. À plusieurs variables, nous verrons que la réponse est oui, localement, si f est de classe $\mathcal{C}^1$ et si la différentielle de f est inversible. Cependant, contrairement au cas des fonctions d'une variable, le fait que f soit localement inversible ne garantit pas que f soit globalement inversible, comme le montre par exemple l'application $z \mapsto z^2 : \mathbb{C} \setminus \{0\} \to \mathbb{C} \setminus \{0\}$.

Dans la première section, nous énoncerons et démontrerons le théorème d'inversion locale. Ce théorème est accompagné de son frère, le théorème des fonctions implicites. C'est ce théorème qui nous servira dans la deuxième section pour définir et étudier les sous-variétés lisses de $\mathbb{R}^n$. Ce sont des ensembles définis par des équations, tels les cercles d'équation $(x - x_0)^2 + (y - y_0)^2 - r^2 = 0$ dans $\mathbb{R}^2$.

La philosophie qui dirigera nos arguments dans ces deux premières sections est que, pour comprendre localement le comportement d'applications de plusieurs variables, il suffit assez souvent de comprendre le comportement de la différentielle. La différentielle d'une application est une application linéaire. On comprend généralement assez bien son comportement grâce au cours d'algèbre linéaire.

Dans la troisième section, nous continuerons d'étudier le comportement d'une application de plusieurs variables en la remplaçant par des objets plus simples. Nous nous concentrerons sur le cas des applications à valeurs réelles. Nous expliquerons en particulier comment étudier les maxima et les minima d'une fonction de plusieurs variables à l'aide du polynôme de Taylor d'ordre 2. Nous ferons alors appel au cours d'algèbre bilinéaire, plus particulièrement aux formes quadratiques (module I.3).

Dans la quatrième et dernière section, nous tenterons de montrer le lien qui existe entre dérivation et intégration d'une fonction de plusieurs variables. Nous montrerons comment le théorème fondamental du calcul différentiel et intégral (théorème 93 de la page 606 et son corollaire) peut se généraliser aux fonctions de deux variables réelles. Nous nous intéresserons en particulier à la formule de Green-Riemann qui relie l'intégrale d'une forme différentielle le long d'une courbe fermée simple à l'intégrale de sa dérivée sur le domaine délimité par la courbe.

La plupart des résultats que nous obtiendrons dans ce module est valable pour des applications $f : U \subset X \to Y$ où X et Y sont des espaces affines de dimension finie et $U \subset X$ est un ouvert de X, par exemple si $X = \mathbb{C}^n$ et/ou $Y = \mathbb{C}^m$.

1 Théorème d'inversion locale Théorème des fonctions implicites

1.1 Rappels

Dans le module IV.2 de [L1], nous avons vu que si $f : I \to \mathbb{R}$ est dérivable sur un intervalle I et si $f' > 0$ sur I, alors :

- d'une part, f est strictement monotone sur I ; elle est donc bijective de I sur $f(I)$ et admet une fonction réciproque $f^{-1} : f(I) \to I$;
- d'autre part, f^{-1} est dérivable sur $f(I)$ et :

$$\forall y \in f(I), \quad \big(f^{-1}\big)'(y) = \frac{1}{f'\big(f^{-1}(y)\big)}\cdot$$

Nous nous proposons de généraliser ce genre de résultat au cas de fonctions vectorielles de plusieurs variables.

Rappelons ce qui est connu lorsque $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ est une application linéaire. Dans le module II.7 de [L1], nous avons étudié les systèmes linéaires. Résoudre un système linéaire de m équations à n inconnues revient à rechercher les solutions d'une équation $\mathbf{f}(\mathbf{x}) = \mathbf{b}$ avec $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ une application linéaire, $\mathbf{x} \in \mathbb{R}^n$ et $\mathbf{b} \in \mathbb{R}^m$. Supposons que le rang du système soit égal à m (ce qui impose la condition $m \leqslant n$).

Si $m = n$, l'application $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ est inversible. Elle admet une application réciproque $\mathbf{f}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ qui est l'application linéaire inverse.

Si $m < n$ et si le système admet une solution $\mathbf{x}_0 \in \mathbb{R}^n$, alors le système est *compatible*. Dans ce cas, on peut trouver m inconnues principales que l'on peut exprimer en fonction des $n - m$ autres inconnues.

Nous allons étudier ce qui persiste lorsque l'application $\mathbf{f}$ n'est plus supposée linéaire, mais seulement différentiable.

Rappelons que dans le cas d'une fonction de plusieurs variables, la notion de dérivabilité est remplacée par la notion de différentiabilité (voir le module IV.7 de [L1]) : une application $\mathbf{f} : U \to \mathbb{R}^m$, définie sur un ouvert $U \subset \mathbb{R}^n$, est différentiable en un point $\mathbf{a} \in U$ s'il existe une application linéaire $\mathbf{L} : \mathbb{R}^n \to \mathbb{R}^m$ telle que :

$$\mathbf{f}(\mathbf{a} + \mathbf{h}) = \mathbf{f}(\mathbf{a}) + \mathbf{L}(\mathbf{h}) + \|\mathbf{h}\|\varepsilon(\mathbf{h}) \quad \text{avec} \quad \varepsilon(\mathbf{h}) \underset{\mathbf{h}\to\mathbf{0}}{\longrightarrow} \mathbf{0}.$$

L'application linéaire $\mathbf{L}$ est alors appelée la différentielle de $\mathbf{f}$ en $\mathbf{a}$. On la note $D_{\mathbf{a}}\mathbf{f}$. De plus, l'application $\mathbf{f}$ admet en $\mathbf{a}$ des dérivées partielles par rapport à toutes les variables et la matrice de $D_{\mathbf{a}}\mathbf{f}$ dans la base canonique est la matrice dont la $j^{ème}$ colonne est le vecteur $\dfrac{\partial \mathbf{f}}{\partial x_j}(\mathbf{a})$. On appelle cette matrice la matrice Jacobienne de $\mathbf{f}$ en $\mathbf{a}$. On la note $J_{\mathbf{a}}\mathbf{f}$.

Si $m = n$ le déterminant de la matrice Jacobienne s'appelle le *Jacobien* de $\mathbf{f}$.

Si $\mathbf{f} : U \subset \mathbb{R}^n \to V \subset \mathbb{R}^m$ et $\mathbf{g} : V \to W \subset \mathbb{R}^p$ sont deux applications différentiables, alors $\mathbf{g} \circ \mathbf{f} : U \to \mathbb{R}^p$ est différentiable et pour tout $\mathbf{a} \in U$, on a :

$$J_{\mathbf{a}}(\mathbf{g} \circ \mathbf{f}) = \big(J_{\mathbf{f}(\mathbf{a})}\mathbf{g}\big) \cdot \big(J_{\mathbf{a}}\mathbf{f}\big).$$

Cette formule traduit la relation plus intrinsèque :

$$D_{\mathbf{a}}(\mathbf{g}\circ\mathbf{f}) = \big(D_{\mathbf{f}(\mathbf{a})}\mathbf{g}\big)\circ D_{\mathbf{a}}\mathbf{f}.$$

Si l'on note $(x_1,\dots x_n)$ les points de U, $(y_1,\dots,y_m)$ les points de V et $(f_1,\dots,f_m)$ les coordonnées de $\mathbf{f}$, cela peut également s'écrire de la façon suivante :

$$\forall j\in[\![1,n]\!],\quad \frac{\partial \mathbf{g}\circ\mathbf{f}}{\partial x_j}(\mathbf{a}) = \sum_{i=1}^{m}\frac{\partial \mathbf{g}}{\partial y_i}\big(\mathbf{f}(\mathbf{a})\big)\cdot\frac{\partial f_i}{\partial x_j}(\mathbf{a}).$$

Définition 1. On dit que $\mathbf{f}$ est une application continûment différentiable sur U (on dit aussi de classe $\mathcal{C}^1$) si toutes les dérivées partielles de $\mathbf{f}$ existent et sont continues sur U. Dans ce cas, $\mathbf{f}$ est différentiable sur U.

L'existence de dérivées partielles en tout point de U ne suffit pas pour garantir la différentiabilité de $\mathbf{f}$ (ni même sa continuité). Par exemple la fonction $(x,y)\mapsto\dfrac{xy}{x^2+y^2}$ prolongée par 0 à l'origine n'y est pas continue mais admet des dérivées partielles en tout point de $\mathbb{R}^2$.

1.2 Énoncés des théorèmes

Le premier résultat concerne le cas où la dimension de l'espace de départ est égale à la dimension de l'espace d'arrivée. Le théorème d'inversion locale affirme que si la différentielle de $\mathbf{f}$ est inversible, alors l'application $\mathbf{f}$ est localement inversible.

Définition 2. Soient V et W deux ouverts de $\mathbb{R}^n$. Une bijection continûment différentiable $\mathbf{f}:V\to W$ dont l'inverse $\mathbf{f}^{-1}:W\to V$ est également continûment différentiable est un *difféomorphisme de classe* $\mathcal{C}^1$.

Exercice 1.

Montrer que l'application $\mathbf{f}:\mathbb{R}^2\to\mathbb{R}^2$ définie par $\mathbf{f}(x,y) := \big(x+y+\sin(x), y+\sin(x)\big)$ est un difféomorphisme de classe $\mathcal{C}^1$.

Solution. L'application $\mathbf{f}:\mathbb{R}^2\to\mathbb{R}^2$ est une bijection dont l'inverse est obtenue en résolvant le système :

$$\begin{cases} X = x+y+\sin(x)\\ Y = y+\sin(x).\end{cases}$$

Les solutions sont :

$$\begin{cases} x = X-Y\\ y = Y-\sin(X-Y).\end{cases}$$

On a donc $\mathbf{f}^{-1}(X,Y) = \big(X-Y, Y-\sin(X-Y)\big)$. Les matrices Jacobiennes de $\mathbf{f}$ et $\mathbf{f}^{-1}$ sont :

$$J_{(x,y)}\mathbf{f} = \begin{pmatrix} 1+\cos(x) & 1\\ \cos(x) & 1\end{pmatrix}\quad\text{et}\quad J_{(X,Y)}\mathbf{f}^{-1} = \begin{pmatrix} 1 & -1\\ -\cos(X-Y) & 1+\cos(X-Y)\end{pmatrix}$$

ce qui montre que $\mathbf{f}$ et $\mathbf{f}^{-1}$ sont de classe $\mathcal{C}^1$ puisque les dérivées partielles sont toutes continues.

Théorème 1 (Inversion locale). Soit U un ouvert de $\mathbb{R}^n$ et $\mathbf{f} : U \to \mathbb{R}^n$ une application de classe $\mathcal{C}^1$ dont la différentielle est inversible en un point $\mathbf{x}_0 \in U$. Alors, il existe un ouvert $V \subset U$ contenant $\mathbf{x}_0$ et un ouvert $W \subset \mathbb{R}^n$ contenant $\mathbf{f}(\mathbf{x}_0)$ tels que $\mathbf{f} : V \to W$ soit un difféomorphisme de classe $\mathcal{C}^1$.

Démonstration en dimension n = 1 En dimension 1, ce théorème découle facilement de ce que nous savons déjà. En effet, si $f : I \to \mathbb{R}$ est une application de classe $\mathcal{C}^1$ sur un intervalle I et si $f'(x_0) \neq 0$ en un point $x_0 \in I$, alors il existe un intervalle $J \subset I$ contenant x_0 sur lequel f' ne s'annule pas (par continuité de f'). L'application $f : J \to f(J)$ est alors un difféomorphisme de classe $\mathcal{C}^1$. ∎

L'hypothèse que f soit de classe $\mathcal{C}^1$ et pas seulement différentiable est importante. Par exemple, l'application $f : \mathbb{R} \to \mathbb{R}$ définie par :

$$f(x) := \begin{cases} 0 & \text{si } x = 0 \\ \dfrac{x}{2} + x^2 \sin \dfrac{1}{x} & \text{si } x \neq 0. \end{cases}$$

est partout dérivable. La dérivée en 0 est égale à $1/2 \neq 0$ mais la fonction f n'est monotone sur aucun voisinage de 0 puisque sa dérivée $x \mapsto \dfrac{1}{2} + 2x \sin \dfrac{1}{x} - \cos \dfrac{1}{x}$ change de signe dans tout voisinage de 0.

La démonstration en dimension $n > 1$ est autrement plus difficile (déjà pour $n = 2$) et nous l'exposerons ultérieurement (section 1.4).

> ***Remarque***. Nous avons vu en dimension 1 qu'il se peut qu'une application dérivable $f : I \to \mathbb{R}$ admette une réciproque $f^{-1} : f(I) \to I$ même si la dérivée de f s'annule en un point $x_0 \in I$ (auquel cas la différentielle n'y est pas inversible). C'est par exemple le cas pour la fonction $x \mapsto x^3$ dont la dérivée s'annule en 0. Cela reste *a fortiori* vrai en dimension supérieure. Le théorème d'inversion locale ne couvre pas ces cas.

Le théorème d'inversion locale est complété par la proposition suivante qui relie les différentielles de $\mathbf{f}$ et de $\mathbf{f}^{-1}$.

Proposition 2. Si $\mathbf{f} : V \to W$ est un difféomorphisme de classe $\mathcal{C}^1$, pour tout point $\mathbf{x} \in V$, la différentielle $D_{\mathbf{x}}\mathbf{f}$ est inversible et :

$$D_{\mathbf{f}(\mathbf{x})}\left(\mathbf{f}^{-1}\right) = \left(D_{\mathbf{x}}\mathbf{f}\right)^{-1}.$$

Démonstration.
Il suffit de différentier l'égalité $\mathbf{f}^{-1} \circ \mathbf{f} = \mathrm{Id}$ en $\mathbf{x} \in V$, ce qui donne (par application de la règle donnant la différentielle d'une composition, page 705) $D_{\mathbf{f}(\mathbf{x})}\left(\mathbf{f}^{-1}\right) \circ D_{\mathbf{x}}\mathbf{f} = \mathrm{Id}$. ∎

Corollaire 3 (Invariance de la dimension). Soient U un ouvert non vide de $\mathbb{R}^m$ et V un ouvert non vide de $\mathbb{R}^n$, avec $m \neq n$. Il n'existe alors aucun difféomorphisme de classe $\mathcal{C}^1$ de U sur V.

Remarque. On peut démontrer qu'il n'existe aucun *homéomorphisme* de U sur V (invariance *topologique* de la dimension), mais c'est nettement plus difficile.

Le second résultat concerne le cas où $m < n$. Si $\mathbf{F} := (F_1, F_2, \ldots, F_m)$ est une application d'un ouvert $U \subset \mathbb{R}^n$ à valeurs dans $\mathbb{R}^m$, résoudre l'équation :

$$\mathbf{F}(x_1, x_2, \ldots, x_n) = (b_1, b_2, \ldots, b_m)$$

revient à résoudre un système (non linéaire) de m équations à n inconnues :

$$\left\{ \begin{array}{l} F_1(x_1, x_2, \ldots, x_n) = b_1 \\ F_2(x_1, x_2, \ldots, x_n) = b_2 \\ \qquad\qquad \vdots \\ F_m(x_1, x_2, \ldots, x_n) = b_m. \end{array} \right.$$

On a moins d'équations que d'inconnues.

Exemple. Considérons l'application $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ définie par :

$$\mathbf{F}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x^2 + y^2 + z^2 \\ x + y \end{pmatrix}.$$

Chercher les solutions de $\mathbf{F}(x, y, z) = (a, b)$ revient à résoudre le système de 2 équations à 3 inconnues :

$$\left\{ \begin{array}{l} x^2 + y^2 + z^2 = a \\ x + y = b. \end{array} \right.$$

Si $a < 0$, il n'y a pas de solution. Si $a > 0$, on est en train de rechercher l'intersection de la sphère centré à l'origine de rayon $\sqrt{a}$ avec le plan d'équation $x + y = b$. Il se peut ici encore que l'ensemble des solutions soit vide. Sinon, c'est soit un point (si $a = 0$ ou si le plan est tangent à la sphère), soit un cercle.

Le théorème des fonctions implicites exprime que si l'équation $\mathbf{F}(\mathbf{x}) = \mathbf{b}$ admet une solution $\mathbf{x}_0 \in U$ et si la différentielle de $\mathbf{F}$ en $\mathbf{x}_0$ est de rang maximal m, alors on peut localement exprimer m variables bien choisies en fonction des $n - m$ autres variables.
Pour faciliter la lecture, on pose $p := n - m$ et l'on note $(\mathbf{x}, \mathbf{y})$ un point de $\mathbb{R}^n = \mathbb{R}^p \times \mathbb{R}^m$ avec $\mathbf{x} := (x_1, \ldots, x_p) \in \mathbb{R}^p$ et $\mathbf{y} := (y_1, \ldots, y_m) \in \mathbb{R}^m$.

Théorème 4 (Fonctions implicites).
Soit U un ouvert de $\mathbb{R}^{p+m}$ et $\mathbf{F} := (F_1, \dots, F_m) : U \to \mathbb{R}^m$ une application de classe $\mathcal{C}^1$. Supposons qu'en un point $(\mathbf{a}, \mathbf{b}) \in U$, la matrice suivante soit inversible :

$$\begin{pmatrix} \dfrac{\partial F_1}{\partial y_1}(\mathbf{a},\mathbf{b}) & \cdots & \dfrac{\partial F_1}{\partial y_m}(\mathbf{a},\mathbf{b}) \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial y_1}(\mathbf{a},\mathbf{b}) & \cdots & \dfrac{\partial F_m}{\partial y_m}(\mathbf{a},\mathbf{b}) \end{pmatrix}.$$

Alors, il existe un voisinage U' de $(\mathbf{a}, \mathbf{b})$ dans $\mathbb{R}^{p+m}$, un voisinage V de $\mathbf{a}$ dans $\mathbb{R}^p$ et une unique application $\mathbf{f} : V \to \mathbb{R}^m$ telle que :

$$(\mathbf{x}, \mathbf{y}) \in U' \text{ et } \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{F}(\mathbf{a}, \mathbf{b}) \qquad \Longleftrightarrow \qquad \mathbf{x} \in V \text{ et } \mathbf{y} = \mathbf{f}(\mathbf{x}).$$

L'application $\mathbf{f}$ est de classe $\mathcal{C}^1$.

La démonstration de ce théorème est assez difficile et nous la réservons pour plus loin (section 1.4). Notons qu'il s'applique (et que nous l'appliquerons) même quand les « variables principales » ne sont pas les p premières de la liste, voir par exemple la proposition 9 de la page 722.
La proposition suivante complète l'énoncé du théorème des fonctions implicites.

Proposition 5. Supposons que $F : U \subset \mathbb{R}^{p+m} \to \mathbb{R}^m$ vérifie les hypothèses du théorème des fonctions implicites en un point $\mathbf{c} := (\mathbf{a}, \mathbf{b}) \in U$ et soit $\mathbf{f} : V \subset \mathbb{R}^p \to \mathbb{R}^m$ une fonction différentiable qui vérifie :

$$\forall \mathbf{x} \in V, \qquad \mathbf{F}\big(\mathbf{x}, \mathbf{f}(\mathbf{x})\big) = \mathbf{F}(\mathbf{a}, \mathbf{b}).$$

Alors :

$$J_{\mathbf{a}}\mathbf{f} = -\begin{pmatrix} \dfrac{\partial F_1}{\partial y_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_1}{\partial y_m}(\mathbf{c}) \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial y_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_m}{\partial y_m}(\mathbf{c}) \end{pmatrix}^{-1} \cdot \begin{pmatrix} \dfrac{\partial F_1}{\partial x_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_1}{\partial x_p}(\mathbf{c}) \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial x_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_m}{\partial x_p}(\mathbf{c}) \end{pmatrix}.$$

Démonstration. Soit $\mathbf{g} : V \to \mathbb{R}^{p+m}$ l'application qui à $\mathbf{x}$ associe $\big(\mathbf{x}, \mathbf{f}(\mathbf{x})\big)$.
C'est une application de classe $\mathcal{C}^1$ dont la matrice Jacobienne en $\mathbf{a}$ est obtenue en superposant les matrices I_p et $J_{\mathbf{a}}\mathbf{f}$: $J_{\mathbf{a}}\mathbf{g} = \begin{pmatrix} I_p \\ J_{\mathbf{a}}\mathbf{f} \end{pmatrix}$.

La matrice Jacobienne de $\mathbf{F}$ en $\mathbf{c} = \mathbf{g}(\mathbf{a})$ est $J_{\mathbf{c}}\mathbf{F} = \begin{pmatrix} A & B \end{pmatrix}$ avec

$$A := \begin{pmatrix} \dfrac{\partial F_1}{\partial x_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_1}{\partial x_p}(\mathbf{c}) \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial x_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_m}{\partial x_p}(\mathbf{c}) \end{pmatrix} \quad \text{et} \quad B := \begin{pmatrix} \dfrac{\partial F_1}{\partial y_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_1}{\partial y_m}(\mathbf{c}) \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial y_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_m}{\partial y_m}(\mathbf{c}) \end{pmatrix}.$$

L'application $\mathbf{F} \circ \mathbf{g}$ est constante, égale à $\mathbf{F}(\mathbf{c})$. Sa différentielle est donc l'application nulle. Par conséquent :

$$\mathbf{0} = J_{\mathbf{g}(\mathbf{a})}\mathbf{F} \cdot J_{\mathbf{a}}\mathbf{g} = \begin{pmatrix} A & B \end{pmatrix} \cdot \begin{pmatrix} I_p \\ J_{\mathbf{a}}\mathbf{f} \end{pmatrix} = A + B \cdot J_{\mathbf{a}}\mathbf{f}.$$

On en déduit que $B \cdot J_{\mathbf{a}}\mathbf{f} = -A$ puis que $J_{\mathbf{a}}\mathbf{f} = -B^{-1} \cdot A$. ∎

1.3 Exemples

Avant d'exposer les démonstrations du théorème d'inversion locale et du théorème des fonctions implicites, nous allons en présenter des cas particuliers et des conséquences.

Nous commencerons par illustrer ce que dit le théorème d'inversion locale dans le cas d'une fonction $f : U \to \mathbb{C}$, holomorphe sur un ouvert $U \subset \mathbb{C}$. Nous reviendrons sur ce sujet dans le module II.6.

Proposition 6. Soit f une fonction holomorphe dans un ouvert U de $\mathbb{C}$ telle que f' soit continue. Soit z_0 un point de U tel que $f'(z_0) \neq 0$. Alors, il existe un ouvert $V \subset U$ contenant z_0 et un ouvert $W \subset \mathbb{C}$ contenant $f(z_0)$ tels que $f : V \to W$ soit une bijection holomorphe : la réciproque $f^{-1} : W \to V$ est holomorphe et sa dérivée est continue. De plus, pour tout $z \in W$, on a :

$$\big(f^{-1}\big)'(z) = \frac{1}{f'\big(f^{-1}(z)\big)}.$$

Remarque. Cela est cohérent avec les résultats concernant les fonctions dérivables d'une variable réelle.

Démonstration. On identifie la fonction $f : x + iy \mapsto f(x + iy)$ avec la fonction :

$$\mathbf{f} : (x, y) \mapsto \big(P(x, y), Q(x, y)\big)$$

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 fonction f est $\mathbb{C}$-dérivable en $z_0 := x_0 + iy_0$ si, et seulement si, $\mathbf{f}$ est différentiable en (x_0, y_0) et :

$$J_{(x_0,y_0)}\mathbf{f} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \quad \text{avec} \quad (a, b) \in \mathbb{R}^2.$$

On a alors $f'(z_0) = a + ib$.

Par conséquent, lorsque f est holomorphe, la fonction $\mathbf{f}$ est de classe $\mathcal{C}^1$ si, et seulement si, la dérivée $f'(z)$ est une fonction continue. La différentielle de $\mathbf{f}$ en (x_0, y_0) est inversible si, et seulement si, le déterminant $a^2 + b^2 = \big|f'(z_0)\big|^2$ de la matrice Jacobienne est non nul, donc si, et seulement si, $f'(z_0) \neq 0$.

Le théorème d'inversion locale implique donc que si $f : U \to \mathbb{C}$ est une fonction holomorphe dans un ouvert $U \subset \mathbb{C}$, si f' est une fonction continue sur U et si $z_0 \in U$ est un

point tel que $f'(z_0) \neq 0$, alors il existe un voisinage V de z_0 et un voisinage W de $f(z_0)$ de sorte que $f : U \to V$ soit un difféomorphisme de classe $\mathcal{C}^1$. Plus précisément, l'application $\mathbf{f}$ admet localement une application réciproque $\mathbf{g} := (g_1, g_2) := \mathbf{f}^{-1}$ définie dans un voisinage de $\big(\operatorname{Re} f(z_0), \operatorname{Im} f(z_0)\big)$. Si l'on pose $g(x+iy) = g_1(x,y) + ig_2(x,y)$, on définit au voisinage de $f(z_0)$ un inverse local g de f.

De plus, le théorème d'inversion locale dit que la fonction réciproque $\mathbf{g}$ est continûment différentiable et que :

$$J_{\mathbf{f}(x,y)}\mathbf{g} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}^{-1} = \frac{1}{a^2+b^2} \begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

On voit donc que l'application réciproque g est holomorphe et que :

$$g'\big(f(z)\big) = \frac{a-ib}{a^2+b^2} = \frac{\overline{f'(z)}}{\big|f'(z)\big|^2} = \frac{1}{f'(z)}.$$

En ce qui concerne le théorème des fonctions implicites, nous allons maintenant regarder le cas particulier $m = 1$ et $n = 2$. Dans ce cas, le théorème des fonctions implicites se formule de la manière suivante.

Proposition 7. Soit U un ouvert de $\mathbb{R}^2$ et $F : U \to \mathbb{R}$ une application de classe $\mathcal{C}^1$. Supposons que $(a,b) \in U$ et $\dfrac{\partial F}{\partial y}(a,b) \neq 0$. Alors, il existe un ouvert $U' \subset U$ contenant (a,b), un intervalle ouvert I contenant a, et une unique application $f : I \to R$ telle que :

$$(x,y) \in U' \text{ et } F(x,y) = F(a,b) \qquad \Longleftrightarrow \qquad x \in I \text{ et } y = f(x).$$

L'application f est de classe $\mathcal{C}^1$, et pour tout $x \in I$, on a $f'(x) = -\dfrac{\dfrac{\partial F}{\partial x}\big(x, f(x)\big)}{\dfrac{\partial F}{\partial y}\big(x, f(x)\big)}$.

Autrement dit, au voisinage de (a,b), l'ensemble des solutions de l'équation :

$$F(x,y) = F(a,b)$$

est le graphe d'une application $x \mapsto y = f(x)$ qui est de classe $\mathcal{C}^1$. C'est une courbe « lisse » (ce mot sera défini en toute généralité à la section 2).

Exemple. Considérons l'application $F : \mathbb{R}^2 \to \mathbb{R}^2$ définie par $f(x,y) = x^2 + y^2$. Pour tout point $(a,b) \in \mathbb{R}^2$, l'ensemble des points (x,y) tels que $F(x,y) = F(a,b)$ est

- le singleton $\big\{(0,0)\big\}$ si $(a,b) = (0,0)$;
- le cercle centré à l'origine et passant par (a,b) sinon.

Notons que $\dfrac{\partial F}{\partial x}(x,y) = 2x$ et $\dfrac{\partial F}{\partial y}(x,y) = 2y$. Ces deux fonctions sont continues sur $\mathbb{R}^2$. Par conséquent, F est une application continûment différentiable sur $\mathbb{R}^2$.

Si (a,b) n'est pas un point de l'axe des abscisses, $b \neq 0$ et $\dfrac{\partial F}{\partial y}(a,b) = 2b \neq 0$. Le théorème des fonctions implicites nous dit que dans ce cas, le cercle centré à l'origine et passant par (a,b) est localement le graphe d'une fonction $x \mapsto y = f(x)$. Dans le cas présent, on peut même expliciter cette fonction. En effet, il s'agit de résoudre l'équation $x^2+y^2 = a^2+b^2$. Au voisinage de tout point (a,b) avec $b \neq 0$, l'ensemble des solutions est donné par $y = f(x) := \sqrt{a^2+b^2-x^2}$ si $b > 0$ et $y = f(x) := -\sqrt{a^2+b^2-x^2}$ si $b < 0$. Notons que dans les deux cas, on obtient une fonction f qui est de classe $\mathcal{C}^1$ au voisinage de $x = a$ et dont la dérivée est :

$$f'(x) = \begin{cases} \dfrac{-2x}{2\sqrt{a^2+b^2-x^2}} & \text{si } b > 0 \\ -\dfrac{-2x}{2\sqrt{a^2+b^2-x^2}} & \text{si } b < 0. \end{cases}$$

Dans les deux cas, on a bien $f'(x) = -\dfrac{\dfrac{\partial F}{\partial x}(x,y)}{\dfrac{\partial F}{\partial y}(x,y)}$.

Si (a,b) n'est pas un point de l'axe des ordonnées, on peut échanger le rôle des variables. En effet, dans ce cas $a \neq 0$ et $\dfrac{\partial F}{\partial x}(a,b) = 2a \neq 0$. Le théorème des fonctions implicites nous dit alors qu'au voisinage de (a,b), pour résoudre l'équation $x^2+y^2 = a^2+b^2$, il est possible d'exprimer x comme fonction de y. C'est bien entendu le cas puisque l'on peut alors écrire $x = \sqrt{a^2+b^2-y^2}$ si $a > 0$, et $x = -\sqrt{a^2+b^2-y^2}$ si $a < 0$.
Notons enfin qu'au voisinage d'un point $(a,0)$ de l'axe des abscisses, on ne peut pas exprimer l'ensemble des solutions l'équation $x^2+y^2 = a^2$ comme le graphe d'une application $x \mapsto y = f(x)$. En effet, l'équation admet deux solutions distinctes pour $x < a$ et aucune solution pour $x > a$. Le théorème des fonctions implicites n'est pas pris en défaut puisque $\dfrac{\partial F}{\partial y}(a,0) = 0$. De même, au voisinage d'un point $(0,b)$ de l'axe des ordonnées, on ne peut pas exprimer l'ensemble des solutions comme le graphe d'une application $y \mapsto x = f(y)$.

Plutôt que de considérer le cas d'une fonction $F : \mathbb{R}^2 \to \mathbb{R}$, on peut considérer le cas d'une fonction $F : \mathbb{R}^3 \to \mathbb{R}$.

Exercice 2.

Soit $F : \mathbb{R}^3 \to \mathbb{R}$ une fonction des 3 variables (x,y,z), de classe $\mathcal{C}^1$. Donner une condition qui garantit que l'ensemble des solutions est, au voisinage d'un point $(a,b,c) \in \mathbb{R}^3$, le graphe d'une application $(x,y) \mapsto z = f(x,y)$.

Solution. Si $\dfrac{\partial F}{\partial z}(a,b,c) \neq 0$, alors on peut appliquer le théorème des fonctions implicites : au voisinage du point (a,b,c), l'ensemble des solutions est le graphe d'une application $(x,y) \mapsto z = f(x,y)$.

Remarque. Dans l'exercice précédent, au voisinage du point (a, b, c), l'ensemble des solutions est une surface. Ce n'est pas une courbe. Cela n'a rien d'étonnant. Rappelons par exemple qu'une équation cartésienne dans $\mathbb{R}^3$ décrit un plan. Si l'on souhaite caractériser une droite dans $\mathbb{R}^3$, il faut deux équations cartésiennes. De manière similaire, pour caractériser une courbe dans $\mathbb{R}^3$ à l'aide d'une équation de la forme $\mathbf{f}(x, y, z) = 0$, il faut en général que $\mathbf{f}$ soit une application à valeurs dans $\mathbb{R}^2$, ce qui correspond à se donner deux équations. Une courbe est alors caractérisée comme l'intersection de deux surfaces.

Exemple. Considérons la fonction $F : \mathbb{R}^3 \to \mathbb{R}$ définie par $F(x, y, z) = x^2+y^2+z^2$. Pour tout point $(a, b, c) \in \mathbb{R}^3$, l'ensemble des solutions de l'équation $F(x, y, z) = F(a, b, c)$ est :

- le singleton $\big\{(0, 0, 0)\big\}$ si $(a, b, c) = (0, 0, 0)$;
- la sphère centrée à l'origine et passant par (a, b, c) sinon.

Comme dans le cas de l'équation d'un cercle de $\mathbb{R}^2$, on montre facilement que F est une fonction de classe $\mathcal{C}^1$. La dérivée partielle de F par rapport à z est non nulle précisément aux points qui n'appartiennent pas au plan d'équation $z = 0$. Tout point $(a, b, c) \in \mathbb{R}^3$ avec $c > 0$ appartient à une demi-sphère qui est le graphe de l'application :

$$(x, y) \mapsto \sqrt{a^2 + b^2 + c^2 - x^2 - y^2}.$$

Tout point $(a, b, c) \in \mathbb{R}^3$ avec $c < 0$ appartient à une demi-sphère qui est le graphe de l'application :

$$(x, y) \mapsto -\sqrt{a^2 + b^2 + c^2 - x^2 - y^2}.$$

Le théorème des fonctions implicites admet une variante pour les fonctions de plusieurs variables complexes, mais n'ayant défini ni les fonctions analytiques de plusieurs variables complexes ni leurs $\mathbb{C}$-dérivées partielles, nous nous contenterons de l'illustrer sur un exemple utile.

Exercice 3.

Soit $P \in \mathbb{C}[X, Y]$ un polynôme à deux variables complexes. Pour $x \in \mathbb{C}$, on note P_x le polynôme $P_x(y) := P(x, y)$. On suppose que $y_0 \in \mathbb{C}$ est une racine simple du polynôme P_{x_0}, $x_0 \in \mathbb{C}$. Montrer qu'il existe un ouvert U contenant x_0 et une fonction holomorphe $f : U \to \mathbb{C}$ telle que pour tout $x \in U$, $f(x)$ soit une racine du polynôme P_x.

Solution. L'hypothèse de simplicité de la racine équivaut à la non annulation de $P'_{x_0}(y_0)$. Par un calcul analogue à celui de la preuve de la proposition 6, on en déduit que la différentielle de P, vue comme fonction de $\mathbb{R}^4$ dans $\mathbb{R}^2$, est de rang 2 au point (x_0, y_0). Le théorème des fonctions implicites s'applique donc. Un calcul direct montre alors que la différentielle de la fonction implicite $y(x)$ en x_0 est la multiplication par le nombre complexe $-\dfrac{\dfrac{\partial P}{\partial x}(x_0, y_0)}{\dfrac{\partial P}{\partial y}(x_0, y_0)}$, ce qui montre que c'est une fonction holomorphe.

1.3.1 Équivalence des théorèmes

Avant de passer à la démonstration des théorèmes, nous souhaitons mettre en avant que ces deux théorèmes sont équivalents. Les deux théorèmes s'impliquent mutuellement. Il nous suffira donc de ne démontrer que l'un d'entre eux.

Commençons par supposer que le théorème d'inversion locale est vrai et montrons que le théorème des fonctions implicites en découle.

Nous disposons donc d'une application $\mathbf{F} : U \subset \mathbb{R}^{p+m} \to \mathbb{R}^m$ qui est continûment différentiable sur U et d'un point $(\mathbf{a}, \mathbf{b}) \in U$ tel que la matrice :

$$\begin{pmatrix} \dfrac{\partial F_1}{\partial y_1}(\mathbf{a},\mathbf{b}) & \cdots & \dfrac{\partial F_1}{\partial y_m}(\mathbf{a},\mathbf{b}) \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial y_1}(\mathbf{a},\mathbf{b}) & \cdots & \dfrac{\partial F_m}{\partial y_m}(\mathbf{a},\mathbf{b}) \end{pmatrix}$$

soit inversible. Nous allons appliquer le théorème d'inversion locale à l'application :

$$\begin{array}{rccc} \mathbf{g}: & \mathbb{R}^p \times \mathbb{R}^m & \longrightarrow & \mathbb{R}^p \times \mathbb{R}^m \\ & (\mathbf{x},\mathbf{y}) & \longmapsto & \big(\mathbf{x}, \mathbf{F}(\mathbf{x},\mathbf{y})\big). \end{array}$$

La matrice Jacobienne de $\mathbf{g}$ en un point $(\mathbf{x},\mathbf{y})$ est $J_{(\mathbf{x},\mathbf{y})}\mathbf{g} = \begin{pmatrix} I_p & \mathbf{0} \\ B(\mathbf{x},\mathbf{y}) & A(\mathbf{x},\mathbf{y}) \end{pmatrix}$ avec :

$$A := \begin{pmatrix} \dfrac{\partial F_1}{\partial y_1} & \cdots & \dfrac{\partial F_1}{\partial y_m} \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial y_1} & \cdots & \dfrac{\partial F_m}{\partial y_m} \end{pmatrix} \quad \text{et} \quad B := \begin{pmatrix} \dfrac{\partial F_1}{\partial x_1} & \cdots & \dfrac{\partial F_1}{\partial x_p} \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial x_1} & \cdots & \dfrac{\partial F_m}{\partial x_p} \end{pmatrix}.$$

En particulier, toutes les dérivées partielles de $\mathbf{g}$ sont continues. De plus, le déterminant de $J_{(\mathbf{a},\mathbf{b})}\mathbf{g}$ est égal à celui de $A(\mathbf{a},\mathbf{b})$, que l'on a supposée inversible (cela se voit par exemple en développant successivement par rapport aux p premières lignes). D'après le théorème d'inversion locale, il existe un voisinage $U' \subset U$ de $(\mathbf{a},\mathbf{b})$ tel que $\mathbf{g} : U' \to \mathbf{g}(U')$ soit un difféomorphisme de classe $\mathcal{C}^1$. L'application $\mathbf{g}$ admet un inverse local de classe $\mathcal{C}^1$. Quitte à diminuer U', on peut supposer que $\mathbf{g}(U') = V \times W$ avec V un voisinage de $\mathbf{a}$ dans $\mathbb{R}^p$ et W un voisinage de $\mathbf{b}$ dans $\mathbb{R}^m$. Cet inverse est de la forme :

$$\mathbf{g}^{-1} : (\mathbf{x},\mathbf{y}) \mapsto \big(\mathbf{x}, \mathbf{h}(\mathbf{x},\mathbf{y})\big)$$

avec $\mathbf{h} : V \times W \to \mathbb{R}^m$ une fonction de classe $\mathcal{C}^1$. Observons que si $(\mathbf{x},\mathbf{y}) \in U'$:

$$\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{F}(\mathbf{a},\mathbf{b}) \quad \Longleftrightarrow \quad \mathbf{g}(\mathbf{x},\mathbf{y}) = \big(\mathbf{x}, \mathbf{F}(\mathbf{a},\mathbf{b})\big).$$

Par conséquent, l'ensemble des solutions de l'équation $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{F}(\mathbf{a},\mathbf{b})$ dans U' est précisément l'ensemble des points $\mathbf{g}^{-1}\big(\mathbf{x}, \mathbf{F}(\mathbf{a},\mathbf{b})\big)$ avec $\mathbf{x} \in V$, c'est-à-dire les points :

$$\big(\mathbf{x}, \mathbf{f}(\mathbf{x})\big) \quad \text{avec} \quad \begin{array}{rccc} \mathbf{f}: & V & \longrightarrow & \mathbb{R}^m \\ & \mathbf{x} & \longmapsto & \mathbf{h}\big(\mathbf{x}, \mathbf{F}(\mathbf{a},\mathbf{b})\big). \end{array}$$

Nous avons donc montré que le théorème des fonctions implicites découle du théorème d'inversion locale. Montrons maintenant la réciproque.
Nous supposons donc que $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^n$ est une application continûment différentiable dont la dérivée est inversible en un point $\mathbf{x}_0 \in U$. Considérons l'application $\mathbf{F} : U \times \mathbb{R}^n \to \mathbb{R}^n$ définie par :

$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{x}) - \mathbf{y}.$$

Au point $\mathbf{c} := \big(\mathbf{x_0}, \mathbf{f}(\mathbf{x}_0)\big)$, l'application $\mathbf{F}$ prend la valeur $\mathbf{0} \in \mathbb{R}^n$ et la matrice :

$$\begin{pmatrix} \dfrac{\partial F_1}{\partial x_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_1}{\partial x_n}(\mathbf{c}) \\ \vdots & & \vdots \\ \dfrac{\partial F_n}{\partial x_1}(\mathbf{c}) & \cdots & \dfrac{\partial F_n}{\partial x_n}(\mathbf{c}) \end{pmatrix} = J_{\mathbf{c}}\mathbf{f}$$

est inversible. D'après le théorème des fonctions implicites, il existe donc un voisinage U' de $\big(\mathbf{x_0}, \mathbf{f}(\mathbf{x}_0)\big)$ dans $U \times \mathbb{R}^n$, un voisinage W de $\mathbf{f}(\mathbf{x}_0)$ dans $\mathbb{R}^n$ et une unique application $\mathbf{g} : W \to \mathbb{R}^n$ telle que :

$$(\mathbf{x}, \mathbf{y}) \in U' \text{ et } \mathbf{f}(\mathbf{x}) - \mathbf{y} = \mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{y} \in W \text{ et } \mathbf{x} = \mathbf{g}(\mathbf{y}).$$

Autrement dit, pour tout $\mathbf{y} \in W$, l'équation $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ admet une unique solution $\mathbf{x}$ avec $(\mathbf{x}, \mathbf{y}) \in U'$. Posons $V := \mathbf{g}(W)$. Alors, $\mathbf{f} : V \to W$ est un difféomorphisme de classe $\mathcal{C}^1$. En effet, si $\mathbf{x} \in V$, il existe $\mathbf{y} \in W$ avec $\mathbf{x} = \mathbf{g}(\mathbf{y})$, c'est-à-dire $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0} = \mathbf{f}(\mathbf{x}) - \mathbf{y}$. Donc, pour tout $\mathbf{x} \in V$, $\mathbf{y} = \mathbf{f}(\mathbf{x}) \in W$ et $\mathbf{x} = \mathbf{g}(\mathbf{y})$. L'application $\mathbf{f} : V \to W$ est bijective. La fonction réciproque est $\mathbf{g} : W \to V$. Ces deux applications sont de classe $\mathcal{C}^1$.

1.4 Démonstrations des théorèmes

Nous en venons finalement à la démonstration des théorèmes. Il suffit de ne démontrer que l'un des deux. Cependant, il existe (au moins) deux démonstrations indépendantes. Nous souhaitons les présenter toutes les deux.

Première démonstration. Nous allons démontrer le théorème des fonctions implicites par récurrence sur le nombre m d'équations. Pour faciliter la récurrence, nous commençons par modifier le problème. Nous disposons pour commencer d'une application $\mathbf{F} : U \subset \mathbb{R}^{p+m} \to \mathbb{R}^m$ telle que la matrice :

$$B := \begin{pmatrix} \dfrac{\partial F_1}{\partial y_1}(\mathbf{a}, \mathbf{b}) & \cdots & \dfrac{\partial F_1}{\partial y_m}(\mathbf{a}, \mathbf{b}) \\ \vdots & & \vdots \\ \dfrac{\partial F_m}{\partial y_1}(\mathbf{a}, \mathbf{b}) & \cdots & \dfrac{\partial F_m}{\partial y_m}(\mathbf{a}, \mathbf{b}) \end{pmatrix}$$

soit inversible en un point $(\mathbf{a}, \mathbf{b}) \in U$. Soit $\mathbf{L} : \mathbb{R}^m \to \mathbb{R}^m$ l'application linéaire dont la matrice dans la base canonique est B et soit :

$$\mathbf{G} := \mathbf{L}^{-1} \circ \mathbf{F} : U \to \mathbb{R}^m.$$

Les solutions de $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{F}(\mathbf{a},\mathbf{b})$ sont exactement les solutions de $\mathbf{G}(\mathbf{x},\mathbf{y}) = \mathbf{G}(\mathbf{a},\mathbf{b})$. Ce que nous avons gagné, c'est que :

$$\begin{pmatrix} \dfrac{\partial G_1}{\partial y_1}(\mathbf{a},\mathbf{b}) & \cdots & \dfrac{\partial G_1}{\partial y_m}(\mathbf{a},\mathbf{b}) \\ \vdots & & \vdots \\ \dfrac{\partial G_m}{\partial y_1}(\mathbf{a},\mathbf{b}) & \cdots & \dfrac{\partial G_m}{\partial y_m}(\mathbf{a},\mathbf{b}) \end{pmatrix} = I_m.$$

Nous pouvons donc supposer à partir de maintenant que la matrice B est la matrice I_m.

Cas d'une seule équation. L'application $F : U \subset \mathbb{R}^{p+1} \to \mathbb{R}$ est continûment différentiable et $\dfrac{\partial F}{\partial y}(\mathbf{a},b) = 1$. Comme F est de classe $\mathcal{C}^1$, il existe $\delta > 0$ tel que :

$$\|\mathbf{x}-\mathbf{a}\| < \delta \text{ et } |y-b| < \delta \quad \Longrightarrow \quad \frac{\partial F}{\partial y}(\mathbf{x},y) > 0.$$

Alors, pour tout $\varepsilon \in]0,\delta[$:

$$F(\mathbf{a},b-\varepsilon) < F(\mathbf{a},b) < F(\mathbf{a},b+\varepsilon).$$

Comme F est continue, pour chaque $\varepsilon \in]0,\delta[$, il existe $\eta \in]0,\delta[$ tel que :

$$\|\mathbf{x}-\mathbf{a}\| < \eta \quad \Longrightarrow \quad F(\mathbf{x},b-\varepsilon) < F(\mathbf{a},b) < F(\mathbf{x},b+\varepsilon).$$

On note V l'ensemble des points $\mathbf{x} \in \mathbb{R}^p$ tels que $\|\mathbf{x}-\mathbf{a}\| < \eta$. Par choix de δ, quand $\mathbf{x} \in V$, l'application $y \mapsto F(\mathbf{x},y) - F(\mathbf{a},b)$ est strictement croissante sur $[b-\varepsilon, b+\varepsilon]$. En $b-\varepsilon$, elle est strictement négative et en $b+\varepsilon$, elle est strictement positive. Pour tout $\mathbf{x} \in V$, il existe donc un unique point $y \in]b-\varepsilon, b+\varepsilon[$ tel que $F(\mathbf{x},y) - F(\mathbf{a},b) = 0$. On note $f(\mathbf{x})$ ce point. On obtient ainsi une fonction $f : V \to]b-\varepsilon, b+\varepsilon[$ telle que :

$$(\mathbf{x},y) \in V \times]b-\varepsilon,b+\varepsilon[\text{ et } F(\mathbf{x},y) = F(\mathbf{a},b) \quad \Longleftrightarrow \quad \mathbf{x} \in V \text{ et } y = f(\mathbf{x}).$$

Comme $\varepsilon > 0$ peut être choisi arbitrairement petit, on voit que f est continue en $\mathbf{a}$. En remplaçant le point $(\mathbf{a},b)$ par n'importe quel point $\big(\mathbf{x}, f(\mathbf{x})\big)$ avec $\mathbf{x} \in V$, et en reprenant l'argument que nous venons d'utiliser, on voit que f est continue sur V.

Soit maintenant $\mathbf{x}_0$ un point de V et $y_0 := f(\mathbf{x}_0)$. Si $\mathbf{e}_j$ est le $j^{ème}$ vecteur de la base canonique de $\mathbb{R}^p$, lorsque $t \in \mathbb{R}$ tend vers 0, $\Delta_j(t) := f(\mathbf{x}_0 + t\mathbf{e}_j) - y_0$ tend vers 0. Observons que :

$$F\big(\mathbf{x}_0 + t\mathbf{e}_j, y_0 + \Delta_j(t)\big) = F\big(\mathbf{x}_0 + t\mathbf{e}_j, f(\mathbf{x}_0 + t\mathbf{e}_j)\big) = F(\mathbf{x}_0,y_0) \quad \big(= F(\mathbf{a},b)\big).$$

En utilisant la différentiabilité de F en $(\mathbf{x}_0,y_0)$, on obtient :

$$F\big(\mathbf{x}_0+t\mathbf{e}_j, y_0+\Delta_j(t)\big) = F(\mathbf{x}_0,y_0)+\frac{\partial F}{\partial x_j}(\mathbf{x}_0,y_0)\cdot t+\frac{\partial F}{\partial y}(\mathbf{x}_0,y_0)\cdot\Delta_j(t)+\mathrm{o}(t)+\mathrm{o}\big(\Delta_j(t)\big).$$

On a donc :

$$\frac{\partial F}{\partial y}(\mathbf{x}_0,y_0)\cdot\Delta_j(t) + \mathrm{o}\big(\Delta_j(t)\big) = \frac{\partial F}{\partial x_j}(\mathbf{x}_0,y_0)\cdot t + \mathrm{o}(t).$$

Comme $\dfrac{\partial F}{\partial y}(\mathbf{x}_0, y_0) \neq 0$, cela montre que $\Delta_j(t) = \mathrm{O}(t)$ et donc :

$$\Delta_j(t) = -\frac{\dfrac{\partial F}{\partial x_j}(\mathbf{x}_0, y_0)}{\dfrac{\partial F}{\partial y}(\mathbf{x}_0, y_0)} \cdot t + \mathrm{o}(t).$$

On voit donc que pour tout $j \in [\![1, p]\!]$, la dérivée partielle $\dfrac{\partial f}{\partial x_j}(\mathbf{x}_0)$ existe et :

$$\frac{\partial f}{\partial x_j}(\mathbf{x}_0) = \lim_{t\to 0} \frac{\Delta_j(t)}{t} = -\frac{\dfrac{\partial F}{\partial x_j}\big(\mathbf{x}_0, f(\mathbf{x}_0)\big)}{\dfrac{\partial F}{\partial y}\big(\mathbf{x}_0, f(\mathbf{x}_0)\big)}\cdot$$

Comme f et les dérivées partielles de F sont des fonctions continues, toutes les dérivées partielles de f existent et sont continues. Cela montre que f est une fonction de classe $\mathcal{C}^1$ sur V.

Cas de $m+1$ équations. Nous allons maintenant montrer que le cas de $m+1$ équations se ramène au cas de m équations. Supposons donc que $\mathbf{F}$ soit une application continûment différentiable sur un ouvert U de $\mathbb{R}^{p+m+1}$ et notons $(\mathbf{x}, \mathbf{y}, z)$ la variable. Supposons de plus que :

$$\begin{pmatrix} \dfrac{\partial F_1}{\partial y_1}(\mathbf{a}, \mathbf{b}, c) & \cdots & \dfrac{\partial F_1}{\partial y_m}(\mathbf{a}, \mathbf{b}, c) & \dfrac{\partial F_1}{\partial z}(\mathbf{a}, \mathbf{b}, c) \\ \vdots & & \vdots & \vdots \\ \dfrac{\partial F_{m+1}}{\partial y_1}(\mathbf{a}, \mathbf{b}, c) & \cdots & \dfrac{\partial F_{m+1}}{\partial y_m}(\mathbf{a}, \mathbf{b}, c) & \dfrac{\partial F_{m+1}}{\partial z}(\mathbf{a}, \mathbf{b}, c) \end{pmatrix} = I_{m+1}.$$

Comme $\dfrac{\partial F_{m+1}}{\partial z}(\mathbf{a}, \mathbf{b}, c) = 1$, il existe un voisinage U' de $(\mathbf{a}, \mathbf{b})$ et une application continûment différentiable $\varphi : U' \to \mathbb{R}$ tels qu'au voisinage de $(\mathbf{a}, \mathbf{b}, c)$, les solutions de $F_{m+1}(\mathbf{x}, \mathbf{y}, z) = F_{m+1}(\mathbf{a}, \mathbf{b}, c)$ soient les points de la forme $\big(\mathbf{x}, \mathbf{y}, \varphi(\mathbf{x}, \mathbf{y})\big)$. De plus, pour tout $i \in [\![1, m]\!]$, on a :

$$\frac{\partial \varphi}{\partial y_j}(\mathbf{a}, \mathbf{b}) = -\frac{\dfrac{\partial F_{m+1}}{\partial y_j}}{\dfrac{\partial F_{m+1}}{\partial z}} = 0.$$

On voit alors qu'au voisinage de $(\mathbf{a}, \mathbf{b}, c)$, les solutions de $\mathbf{F}(\mathbf{x}, \mathbf{y}, z) = \mathbf{F}(\mathbf{a}, \mathbf{b}, c)$ sont les solutions de $\mathbf{G}(\mathbf{x}, \mathbf{y}) = \mathbf{G}(\mathbf{a}, \mathbf{b})$ avec :

$$\mathbf{G}(\mathbf{x}, \mathbf{y}) := \begin{pmatrix} F_1\big(\mathbf{x}, \mathbf{y}, \varphi(\mathbf{x}, \mathbf{y})\big) \\ \vdots \\ F_m\big(\mathbf{x}, \mathbf{y}, \varphi(\mathbf{x}, \mathbf{y})\big) \end{pmatrix}.$$

De plus pour tout $i \in [\![1, m]\!]$ et tout $j \in [\![1, m]\!]$, on a :

$$\frac{\partial G_i}{\partial y_j}(\mathbf{a}, \mathbf{b}) = \frac{\partial F_i}{\partial y_j}(\mathbf{a}, \mathbf{b}, c) + \frac{\partial F_i}{\partial z}(\mathbf{a}, \mathbf{b}, c) \cdot \frac{\partial \varphi}{\partial y_j}(\mathbf{a}, \mathbf{b}) = \frac{\partial F_i}{\partial y_j}(\mathbf{a}, \mathbf{b}, c).$$

La matrice des dérivées partielles de $\mathbf{G}$ par rapport aux y_j au point $(\mathbf{a},\mathbf{b})$ est donc la matrice I_m. Par récurrence, on voit donc qu'il existe un voisinage V de $\mathbf{a}$ et une application continûment différentiable $\mathbf{g} : V \to \mathbb{R}^m$ tels qu'au voisinage de $(\mathbf{a},\mathbf{b})$, les solutions de $\mathbf{G}(\mathbf{x},\mathbf{y}) = \mathbf{G}(\mathbf{a},\mathbf{b})$ soient les points de la forme $\big(\mathbf{x},\mathbf{g}(\mathbf{x})\big)$. En remontant les calculs, on voit qu'au voisinage de $(\mathbf{a},\mathbf{b},c)$ les solutions de $\mathbf{F}(\mathbf{x},\mathbf{y},z) = \mathbf{F}(\mathbf{a},\mathbf{b},c)$ sont les points de la forme $\big(\mathbf{x},\mathbf{f}(\mathbf{x})\big)$ avec :

$$\mathbf{f}(\mathbf{x}) := \big(\mathbf{g}(\mathbf{x}), \varphi\big(\mathbf{x},\mathbf{g}(\mathbf{x})\big)\big).$$

Seconde démonstration. Nous allons démontrer le théorème d'inversion locale grâce à un théorème de point fixe. Cette démonstration présente l'intérêt de se généraliser à des espaces de Banach de dimension infinie.

Nous disposons d'une application $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^n$ continûment différentiable sur U dont la différentielle en $\mathbf{x}_0 \in U$ est inversible. Notons $\mathbf{L} : \mathbb{R}^n \to \mathbb{R}^n$ cette différentielle et posons $\mathbf{y}_0 := \mathbf{f}(\mathbf{x}_0)$. Nous souhaitons dans un premier temps montrer que tout point $\mathbf{y}$ suffisamment proche de $\mathbf{y}_0$ a un antécédent (et un seul) proche de $\mathbf{x}_0$. Pour cela, considérons l'application $\mathbf{G}_\mathbf{y}$ qui à un point $\mathbf{x} \in U$ associe l'unique antécédent de $\mathbf{y}$ par l'application affine $\mathbf{x}' \mapsto \mathbf{f}(\mathbf{x}) + \mathbf{L}(\mathbf{x}' - \mathbf{x})$ (qui est une approximation grossière de l'application $\mathbf{f}$ au voisinage du point $\mathbf{x}$). L'application $\mathbf{G}_\mathbf{y} : U \to \mathbb{R}^n$ a pour expression :

$$\mathbf{G}_\mathbf{y}(\mathbf{x}) = \mathbf{x} + \mathbf{L}^{-1}\big(\mathbf{y} - \mathbf{f}(\mathbf{x})\big).$$

Observons que les points fixes de l'application $\mathbf{G}_\mathbf{y}$ sont les points $\mathbf{x} \in U$ tels que $\mathbf{f}(\mathbf{x}) = \mathbf{y}$. Nous sommes donc à la recherche des points fixes de $\mathbf{G}_\mathbf{y}$.

L'application $\mathbf{G}_\mathbf{y}$ est continûment différentiable sur U et sa différentielle en $\mathbf{x}$ est :

$$D_\mathbf{x}\mathbf{G}_\mathbf{y} = \mathrm{Id} - \mathbf{L}^{-1} \circ D_\mathbf{x}\mathbf{f}.$$

Cette différentielle ne dépend pas de $\mathbf{y}$. Elle dépend continûment de $\mathbf{x}$ et elle s'annule en $\mathbf{x}_0$. On peut donc trouver $\varepsilon > 0$ tel que :

$$\forall \mathbf{y} \in \mathbb{R}^n,\ \forall \mathbf{x} \in \overline{B}(\mathbf{x}_0,\varepsilon), \quad \big\|D_\mathbf{x}\mathbf{G}_\mathbf{y}\big\| \leqslant \frac{1}{2}.$$

D'après l'inégalité des accroissements finis (théorème 30 du module IV.7 de [L1]) :

$$\mathbf{G}_\mathbf{y}\big(\overline{B}(\mathbf{x}_0,\varepsilon)\big) \subset \overline{B}\big(\mathbf{G}_\mathbf{y}(\mathbf{x}_0),\varepsilon/2\big).$$

Définissons maintenant :

$$W := \big\{\mathbf{y} \in \mathbb{R}^n \;\big|\; \mathbf{y} - \mathbf{y}_0 \in \mathbf{L}\big(B(\mathbf{0},\varepsilon/2)\big)\big\} \quad \text{et} \quad V := \big\{\mathbf{x} \in B(\mathbf{x}_0,\varepsilon) \;\big|\; \mathbf{f}(\mathbf{x}) \in W\big\}.$$

Comme $\mathbf{L}$ et $\mathbf{f}$ sont des applications continues, V et W sont des ouverts de $\mathbb{R}^n$. Ce sont des voisinages respectif de $\mathbf{x}_0$ et $\mathbf{y}_0$. Si $\mathbf{y} \in W$, alors :

$$\mathbf{G}_\mathbf{y}(\mathbf{x}_0) = \mathbf{x}_0 + \mathbf{L}^{-1}(\mathbf{y} - \mathbf{y}_0) \in \overline{B}(\mathbf{x}_0,\varepsilon/2)$$

et

$$B\big(\mathbf{G}_\mathbf{y}(\mathbf{x}_0),\varepsilon/2\big) \subset B(\mathbf{x}_0,\varepsilon) \subset \overline{B}(\mathbf{x}_0,\varepsilon).$$

L'application $\mathbf{G}_\mathbf{y} : \overline{B}(\mathbf{x}_0,\varepsilon) \to B(\mathbf{x}_0,\varepsilon)$ est contractante. Elle admet donc un unique point fixe dans $\overline{B}(\mathbf{x}_0,\varepsilon)$ (théorème 39 de la page 427). Ce point fixe se trouve nécessairement dans $B(\mathbf{x}_0,\varepsilon)$. Cela montre que pour tout $\mathbf{y} \in W$, il existe un unique

point $\mathbf{g}(\mathbf{y}) \in B(\mathbf{x}_0, \varepsilon)$ tel que $\mathbf{f}\big(\mathbf{g}(\mathbf{y})\big) = \mathbf{y}$. Cela montre que l'application $\mathbf{f} : V \to W$ est une bijection et que la réciproque est $\mathbf{g} : W \to V$.

De plus, comme $\varepsilon > 0$ peut être choisi arbitrairement petit, on voit que $\mathbf{g}$ est continue en $\mathbf{x}_0$. Observons également qu'en tout point $\mathbf{x} \in V$, la différentielle $D_{\mathbf{x}}\mathbf{f}$ est inversible (si $D_{\mathbf{x}}\mathbf{f}(\mathbf{h}) = \mathbf{0}$, alors $D_{\mathbf{x}}\mathbf{G}_{\mathbf{y}}(\mathbf{h}) = \mathbf{h}$, ce qui contredit $\big\|D_{\mathbf{x}}\mathbf{G}_{\mathbf{y}}\big\| \leqslant 1/2$). En remplaçant $\mathbf{x}_0$ par n'importe quel point $\mathbf{x} \in V$, on voit donc que $\mathbf{g}$ est continue en $\mathbf{f}(\mathbf{x})$ pour tout $\mathbf{x} \in V$. Autrement dit, $\mathbf{g}$ est continue sur W.

Si $\mathbf{e}_j$ est le $j^{ème}$ vecteur de la base canonique de $\mathbb{R}^n$, lorsque $t \in \mathbb{R}$ tend vers 0, $\Delta_j(t) := f(\mathbf{x}_0 + t\mathbf{e}_j) - \mathbf{y}_0$ tend vers 0. En utilisant la différentiabilité de $\mathbf{f}$ en $\mathbf{x}_0$, on obtient :

$$\mathbf{f}\big(\mathbf{g}(\mathbf{y}_0 + t\mathbf{e}_j)\big) = \mathbf{f}\big(\mathbf{g}(\mathbf{y}_0)\big) + \mathbf{L}\big(\Delta_j(t)\big) + \mathrm{o}\big(\Delta_j(t)\big).$$

Comme $\mathbf{f}\big(\mathbf{g}(\mathbf{y}_0 + t\mathbf{e}_j)\big) = \mathbf{y}_0 + t\mathbf{e}_j$ et $\mathbf{f}\big(\mathbf{g}(\mathbf{y}_0)\big) = \mathbf{y}_0$:

$$t\mathbf{e}_j = \mathbf{L}\big(\Delta_j(t)\big) + \mathrm{o}\big(\Delta_j(t)\big).$$

Comme $\mathbf{L}$ est inversible, $\Delta_j(t) = \mathrm{O}(t)$, et

$$\Delta_j(t) = \mathbf{L}^{-1}(\mathbf{e}_j) \cdot t + \mathrm{o}(t).$$

On voit donc que pour tout $j \in [\![1, p]\!]$, la dérivée partielle $\dfrac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x}_0)$ existe et vaut $\mathbf{L}^{-1}(\mathbf{e}_j)$.

En remplaçant $\mathbf{x}_0$ par n'importe quel point $\mathbf{x} \in V$, on voit donc que $\mathbf{g}$ admet des dérivées partielles dans toutes les directions en tout point $\mathbf{y} \in W$. La matrice des dérivées partielles est :

$$J_{\mathbf{y}}\mathbf{g} = \big(J_{\mathbf{g}(\mathbf{y})}\mathbf{f}\big)^{-1}.$$

Comme $\mathbf{g}$ est continue sur W et comme $\mathbf{f}$ est de classe $\mathcal{C}^1$ sur $\mathbf{g}(W) = V \subset U$, les coefficients de la matrice $J_{\mathbf{g}(\mathbf{y})}\mathbf{f}$ sont des fonctions continues de $\mathbf{y} \in W$. Le déterminant de la matrice ne s'annule pas. Les coefficients de la matrice $J_{\mathbf{y}}\mathbf{g}$ dépendent donc continûment de $\mathbf{y} \in W$. Cela montre que $\mathbf{g}$ est une fonction de classe $\mathcal{C}^1$ sur W. ∎

2 Sous-variétés de $\mathbb{R}^n$

2.1 Définition et exemples

Il existe plusieurs approches à la notion de sous-variété. Ces approches ne définissent pas toujours les mêmes objets. L'approche que nous suivrons ici consiste à définir les sous-variétés de $\mathbb{R}^n$ par des équations.

Définition 3. Un ensemble $X \subset \mathbb{R}^n$ est une *sous-variété lisse* de $\mathbb{R}^n$ de dimension k si pour tout $\mathbf{a} \in X$, il existe un ouvert $U \subset \mathbb{R}^n$ contenant $\mathbf{a}$ et une application $\mathbf{F} : U \to \mathbb{R}^{n-k}$ continûment différentiable telle que $X \cap U$ soit l'ensemble des solutions de l'équation $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ et telle que le rang de $D_{\mathbf{a}}\mathbf{F}$ soit égal à $n - k$.

Le terme « lisse » renvoie à la requête que l'application $\mathbf{F}$ soit une application de classe $\mathcal{C}^1$, mais aussi à la condition sur le rang. Nous verrons plus loin qu'une sous-variété lisse de $\mathbb{R}^n$ admet en tout point un espace tangent. Autrement dit, une variété lisse n'a pas d'« angle ».

Dans $\mathbb{R}^2$, l'ensemble des points (x, y) tels que $y^2 = 0$ est une droite. C'est l'axe des abscisses. Si l'on pose $F(x, y) = y^2$, alors les dérivées partielles de F par rapport à x et à y sont identiquement nulles le long de l'axe des abscisses. Cependant, X est une sous-variété lisse de $\mathbb{R}^2$ de dimension 1 puisque c'est aussi l'ensemble des points (x, y) tels que $G(x, y) := y = 0$ et que $\dfrac{\partial G}{\partial y}$ ne s'annule pas.

Exemples.

1. Un ouvert $U \subset \mathbb{R}^n$ est une sous-variété lisse de $\mathbb{R}^n$ de dimension n. En effet, l'ouvert U est l'ensemble des solutions de l'équation $\mathbf{F}(\mathbf{x}) = 0$ où $\mathbf{F} : U \to \mathbb{R}^0$ est l'application constante égale à 0 et où $\mathbb{R}^0$ est l'espace vectoriel trivial $\{0\}$ qui ne contient que le vecteur nul. En particulier, $\mathbb{R}^n$ est une sous-variété lisse de $\mathbb{R}^n$ de dimension n.
2. Les hyperplans affines de $\mathbb{R}^n$ sont des sous-variétés lisses de dimension $n-1$. Ils sont donnés par une équation de la forme $\mathbf{L}(\mathbf{x}) = \mathbf{b}$ avec $\mathbf{L} : \mathbb{R}^n \to \mathbb{R}$ une forme linéaire non nulle et $\mathbf{b} \in \mathbb{R}^n$. On peut donc considérer l'application affine $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}$ définie par $\mathbf{F}(\mathbf{x}) := \mathbf{L}(\mathbf{x}) - \mathbf{b}$. L'application $\mathbf{F}$ est différentiable sur $\mathbb{R}^n$ et sa différentielle est constante égale à $\mathbf{L}$. Elle est donc partout de rang 1.
3. Plus généralement, les sous-espaces affines de $\mathbb{R}^n$ de dimension k sont des sous-variétés lisses de $\mathbb{R}^n$ de dimension k. Les droites de $\mathbb{R}^n$ sont des sous-variétés lisses de dimension 1.
4. Les cercles de $\mathbb{R}^2$ sont des sous-variétés lisses de $\mathbb{R}^2$ de dimension 1. En effet, le cercle de centre (x_0, y_0) et de rayon $r > 0$ est donné par l'équation :
$$(x - x_0)^2 + (y - y_0)^2 - r^2 = 0.$$
L'application $F : \mathbb{R}^2 \to \mathbb{R}$ définie par $F(x, y) := (x - x_0)^2 + (y - y_0)^2 - r^2$ a pour dérivées partielles :
$$\frac{\partial F}{\partial x}(x, y) = 2(x - x_0) \quad \text{et} \quad \frac{\partial F}{\partial y}(x, y) = 2(y - y_0).$$
L'application F est de classe $\mathcal{C}^1$ sur $\mathbb{R}^2$ et la différentielle ne s'annule qu'au point (x_0, y_0). Par conséquent, la différentielle de F est de rang constant égal à 1 au voisinage du cercle de centre (x_0, y_0) et de rayon $r > 0$. Cela montre que ce cercle est une sous-variété lisse de $\mathbb{R}^2$ de dimension 1.
5. Un argument similaire montre que les sphères de $\mathbb{R}^3$ sont des sous-variétés lisses de $\mathbb{R}^3$ de dimension 2 données par des équations $F(x, y, z) = 0$ avec $F : \mathbb{R}^3 \to \mathbb{R}$ une application de la forme $F(x, y, z) := (x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 - r^2$ dont la différentielle ne s'annule qu'en (x_0, y_0, z_0) et est de rang 1 partout ailleurs.
6. Le graphe d'une application $\mathbf{f} : U \subset \mathbb{R}^k \to \mathbb{R}^{n-k}$ de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbb{R}^k$, c'est-à-dire, l'ensemble des points $(\mathbf{x}, \mathbf{y}) \in U \times \mathbb{R}^{n-k} = \mathbb{R}^n$ tels que $\mathbf{y} = \mathbf{f}(\mathbf{x})$ est une sous-variété lisse de $\mathbb{R}^n$ de dimension k. En effet, c'est l'ensemble des solutions de l'équation $\mathbf{F}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ avec $\mathbf{F} : U \times \mathbb{R}^{n-k} \to \mathbb{R}^{n-k}$

l'application définie par $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{y} - \mathbf{f}(\mathbf{x})$. L'application $\mathbf{F}$ est de classe $\mathcal{C}^1$ et sa différentielle de $\mathbf{F}$ est de rang constant égal à $n-k$ puisque la matrice des dérivées partielles de $\mathbf{F}$ par rapport aux y_j est la matrice identité I_{n-k}.

Des contre-exemples sont présentés dans les exercices en fin de chapitre. La démonstration fait presque toujours intervenir la proposition 9, qui dit qu'*une sous-variété lisse est localement le graphe d'une fonction de classe* $\mathcal{C}^1$, et qui est la réciproque du résultat suivant.

Proposition 8. Soit $U \subset \mathbb{R}^n$ un ouvert de $\mathbb{R}^n$ et $F : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$. Soit $t \in F(U)$ un réel tel que pour tout $\mathbf{x} \in U$ vérifiant $F(\mathbf{x}) = t$, l'une au moins des dérivées partielles $\dfrac{\partial F}{\partial x_j}(\mathbf{x})$ soit non nulle. Alors, l'ensemble :

$$X_t := \big\{\mathbf{x} \in U \;\big|\; F(\mathbf{x}) = t\big\}$$

est une sous-variété lisse de $\mathbb{R}^n$ de dimension $n-1$.

Démonstration. La différentielle $D_{\mathbf{x}}F$ est de rang maximal 1 si, et seulement si, ce n'est pas l'application nulle, donc si, et seulement si, l'une au moins des dérivées partielles $\dfrac{\partial F}{\partial x_j}(\mathbf{x})$ est non nulle. ∎

Exemple. Soit $f : \mathbb{C} \to \mathbb{C}$ la fonction $z \mapsto z^2 - 1$ et pour $r > 0$, soit X_r l'ensemble des points $z \in \mathbb{C}$ tels que $\big|f(z)\big| = r$. Si $r \neq 1$, l'ensemble X_r est une sous-variété lisse de $\mathbb{C} \simeq \mathbb{R}^2$. En effet, cet ensemble est donné par l'équation $F(x,y) = 0$ avec :

$$F(x,y) = (x^2 - y^2 - 1)^2 + (2xy)^2 - r^2.$$

L'application F est polynomiale en x et y, elle est donc de classe $\mathcal{C}^1$ sur $\mathbb{R}^2$. De plus :

$$\frac{\partial F}{\partial x}(x,y) = 4x(x^2 - y^2 - 1) + 8xy^2 = 4x(x^2 + 3y^2 - 1)$$

et :

$$\frac{\partial F}{\partial y}(x,y) = -4y(x^2 - y^2 - 1) + 8x^2y = 4y(x^2 + y^2 + 1).$$

Pour que la différentielle ne soit pas de rang 1, il faut que les deux dérivées partielles soient simultanément nulles. La dérivée partielle par rapport à y ne s'annule que lorsque $y = 0$. Pour que la dérivée partielle par rapport à x soit simultanément nulle, il faut que $x(x^2-1)$ soit nul, ce qui ne se produit que pour $x = 0$ et $x = \pm 1$.

Dans le premier cas, $F(0,0) = 1 - r^2$ ne s'annule que si $r = 1$. Dans le second cas, $F(1,0) = F(-1,0) = -r^2$ ne s'annule que si $r = 0$. Par conséquent, si $r > 0$ et $r \neq 1$, l'ensemble X_r est une sous-variété lisse de $\mathbb{C}$ de dimension 1.

Notons que pour $r = 0$, l'ensemble X_0 est une sous-variété lisse de $\mathbb{C}$ de dimension 0. En effet, $f(z) = 0$ si, et seulement si, $z = \pm 1$. L'ensemble X_0 est donc la réunion de deux points.

Exercice 4.
Soit $F : \mathbb{R}^2 \to \mathbb{R}$ l'application définie par $F(x, y) = x^2 - y^2$. Pour quelles valeurs de $t \in \mathbb{R}$ l'ensemble défini par l'équation $F(x, y) = t$ est-il une sous-variété lisse de $\mathbb{R}^2$?
Solution. Les dérivées partielles de F par rapport à x et à y sont :

$$\frac{\partial F}{\partial x}(x, y) = 2x \quad \text{et} \quad \frac{\partial F}{\partial y}(x, y) = -2y.$$

Elles ne s'annulent simultanément qu'au point $(0, 0)$. En ce point, la fonction F prend la valeur 0. Par conséquent, pour $t \neq 0$, l'ensemble défini par l'équation $F(x, y) = t$ est une sous-variété lisse de $\mathbb{R}^2$ de dimension 1.
Attention, ce n'est pas parce que la différentielle de F s'annule au point $(0, 0)$ que l'ensemble des points (x, y) tels que $F(x, y) = F(0, 0) = 0$ n'est pas une sous-variété lisse de $\mathbb{R}^2$. Cela dit, $(x, y) \in X_0$ si, et seulement si, $(x - y)(x + y) = 0$. L'ensemble X_0 est donc la réunion des deux droites d'équations $y = x$ et $y = -x$ qui se coupent à l'origine. L'ensemble X_0 n'est donc pas une sous-variété lisse de $\mathbb{R}^2$ puisqu'au voisinage de $(0, 0)$, X_0 n'est le graphe ni d'une fonction exprimant y en fonction de x, ni d'une fonction exprimant x en fonction de y (voir la proposition 9 de la page suivante).

Notons que la définition de sous-variété est locale. Si $X \subset \mathbb{R}^n$ est une sous-variété lisse de $\mathbb{R}^n$ et si U est un ouvert de $\mathbb{R}^n$, alors $X \cap U$ est une sous-variété lisse de $\mathbb{R}^n$ de même dimension. Pour montrer que $X \subset \mathbb{R}^n$ est une sous-variété lisse de dimension k, il suffit de montrer que pour chaque point $\mathbf{a} \in X$, il existe un ouvert $U \subset \mathbb{R}^n$ contenant $\mathbf{a}$ tel que $X \cap U$ soit une sous-variété lisse de dimension k.

2.2 Sous-variétés paramétrées

Il existe généralement deux approches à la notion de sous-variété. La première consiste à définir les sous-variétés comme l'ensemble des solutions d'une équation. C'est ce que nous avons fait ici. La seconde consiste à définir une sous-variété comme l'image d'une fonction. On dit alors que l'on a une sous-variété paramétrée.
Nous avons déjà rencontré ces deux approches dans [L1]. Un sous-espace vectoriel F de $\mathbb{R}^n$ peut être caractérisé par un système d'équations cartésiennes. Ce sous-espace est alors décrit comme l'ensemble des solutions d'un système linéaire. C'est le noyau d'une application linéaire. Il est alors facile de vérifier qu'un vecteur donné de $\mathbb{R}^n$ appartient à F. En revanche, pour trouver des vecteurs appartenant à F, pour trouver une base de F par exemple, il faut résoudre un système linéaire.
On peut également donner une équation paramétrique de F. Cela revient à décrire F comme l'image d'une application linéaire, comme l'ensemble des combinaisons linéaires d'une famille de vecteurs. Il est alors généralement facile de produire des vecteurs appartenant à F mais il est plus difficile de vérifier si un vecteur donné appartient à F.
Dans le cas d'une sous-variété lisse X de $\mathbb{R}^n$ donnée par une équation, il est facile de vérifier qu'un point donné de $\mathbb{R}^n$ appartient ou n'appartient pas à X. Il est généralement plus difficile de se représenter X. Si l'on se donne une paramétrisation de X, il est facile de

produire des points appartenant à X (ce qui permet de visualiser X plus facilement), mais il est plus difficile de vérifier si un point particulier de $\mathbb{R}^n$ appartient ou n'appartient pas à X. Puisque le graphe d'une application $\mathbf{f} : U \subset \mathbb{R}^k \to \mathbb{R}^{n-k}$ de classe $\mathcal{C}^1$ est une sous-variété lisse de $\mathbb{R}^n$ de dimension k, on voit qu'un ensemble X qui est localement le graphe d'une application $\mathbf{f}$ de classe $\mathcal{C}^1$ qui exprime $n-k$ variables en fonction des k autres est une sous-variété lisse de $\mathbb{R}^n$ de dimension k.

Proposition 9. Un ensemble $X \subset \mathbb{R}^n$ est une sous-variété lisse de $\mathbb{R}^n$ de dimension k si, et seulement si, pour tout point $\mathbf{a} := (a_1, \ldots, a_n) \in X$, il existe

- un voisinage U de $\mathbf{a}$ dans $\mathbb{R}^n$,
- une permutation σ de $[\![1, n]\!]$,
- un voisinage V de $(a_{\sigma(1)}, \ldots, a_{\sigma(k)})$ dans $\mathbb{R}^k$ et
- une application continûment différentiable $\mathbf{f} : V \to \mathbb{R}^{n-k}$

tels que $\mathbf{x} := (x_1, \ldots, x_n) \in X \cap U$ si, et seulement si, $\mathbf{x} \in V$ et :

$$(x_{\sigma(k+1)}, \ldots, x_{\sigma(n)}) = \mathbf{f}(x_{\sigma(1)}, \ldots, x_{\sigma(k)}).$$

Démonstration. Si X est localement le graphe d'une application $\mathbf{f}$ de classe $\mathcal{C}^1$ exprimant $(x_{\sigma(k+1)}, \ldots, x_{\sigma(n)})$ en fonction de $(x_{\sigma(1)}, \ldots, x_{\sigma(k)})$, alors X est localement l'ensemble des solutions de l'équation $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ avec :

$$\mathbf{F}(\mathbf{x}) = (x_{\sigma(k+1)}, \ldots, x_{\sigma(n)}) - \mathbf{f}(x_{\sigma(1)}, \ldots, x_{\sigma(k)}).$$

La matrice des dérivées partielles de $\mathbf{F}$ par rapport aux $x_{\sigma(j)}$ pour $j \in [\![k+1, n]\!]$ est la matrice identité I_{n-k}. Par conséquent, le rang de F dans U est $n-k$ et X est une sous-variété lisse de $\mathbb{R}^n$ de dimension k.

Réciproquement, si X est une sous-variété lisse de $\mathbb{R}^n$ de dimension k alors au voisinage de chaque point $\mathbf{a} \in X$, l'ensemble X est l'ensemble des solutions de l'équation $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ avec $\mathbf{F}$ continûment différentiable et $D_{\mathbf{a}}\mathbf{F}$ de rang $n-k$. Il existe donc une permutation σ de $[\![1, n]\!]$ telle que la matrice :

$$\begin{pmatrix} \dfrac{\partial F_1}{\partial x_{\sigma(k+1)}} & \cdots & \dfrac{\partial F_1}{\partial x_{\sigma(n)}} \\ \vdots & & \vdots \\ \dfrac{\partial F_{n-k}}{\partial x_{\sigma(k+1)}} & \cdots & \dfrac{\partial F_{n-k}}{\partial x_{\sigma(n)}} \end{pmatrix}$$

soit inversible. Le théorème des fonctions implicites dit que l'ensemble des solutions de l'équation $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ est localement le graphe d'une application de classe $\mathcal{C}^1$ exprimant $(x_{\sigma(k+1)}, \ldots, x_{\sigma(n)})$ en fonction de $(x_{\sigma(1)}, \ldots, x_{\sigma(k)})$. ∎

Exemples.

1. Le graphe X de l'application $x \mapsto |x|$ n'est pas une sous-variété lisse de $\mathbb{R}^2$. Au voisinage du point $(0, 0)$, c'est le graphe d'une unique application exprimant y en fonction de x. Cette fonction est la fonction « valeur absolue » qui n'est pas de classe $\mathcal{C}^1$.

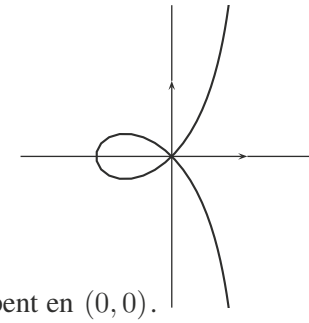
Ce n'est pas le graphe d'une application exprimant x en fonction de y puisqu'il n'y a aucun point (x, y) appartenant à X avec $y < 0$.

2. Dans $\mathbb{R}^2$, la réunion X des deux axes de coordonnées n'est pas une sous-variété lisse. Au voisinage de $(0, 0)$, il y a une infinité de points ayant leur première coordonnée nulle (les points de l'axe des ordonnées) et une infinité de points ayant leur seconde coordonnée nulle (les points de l'axe des abscisses). Au voisinage de $(0, 0)$ on ne peut donc pas exprimer x en fonction de y ou y en fonction de x.

Avant de définir ce qu'est un paramétrage d'une sous-variété de $\mathbb{R}^n$, nous souhaitons considérer le cas des courbes de $\mathbb{R}^2$ pour souligner les problèmes que l'on peut rencontrer. Il est naturel de se demander si toute courbe paramétrée $\gamma : I \subset \mathbb{R} \to \mathbb{R}^2$ définit une sous-variété lisse de $\mathbb{R}^2$ de dimension 1. Plus précisément, on se demande si le support $f(I)$ de la courbe est une sous-variété lisse de$\mathbb{R}^2$.
Si γ n'est pas une fonction de classe $\mathcal{C}^1$, il n'y a aucune raison pour que son support soit « lisse ». Par exemple, le graphe d'une fonction continue $f : \mathbb{R} \to \mathbb{R}$ n'est lisse que si f est de classe $\mathcal{C}^1$.
Même dans le cas où γ est une courbe de classe $\mathcal{C}^1$, il se peut que la courbe ait des points de rebroussement (voir le module III.2 de [L1]) si la dérivée de γ s'annule. Dans ce cas, le support de la courbe n'est pas une sous-variété lisse de $\mathbb{R}^2$.
Si la courbe a des points doubles, c'est-à-dire si γ n'est pas injective, il se peut que le support de la courbe ne soit pas une sous-variété lisse de $\mathbb{R}^2$. Au point de croisement, le support de la courbe ne sera généralement pas localement le graphe d'une fonction $x \mapsto y = f(x)$ ou d'une fonction $y \mapsto x = f(y)$.

Exemple. C'est par exemple le cas de la strophoïde droite d'équation $(x-1)y^2 + x^2(x+1) = 0$ que l'on peut aussi obtenir par la représentation paramétrique :

$$\begin{cases} x = \dfrac{t^2-1}{t^2+1} \\ y = \dfrac{t(t^2-1)}{t^2+1}. \end{cases}$$



Au voisinage de l'origine, on a deux arcs de courbe qui se coupent en $(0, 0)$.

Ces observations motivent la définition suivante.

Définition 4. Un *paramétrage* (global) d'une sous-variété lisse $X \subset \mathbb{R}^n$ de dimension k est une application $\gamma : V \subset \mathbb{R}^k \to X$ vérifiant :
- V est un ouvert de $\mathbb{R}^k$,
- $\gamma : V \to X$ est une bijection de classe $\mathcal{C}^1$ et
- la différentielle $D_{\mathbf{z}}\gamma$ est injective en tout point $\mathbf{z} \in V$.

Remarque. Une sous-variété lisse $X \subset \mathbb{R}^n$ de dimension k n'admet généralement pas de paramétrage global. Par exemple, on ne peut pas trouver d'application continue et bijective entre un intervalle ouvert de $\mathbb{R}$ et le cercle de centre $(0,0)$ et de rayon 1 de $\mathbb{R}^2$. La proposition 9 de la page 722 montre que toute sous-variété lisse X de $\mathbb{R}^n$ admet des paramétrages locaux au voisinage de chaque point, c'est-à-dire que pour chaque point $\mathbf{a} \in X$, il existe un voisinage U de $\mathbf{a}$ dans $\mathbb{R}^n$ tel que $X \cap U$ admette un paramétrage.

Exercice 5.

Montrer qu'il n'existe pas d'application continue et bijective entre $]a,b[$ et le cercle de centre $(0,0)$ et de rayon 1 de $\mathbb{R}^2$.

Solution.

Supposons qu'il existe une telle application continue et bijective $f : t \mapsto \big(x(t), y(t)\big)$. Posons $c := f^{-1}(-1,0) \in]a,b[$ et pour $t \neq c$, posons :

$$g(t) := 2\arctan \frac{y(t)}{x(t) + \sqrt{x^2(t) + y^2(t)}}.$$

le nombre $g(t) \in]-\pi, \pi[$ est l'argument du nombre complexe $x(t) + iy(t) \in \mathbb{U} \setminus \{-1\}$. L'application g est continue sur $]a,b[\setminus \{c\}$ comme composée d'applications continues. C'est une bijection continue entre $]a,c[\cup]c,b[$ et $]-\pi,\pi[$. En particulier, g est strictement monotone sur $]a,c[$ et sur $]c,b[$. Alors, $g\big(]a,c[\big) =]\theta_1,\theta_2[$ et $g\big(]c,b[\big) =]\theta_3,\theta_4[$. Puisque g est bijective, ces deux intervalles sont non vides et disjoints. De plus, $]-\pi,\pi[=]\theta_1,\theta_2[\cup]\theta_3,\theta_4[$. Cela n'est pas possible puisque soit $-\pi < \theta_2 \leqslant \theta_3 < \pi$ et dans ce cas, la réunion des deux intervalles ne contient pas θ_2. Soit $-\pi < \theta_4 \leqslant \theta_1 < \pi$ et dans ce cas, la réunion des deux intervalles ne contient pas θ_1.

2.3 Espaces tangents

Pour étudier le comportement d'une application de plusieurs variables $\mathbf{f}$ au voisinage d'un point $\mathbf{x}$, nous avons d'abord étudié le comportement de la différentielle $D_{\mathbf{x}}\mathbf{f}$ puis nous avons montré (théorème d'inversion locale et théorème des fonctions implicites) que le comportement de $\mathbf{f}$ et de $D_{\mathbf{x}}\mathbf{f}$ sont intimement liés.

Pour étudier la géométrie d'une sous-variété lisse X de $\mathbb{R}^n$ au voisinage d'un point $\mathbf{a} \in X$, il est utile dans un premier temps de remplacer X par un objet linéaire : l'espace tangent à X en $\mathbf{a}$.

Commençons par définir l'espace tangent à $\mathbb{R}^n$ en un point $\mathbf{a} \in \mathbb{R}^n$. Rappelons que $\mathbb{R}^n$ est un espace affine de direction l'espace vectoriel $\mathbb{R}^n$ (pour ce vocabulaire, voir le module III.1 de [L1]). À un couple $(\mathbf{a},\mathbf{b})$ de points de l'espace affine $\mathbb{R}^n$, on associe le vecteur $\overrightarrow{\mathbf{ab}}$ de l'espace vectoriel $\mathbb{R}^n$: si $\mathbf{a} := (a_1,\ldots,a_n)$ et $\mathbf{b} := (b_1,\ldots,b_n)$, alors $\overrightarrow{\mathbf{ab}} := (b_1 - a_1, \ldots, b_n - a_n)$. Si $\mathbf{a}$ est un point de l'espace affine $\mathbb{R}^n$ et si $\mathbf{h}$ est un vecteur de l'espace vectoriel $\mathbb{R}^n$, il existe un unique point $\mathbf{b}$ de l'espace affine $\mathbb{R}^n$ tel que $\overrightarrow{\mathbf{ab}} = \mathbf{h}$: si $\mathbf{a} := (a_1,\ldots,a_n)$ et si $\mathbf{h} := (h_1,\ldots,h_n)$, ce point est $\mathbf{b} := (a_1 + h_1, \ldots, a_n + h_n)$. On le note $\mathbf{b} = \mathbf{a} + \mathbf{h}$.

En particulier, lorsque nous écrivons la formule :

$$\mathbf{f}(\mathbf{a}+\mathbf{h}) = \mathbf{f}(\mathbf{a}) + D_{\mathbf{a}}\mathbf{f}(\mathbf{h}) + \|\mathbf{h}\| \cdot \varepsilon(\mathbf{h}),$$

avec $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ une application différentiable, nous considérons implicitement que $\mathbf{a}$ est un point de l'espace affine $\mathbb{R}^n$, que $\mathbf{f}(\mathbf{a})$ est un point de l'espace affine $\mathbb{R}^m$, que $\mathbf{h}$ est un vecteur de l'espace vectoriel $\mathbb{R}^n$ et que $D_{\mathbf{a}}\mathbf{f}(\mathbf{h})$ et $\varepsilon(\mathbf{h})$ sont des vecteurs de l'espace vectoriel $\mathbb{R}^m$, le dernier étant multiplié par le scalaire $\|\mathbf{h}\|$ dans la formule.
Ce point de vue n'est pas tout à fait ce qui convient pour ce que nous voulons faire. Il est préférable de penser à $\mathbf{h}$ comme à un vecteur « attaché » au point $\mathbf{a}$, de sorte qu'il est possible de parler du point $\mathbf{a}+\mathbf{h}$ mais qu'il n'est pas possible de parler du point $\mathbf{b}+\mathbf{h}$ si $\mathbf{b} \neq \mathbf{a}$. De même, il est préférable de penser à $D_{\mathbf{a}}\mathbf{f}(\mathbf{h})$ et $\varepsilon(\mathbf{h})$ comme à des vecteurs « attachés » au point $\mathbf{f}(\mathbf{a})$, de sorte qu'il est possible d'ajouter le vecteur $D_{\mathbf{a}}\mathbf{f}(\mathbf{h})+\|\mathbf{h}\|\cdot\varepsilon(\mathbf{h})$ au point $\mathbf{f}(\mathbf{a})$ (mais pas à un autre point de $\mathbb{R}^m$).
Il s'agit donc de distinguer le plan tangent à $\mathbb{R}^n$ aux points $\mathbf{a}$ et $\mathbf{b}$, bien que tous deux s'identifient naturellement à $\mathbb{R}^n$.

Définition 5. L'*espace tangent* à $\mathbb{R}^n$ au point $\mathbf{a}$ est l'espace vectoriel $T_{\mathbf{a}}\mathbb{R}^n$ des couples $(\mathbf{a},\mathbf{b})$ avec $\mathbf{b} \in \mathbb{R}^n$ muni de la loi interne $+$ et de la loi externe $\cdot$ suivantes :

- si $\mathbf{h} := (\mathbf{a},\mathbf{b}) \in T_{\mathbf{a}}\mathbb{R}^n$ et $\mathbf{k} := (\mathbf{a},\mathbf{c}) \in T_{\mathbf{a}}\mathbb{R}^n$, alors $\mathbf{h}+\mathbf{k} := (\mathbf{a},\mathbf{d}) \in T_{\mathbf{a}}\mathbb{R}^n$ avec $\mathbf{d} := \mathbf{a} + \left(\overrightarrow{\mathbf{ab}} + \overrightarrow{\mathbf{ac}}\right)$;
- si $\mathbf{h} := (\mathbf{a},\mathbf{b}) \in T_{\mathbf{a}}\mathbb{R}^n$ et $\lambda \in \mathbb{R}$, alors $\lambda\cdot\mathbf{h} := (\mathbf{a},\mathbf{c}) \in T_{\mathbf{a}}\mathbb{R}^n$ avec $\mathbf{c} := \mathbf{a}+\lambda\cdot\overrightarrow{\mathbf{ab}}$.

L'application $(\mathbf{a},\mathbf{b}) \in T_{\mathbf{a}}\mathbb{R}^n \mapsto \overrightarrow{\mathbf{ab}} \in \mathbb{R}^n$ fournit un isomorphisme canonique d'espaces vectoriels. Le fait de garder en mémoire le point $\mathbf{a}$ permet de distinguer des espaces tangents à $\mathbb{R}^n$ en des points distincts. Si $(\mathbf{e}_1,\ldots,\mathbf{e}_n)$ est la base canonique de $\mathbb{R}^n$, alors $\big((\mathbf{a},\mathbf{a}+\mathbf{e}_1),\ldots,(\mathbf{a},\mathbf{a}+\mathbf{e}_n)\big)$ est la base canonique de $T_{\mathbf{a}}\mathbb{R}^n$.
Nous invitons dorénavant le lecteur à penser que la différentielle $D_{\mathbf{a}}\mathbf{f}$ d'une application $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est une application linéaire $D_{\mathbf{a}}\mathbf{f} : T_{\mathbf{a}}\mathbb{R}^n \to T_{\mathbf{f}(\mathbf{a})}\mathbb{R}^m$ de l'espace tangent à $\mathbb{R}^n$ en $\mathbf{a}$ vers l'espace tangent à $\mathbb{R}^m$ en $\mathbf{f}(\mathbf{a})$. La matrice de cette application linéaire dans les bases canoniques respectives de $T_{\mathbf{a}}\mathbb{R}^n$ et $T_{\mathbf{f}(\mathbf{a})}\mathbb{R}^m$ est la matrice Jacobienne $J_{\mathbf{a}}\mathbf{f}$.
Considérons maintenant le cas où $X \subset \mathbb{R}^n$ est une sous-variété lisse de $\mathbb{R}^n$.

Proposition 10. Soit X une sous-variété lisse de $\mathbb{R}^n$ de dimension k et soit $\mathbf{a}$ un point de X. Supposons d'une part que X soit définie par l'équation $\mathbf{F}(\mathbf{x}) = 0$ dans un voisinage $U \subset \mathbb{R}^n$ de $\mathbf{a}$ avec $\mathbf{F} : U \to \mathbb{R}^{n-k}$ une application continûment différentiable et $D_{\mathbf{a}}\mathbf{F}$ de rang $n-k$. Supposons d'autre part que $\gamma : V \subset \mathbb{R}^k \to \mathbb{R}^n$ soit un paramétrage local de X au voisinage de $\mathbf{a}$ avec $\gamma(\mathbf{z}_0) = \mathbf{a}$. Alors :

$$\operatorname{Ker} D_{\mathbf{a}}\mathbf{F} = \operatorname{Im} D_{\mathbf{z}_0}\gamma.$$

Démonstration. La différentielle $D_{\mathbf{z}_0}\gamma : T_{\mathbf{z}_0}\mathbb{R}^k \to T_{\mathbf{a}}\mathbb{R}^n$ est injective, donc de rang k d'après le théorème du rang. L'image de $D_{\mathbf{z}_0}\gamma$ est donc un sous espace vectoriel de $T_{\mathbf{a}}\mathbb{R}^n$ de dimension k.

La différentielle $D_{\mathbf{a}}\mathbf{F} : T_{\mathbf{a}}\mathbb{R}^n \to T_{\mathbf{f(a)}}\mathbb{R}^{n-k}$ est de rang $n-k$, donc son noyau est un sous-espace vectoriel de $T_{\mathbf{a}}\mathbb{R}^n$ de dimension k.

Au voisinage de $\mathbf{z}_0$, l'application $\mathbf{F} \circ \gamma$ est identiquement nulle. Donc, $D_{\mathbf{a}}\mathbf{F} \circ D_{\mathbf{z}_0}\gamma$ est l'application nulle. En particulier, l'image de $D_{\mathbf{z}_0}\gamma$ est contenue dans le noyau de $D_{\mathbf{a}}\mathbf{F}$.

Cela montre que les deux sous-espaces vectoriels de $T_{\mathbf{a}}\mathbb{R}^n$ sont égaux. ∎

En particulier, l'image de $D_{\mathbf{z}_0}\gamma$ ne dépend pas du paramétrage local γ envoyant $\mathbf{z}_0$ sur $\mathbf{a}$. La définition suivante a donc un sens.

Définition 6.

Si X est une sous-variété lisse de $\mathbb{R}^n$ de dimension k et si $\gamma : V \subset \mathbb{R}^k \to \mathbb{R}^n$ est un paramétrage local de X au voisinage d'un point $\mathbf{a} \in X$, envoyant $\mathbf{z}_0$ sur $\mathbf{a}$, alors l'*espace tangent* $T_{\mathbf{a}}X$ à X au point $\mathbf{a}$ est l'image de l'application linéaire $D_{\mathbf{z}_0}\gamma : T_{\mathbf{z}_0}\mathbb{R}^k \to T_{\mathbf{a}}\mathbb{R}^n$.

La proposition précédente montre que l'on peut donner la définition équivalente.

Définition 7.

Si X est une sous-variété lisse de $\mathbb{R}^n$ de dimension k définie par l'équation $\mathbf{F}(\mathbf{x}) = 0$ dans un voisinage $U \subset \mathbb{R}^n$ d'un point $\mathbf{a} \in X$ avec $\mathbf{F} : U \to \mathbb{R}^{n-k}$ une application continûment différentiable et $D_{\mathbf{a}}\mathbf{F}$ de rang $n-k$, alors l'*espace tangent* $T_{\mathbf{a}}X$ à X au point $\mathbf{a}$ est le noyau de l'application linéaire $D_{\mathbf{a}}\mathbf{F} : T_{\mathbf{a}}\mathbb{R}^n \to T_{\mathbf{F(a)}}\mathbb{R}^{n-k}$.

Il ne faut pas confondre l'espace tangent à une courbe de $\mathbb{R}^2$ ou de $\mathbb{R}^3$ et la tangente à la courbe. L'espace tangent est un espace vectoriel, tandis que la tangente est un espace affine.

Exercice 6.

1. Pour quelles valeurs de $c \in \mathbb{R}$ l'ensemble X_c d'équation $x^2 - y^3 = c$ est-il une sous-variété lisse de $\mathbb{R}^2$?
2. Pour de telles valeurs de c, donner une équation de l'espace tangent à X_c en un point (x_0, y_0).
3. Donner également une équation de la tangente à un tel X_c en un point (x_0, y_0).

Solution.

1. Considérons la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x,y) := x^2 - y^3 - c$.

   Alors, $\dfrac{\partial f}{\partial x}(x,y) = 2x$ et $\dfrac{\partial f}{\partial y}(x,y) = -3y^2$. Les deux dérivées partielles ne s'annulent simultanément qu'au point $(0,0)$. Par conséquent, l'ensemble des solutions de l'équation $f(x,y) = 0$ est lisse dès qu'il ne contient pas le point $(0,0)$, c'est-à-dire dès que $c \neq 0$. Pour $c = 0$, cet ensemble n'est pas lisse en $(0,0)$ comme on peut le voir à l'aide de la proposition 9 de la page 722.

2. Si $(x_0, y_0) \neq (0,0)$, l'espace tangent à X_c en (x_0, y_0) est le noyau de la différentielle :
$$D_{(x_0,y_0)}f : (h_1, h_2) \mapsto 2x_0h_1 - 3y_0^2h_2.$$
C'est donc la droite d'équation cartésienne $2x_0x - 3y_0^2y = 0$.
3. La tangente est la droite passant par le point (x_0, y_0) et dirigée par l'espace tangent à X_c en (x_0, y_0). Son équation cartésienne est $2x_0x - 3y_0^2y = 2x_0^2 - 3y_0^3$.

Supposons maintenant que $F : U \subset \mathbb{R}^n \to \mathbb{R}$ soit une fonction continûment différentiable sur un ouvert U de $\mathbb{R}^n$.
Si $\mathbf{a}$ est un point de X, la différentielle $D_{\mathbf{a}}F : T_{\mathbf{a}}X \to T_{F(\mathbf{a})}\mathbb{R} \simeq \mathbb{R}$ est une forme linéaire. Si l'on munit $T_{\mathbf{a}}\mathbb{R}^n$ d'un produit scalaire, il existe (théorème 10 de la page 147) un vecteur $\overrightarrow{\operatorname{grad}}\, F(\mathbf{a})$ tel que :
$$\forall \mathbf{h} \in T_{\mathbf{a}}\mathbb{R}^n, \quad D_{\mathbf{a}}F(\mathbf{h}) = \big(\overrightarrow{\operatorname{grad}}\, F(\mathbf{a}) \;\big|\; \mathbf{h}\big).$$
Si le produit scalaire est le produit scalaire usuel :
$$(\,\mathbf{h} \mid \mathbf{k}\,) = h_1k_1 + \cdots + h_kk_n$$
avec $(h_1, \ldots, h_n)$ et $(k_1, \cdots, k_n)$ les coordonnées de $\mathbf{h}$ et $\mathbf{k}$ dans la base canonique de $T_{\mathbf{a}}\mathbb{R}^n$, alors :
$$D_{\mathbf{a}}F(\mathbf{h}) = \frac{\partial F}{\partial x_1}(\mathbf{a})h_1 + \cdots + \frac{\partial F}{\partial x_n}(\mathbf{a})h_n = \big(\overrightarrow{\operatorname{grad}}\, F(\mathbf{a}) \;\big|\; \mathbf{h}\big)$$
avec :
$$\overrightarrow{\operatorname{grad}}\, F(\mathbf{a}) = \begin{pmatrix} \dfrac{\partial F}{\partial x_1}(\mathbf{a}) \\ \vdots \\ \dfrac{\partial F}{\partial x_n}(\mathbf{a}) \end{pmatrix}.$$
L'ensemble X des solutions de l'équation $F(\mathbf{x}) = F(\mathbf{a})$ est une sous-variété lisse de $\mathbb{R}^n$ de dimension $n-1$ au voisinage de $\mathbf{a}$ si, et seulement si, le vecteur $\overrightarrow{\operatorname{grad}}\, F(\mathbf{a})$ est non nul. L'espace tangent $T_{\mathbf{a}}X$ est alors l'orthogonal de $\overrightarrow{\operatorname{grad}}\, F(\mathbf{a})$. C'est un hyperplan de $T_{\mathbf{a}}\mathbb{R}^n$. On dit que X est une hypersurface de $\mathbb{R}^n$.

Exercice 7.
Soit X une sous-variété lisse de $\mathbb{R}^n$.

1. Soit $c : [a, b] \to X$ une courbe de classe $\mathcal{C}^1$. Montrer que $c'(a) \in T_{c(a)}X$.
2. Soit $\mathbf{a}$ un point de X et $\mathbf{h}$ un vecteur de $T_{\mathbf{a}}X$. Montrer qu'il existe une courbe $c : [a, b] \to X$ de classe $\mathcal{C}^1$ telle que $c(a) = \mathbf{a}$ et $c'(a) = \mathbf{h}$.

Solution.

1. Posons $\mathbf{a} := c(a) \in X$. Soit U un voisinage de $\mathbf{a}$ dans $\mathbb{R}^n$ tel que $X \cap U$ admette un paramétrage $\gamma : V \subset \mathbb{R}^k \to X$. Comme γ est une bijection, il existe une application $h : [a, b] \to V$ telle que $c = \gamma \circ h$. On a alors :
$$c'(a) = D_{h(a)}\gamma\big(h'(t)\big) \in \operatorname{Im}(D_{h(a)}\gamma) = T_{\mathbf{a}}X.$$

2. Soit $U \subset \mathbb{R}^n$ un voisinage du point $\mathbf{a}$ tel que $X \cap U$ admette un paramétrage $\gamma : V \subset \mathbb{R}^k \to X$. Soit $\mathbf{u}_0$ l'antécédent de $\mathbf{a}$. Comme $\mathbf{h} \in \mathrm{Im}(D_{\mathbf{u}_0}\gamma)$, il existe un vecteur $\mathbf{k} \in T_{\mathbf{u}_0}\mathbb{R}^k$ tel que $\mathbf{h} = D_{\mathbf{u}_0\gamma(\mathbf{k})}$.

   Il suffit alors de considérer la courbe $c : [0, \varepsilon] \to X$ définie par $c(t) = \gamma(\mathbf{u}_0 + t\mathbf{k})$ avec ε suffisamment petit pour que $\mathbf{u}_0 + t\mathbf{k}$ appartienne à l'ensemble de définition de γ pour tout $t \in [0, \varepsilon]$.

3 Développement de Taylor

Dans toute cette partie f désignera une fonction de plusieurs variables à valeurs dans $\mathbb{R}$. On sait (module IV.7 de [L1]) que si f est continue sur un compact K, alors f est bornée et atteint ses bornes. Peut-on déterminer ces bornes ? C'est ce problème qui guidera nos pas ici.

Commençons par voir ce que nous pouvons dire dans le cas d'une fonction d'une variable réelle. D'abord, si f est définie sur un intervalle ouvert I, y est dérivable et admet un extremum local en $a \in I$, alors $f'(a) = 0$. Cela reste vrai pour les fonctions de plusieurs variables.

Définition 8. Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ sur un ouvert U de $\mathbb{R}^n$. On dit que $\mathbf{a}$ est un *point critique* de f si $D_{\mathbf{a}}f$ s'annule en $\mathbf{a}$.

Exercice 8.

Soit $U \subset \mathbb{C}$ un ouvert et $f : U \to \mathbb{C}$ une fonction holomorphe. Montrer que les points critiques de f sont exactement les points où f' s'annule.

Solution. La différentielle de f en un point z est l'application linéaire $h \mapsto f'(z) \cdot h$. Cette application est l'application nulle si, et seulement si, $f'(z) = 0$.

Théorème 11 (Condition nécessaire d'existence d'un extremum local).
Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ sur un ouvert U de $\mathbb{R}^n$.
Si f présente en $\mathbf{a}$ un extremum local, alors $\mathbf{a}$ est un point critique de f.

Démonstration. Considérons les fonctions $g_j : t \mapsto f(\mathbf{a} + t\mathbf{e}_j)$, où $\mathbf{e}_j$ est le $j^{ème}$ vecteur de la base canonique de $\mathbb{R}^n$. Ces fonctions admettent un extremum local en $t = 0$, et :

$$\frac{\partial f}{\partial x_j}(\mathbf{a}) = g_j'(0) = 0.$$

Cela montre que les dérivées partielles de f en $\mathbf{a}$ sont toutes nulles. La matrice Jacobienne $J_{\mathbf{a}}f$ est donc la matrice nulle. Par conséquent, la différentielle $D_{\mathbf{a}}f$ est l'application linéaire nulle. ∎

Bien entendu, la condition d'annulation de la différentielle ne suffit pas à garantir l'existence d'un extremum local, l'exemple le plus fréquemment donné étant celui de la fonction $f : x \mapsto x^3$ dont la dérivée s'annule en 0 sans que f y admette un extremum local.

Pour les fonctions d'une variable réelle, on peut, dans bien des cas, utiliser le développement de Taylor de f en a pour savoir si f admet ou n'admet pas d'extremum local en a. Plus précisément, on dispose du corollaire suivant du développement de Taylor-Young d'une fonction de classe $\mathcal{C}^n$ sur un intervalle I (voir le module IV.8 de [L1]) : si l'une des dérivées de f est non nulle en $a \in I$, alors :

$$f(x) - f(a) \underset{x\to a}{\sim} \frac{(x-a)^p}{p!} f^{(p)}(a)$$

où $p \in [\![0, n]\!]$ est le plus petit entier tel que $f^{(p)}(a) \neq 0$. Comme conséquence, on obtient le résultat suivant.

Proposition 12.
Soit $f : I \to \mathbb{R}$ une fonction de classe $\mathcal{C}^n$ $(n \geqslant 2)$ sur un intervalle ouvert I. Supposons que $a \in I$ et $p \in [\![2, n]\!]$ soient tels que $f^{(k)}(a) = 0$ pour $k \in [\![1, p-1]\!]$ et $f^{(p)}(a) \neq 0$. Si p est impair, f n'admet pas d'extremum local en a. Si p est pair et $f^{(p)}(a) > 0$, alors f admet un minimum local en a. Si p est pair et $f^{(p)}(a) < 0$, alors f admet un maximum local en a.

Démonstration. Notons que :

$$f(x) - f(a) \underset{a}{\sim} \frac{(x-a)^p}{p!} f^{(p)}(a).$$

En particulier, au voisinage de a les deux fonctions sont de même signe.

Si p est impair, $\dfrac{(x-a)^p}{p!} f^{(p)}(a)$ change de signe avec $x - a$. Par conséquent, $f(x) - f(a)$ change de signe au voisinage de a. Cela montre que f n'admet pas de maximum ou de minimum local en a.

Si p est pair, alors au voisinage de a, $f(x) - f(a)$ est du même signe que $f^{(p)}(a)$:

- si $f^{(p)}(a) > 0$, alors $f(x) - f(a)$ est positif au voisinage de a et f admet un minimum local en a ;
- si $f^{(p)}(a) < 0$, alors $f(x) - f(a)$ est négatif au voisinage de a et f admet un maximum local en a.

Dans le cas particulier (le plus fréquent) où $p = 2$, on obtient le corollaire suivant.

Corollaire 13. Soit $f : I \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ sur un intervalle ouvert I. Soit $a \in I$ un point tel que $f'(a) = 0$. Si $f''(a) > 0$, f admet un minimum local en 0. Si $f''(a) < 0$, f admet un maximum local en 0.

Si $f''(a) = 0$, on ne peut rien conclure, comme le montrent les exemples $f(x) = x^p$, avec $p > 2$.

C'est ce corollaire que nous finirons par généraliser aux fonctions de n variables. Auparavant, il nous faut définir ce qu'est une fonction de classe $\mathcal{C}^k$ sur un ouvert U de $\mathbb{R}^n$. Il nous faut également généraliser la notion de développement de Taylor.

3.1 Fonctions de classe $\mathcal{C}^k$

Commençons par définir ce qu'est une fonction de classe $\mathcal{C}^2$ sur un ouvert U de $\mathbb{R}^n$.

Définition 9. Une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est *de classe* $\mathcal{C}^2$ sur U si f est de classe $\mathcal{C}^1$ sur U et si les dérivées partielles $\dfrac{\partial f}{\partial x_j}$, $j \in [\![1, n]\!]$ sont de classe $\mathcal{C}^1$ sur U.

Si f est une fonction de classe $\mathcal{C}^2$ sur U, nous utiliserons la notation :

$$\frac{\partial^2 f}{\partial x_i \partial x_j} := \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right).$$

Les fonctions $\dfrac{\partial^2 f}{\partial x_i \partial x_j}$ sont appelées les dérivées partielles d'ordre 2 de f. Si les dérivées partielles d'ordre 2 de f sont elles-même de classe $\mathcal{C}^1$, on dit que f est de classe $\mathcal{C}^3$. Les dérivées partielles des dérivées partielles d'ordre 2 de f sont appelées les dérivées partielles d'ordre 3, et ainsi de suite. On définit donc par récurrence la notion de fonction de classe $\mathcal{C}^k$.

Définition 10.
Une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est *de classe* $\mathcal{C}^k$ sur U si f est de classe $\mathcal{C}^1$ sur U et si les dérivées partielles $\dfrac{\partial f}{\partial x_j}$, $j \in [\![1, n]\!]$ sont de classe $\mathcal{C}^{k-1}$ sur U.
Les dérivées partielles d'ordre k de f sont les dérivées partielles d'ordre $k-1$ des fonctions $\dfrac{\partial f}{\partial x_j}$, $j \in [\![1, n]\!]$.
La fonction f est de classe $\mathcal{C}^\infty$ si elle est de classe $\mathcal{C}^k$ pour tout entier $k \geqslant 1$.

Exemple. Une fonction de 2 variables $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ de classe $\mathcal{C}^n$ a 2 dérivées partielles d'ordre 1 :

$$\frac{\partial f}{\partial x} \quad \text{et} \quad \frac{\partial f}{\partial y}.$$

Elle a 4 dérivées partielles d'ordre 2 :

$$\frac{\partial^2 f}{\partial x \partial x}, \quad \frac{\partial^2 f}{\partial x \partial y}, \quad \frac{\partial^2 f}{\partial y \partial x} \quad \text{et} \quad \frac{\partial^2 f}{\partial y \partial y}.$$

Elle a 8 dérivées partielles d'ordre 3 :

$$\frac{\partial^2 f}{\partial x \partial x \partial x}, \quad \frac{\partial^2 f}{\partial x \partial x \partial y}, \quad \frac{\partial^2 f}{\partial x \partial y \partial x}, \quad \frac{\partial^2 f}{\partial x \partial y \partial y},$$
$$\frac{\partial^2 f}{\partial y \partial x \partial x}, \quad \frac{\partial^2 f}{\partial y \partial x \partial y}, \quad \frac{\partial^2 f}{\partial y \partial y \partial x} \quad \text{et} \quad \frac{\partial^2 f}{\partial y \partial y \partial y}.$$

Elle a 2^n dérivées partielles d'ordre n.

Il devient impératif de trouver des notations pour simplifier l'écriture des dérivées partielles de f. Le théorème de Schwarz suivant va nous simplifier le travail.

Théorème 14 (Schwarz). Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est une fonction de classe $\mathcal{C}^1$ dont les dérivées partielles $\dfrac{\partial f}{\partial x_j}$, $j \in [\![1, n]\!]$ sont différentiables. Alors :

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

pour tout couple $(i, j) \in [\![1, n]\!]^2$.

Si f est de classe $\mathcal{C}^2$, les dérivées partielles de f sont automatiquement de classe $\mathcal{C}^1$. La conséquence la plus importante du théorème de Schwarz est que, si f est de classe $\mathcal{C}^k$, $k \geqslant 2$, toute dérivée partielle d'ordre $p \leqslant k$ de f peut s'écrire sous la forme :

$$\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \left(\frac{\partial}{\partial x_2}\right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} f$$

avec $\alpha_1 + \alpha_2 + \cdots + \alpha_n = p$. Autrement dit, les dérivées partielles d'ordre inférieur à k ne dépendent pas de l'ordre dans lequel elles sont calculées. Nous utiliserons la notation abrégée :

$$\frac{\partial^{|\alpha|} f}{\partial \mathbf{x}^\alpha} = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}$$

avec $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_n)$ et $|\alpha| := \alpha_1 + \alpha_2 + \ldots + \alpha_n$.

Démonstration du théorème de Schwarz. Soient $\mathbf{e}_i$ le $i^{\text{ème}}$ vecteur de la base canonique de $\mathbb{R}^n$ et $\mathbf{e}_j$ le $j^{\text{ème}}$ vecteur. Nous allons étudier le comportement de :

$$A(t) := f(\mathbf{a} + t\mathbf{e}_i + t\mathbf{e}_j) - f(\mathbf{a} + t\mathbf{e}_i) - f(\mathbf{a} + t\mathbf{e}_j) + f(\mathbf{a})$$

quand $t \in \mathbb{R}$ tend vers 0. Notons tout d'abord que :

$$A(t) = g(t) - g(0) \quad \text{avec} \quad g(s) := f(\mathbf{a} + s\mathbf{e}_i + t\mathbf{e}_j) - f(\mathbf{a} + s\mathbf{e}_i).$$

D'après le théorème des accroissements finis, $A(t) = t \cdot g'(c)$ avec $c \in [0, t]$. Or :

$$\begin{aligned} g'(c) &= \frac{\partial f}{\partial x_i}(\mathbf{a} + c\mathbf{e}_i + t\mathbf{e}_j) - \frac{\partial f}{\partial x_i}(\mathbf{a} + c\mathbf{e}_i) \\ &= \left(\frac{\partial f}{\partial x_i}(\mathbf{a} + c\mathbf{e}_i + t\mathbf{e}_j) - \frac{\partial f}{\partial x_i}(\mathbf{a})\right) - \left(\frac{\partial f}{\partial x_i}(\mathbf{a} + c\mathbf{e}_i) - \frac{\partial f}{\partial x_i}(\mathbf{a})\right) \\ &= \left(\frac{\partial^2 f}{\partial x_i^2}(\mathbf{a}) \cdot c + \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}) \cdot t + \mathrm{o}\big(\sqrt{c^2 + t^2}\big)\right) - \left(\frac{\partial^2 f}{\partial x_i^2}(\mathbf{a}) \cdot c + \mathrm{o}(c)\right). \end{aligned}$$

On a donc, en utilisant $c \in [0, t]$:

$$A(t) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}) \cdot t^2 + \mathrm{o}(t^2).$$

En écrivant :

$$A(t) = h(t) - h(0) \quad \text{avec} \quad h(s) := f(\mathbf{a} + t\mathbf{e}_i + s\mathbf{e}_j) - f(\mathbf{a} + s\mathbf{e}_j),$$

on obtient de manière similaire :

$$A(t) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a}) \cdot t^2 + \mathrm{o}(t^2).$$

Par conséquent :

$$A(t) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}) \cdot t^2 + \mathrm{o}(t^2) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a}) \cdot t^2 + \mathrm{o}(t^2).$$

On obtient l'égalité recherchée en divisant par t^2 et en faisant tendre t vers 0. ∎

Exercice 9.

Étudier la validité du théorème de Schwarz sur l'exemple suivant dû à Péano :

$$f : \mathbb{R}^2 \to \mathbb{R} \text{ définie par } f(x,y) := \begin{cases} 0 & \text{si } (x,y) = (0,0) \\ xy\dfrac{x^2 - y^2}{x^2 + y^2} & \text{si } (x,y) \neq (0,0). \end{cases}$$

Solution. La fonction f est de classe $\mathcal{C}^1$ sur $\mathbb{R}^2$ mais les dérivées partielles ne sont pas différentiables en $(0,0)$. On a :

$$\frac{\partial f}{\partial x}(x,y) = \frac{4x^2y^3 + x^4y - y^5}{(x^2+y^2)^2} \quad \text{et} \quad \frac{\partial f}{\partial y}(x,y) = \frac{x^5 - 4x^3y^2 - xy^4}{(x^2+y^2)^2}$$

quand $(x,y) \neq (0,0)$ et $\dfrac{\partial f}{\partial x}(0,0) = \dfrac{\partial f}{\partial y}(0,0) = 0$. Observons que :

$$\frac{\partial f}{\partial x}(0,y) = -y \quad \text{et} \quad \frac{\partial f}{\partial y}(x,0) = x.$$

Par conséquent :

$$\frac{\partial^2 f}{\partial y \partial x}(0,y) = -1 \quad \text{et} \quad \frac{\partial^2 f}{\partial x \partial y}(x,0) = 1.$$

En $(0,0)$ les dérivées croisées ne sont pas égales.

Il est facile de vérifier que si $f : U \to \mathbb{R}$ et $g : U \to \mathbb{R}$ sont deux fonctions de classe $\mathcal{C}^k$ sur un ouvert U de $\mathbb{R}^n$, la somme $f + g$ est elle-même de classe $\mathcal{C}^k$ et les dérivées partielles de la somme sont égales à la somme des dérivées partielles :

$$\forall \alpha \in \mathbb{N}^n \text{ tel que } |\alpha| \leqslant k, \quad \frac{\partial^{|\alpha|}(f+g)}{\partial \mathbf{x}^\alpha} = \frac{\partial^{|\alpha|} f}{\partial \mathbf{x}^\alpha} + \frac{\partial^{|\alpha|} f}{\partial \mathbf{x}^\alpha}.$$

En ce qui concerne le produit $f \cdot g$, la formule de Leibniz ci-dessous garantit que le produit est de classe $\mathcal{C}^k$ et exprime les dérivées partielles du produit en fonction des produits des dérivées partielles. Nous renvoyons le lecteur à la proposition 11 de la page 327 pour la démonstration.

Proposition 15 (Formule de Leibniz). Soit $f : U \to \mathbb{R}$ et $g : U \to \mathbb{R}$ deux fonction de classe $\mathcal{C}^k$ sur un ouvert U de $\mathbb{R}^n$. Alors, $f \cdot g$ est de classe $\mathcal{C}^k$ et :

$$\forall \alpha \in \mathbb{N}^n \text{ tel que } |\alpha| \leqslant k, \quad \frac{1}{\alpha!}\frac{\partial^{|\alpha|}(f \cdot g)}{\partial \mathbf{x}^\alpha} = \sum_{\beta+\gamma=\alpha} \left(\frac{1}{\beta!}\frac{\partial^{|\beta|} f}{\partial \mathbf{x}^\beta}\right) \cdot \left(\frac{1}{\gamma!}\frac{\partial^{|\gamma|} f}{\partial \mathbf{x}^\gamma}\right).$$

Pour finir, nous introduisons la notion d'application de classe $\mathcal{C}^k$. Comme dans le cas des fonctions, la définition se fait par récurrence.

Définition 11. Une application $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ est *de classe* $\mathcal{C}^k$ sur U si $\mathbf{f}$ est de classe $\mathcal{C}^1$ sur U et si les dérivées partielles $\dfrac{\partial \mathbf{f}}{\partial x_j}$ sont de classe $\mathcal{C}^{k-1}$.

Le lecteur vérifiera facilement qu'il est équivalent de demander que les coordonnées de $\mathbf{f}$ soient des fonctions de classe $\mathcal{C}^k$.

Définition 12. Un *difféomorphisme de classe* $\mathcal{C}^k$ est un difféomorphisme $\mathbf{f} : U \to V$ de classe $\mathcal{C}^1$ tel que $\mathbf{f}$ et $\mathbf{f}^{-1}$ soient de classe $\mathcal{C}^k$.

On verra (exercice 10 de la page suivante) qu'il suffit de demander que $\mathbf{f}$ soit de classe $\mathcal{C}^k$.

Théorème 16. Supposons $1 \leqslant k \leqslant +\infty$.

1. Si $\mathbf{f} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ et $\mathbf{g} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ sont de classe $\mathcal{C}^k$, alors $\mathbf{f} + \mathbf{g}$ est de classe $\mathcal{C}^k$.
2. Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ et $\mathbf{g} : U \subset \mathbb{R}^n \to \mathbb{R}^m$ sont de classe $\mathcal{C}^k$, alors $f\mathbf{g}$ est de classe $\mathcal{C}^k$.
3. Si $\mathbf{f} : U \subset \mathbb{R}^n \to V \subset \mathbb{R}^m$ et $\mathbf{g} : V \subset \mathbb{R}^m \to \mathbb{R}^p$ sont des applications de classe $\mathcal{C}^k$ alors $\mathbf{g} \circ \mathbf{f} : U \to \mathbb{R}^p$ est de classe $\mathcal{C}^k$.
4. Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est de classe $\mathcal{C}^k$ et ne s'annule pas, alors $1/f$ est de classe $\mathcal{C}^k$.

Démonstration. Les démonstrations se font par récurrence sur k.

Pour $k := 1$, elles sont toutes vraies : nous les avons démontrées dans [L1] pour les fonctions de classe $\mathcal{C}^1$. Supposons qu'elles soient vraies à l'ordre $k-1$ avec $k \geqslant 2$.

1. Si $\mathbf{f}$ et $\mathbf{g}$ sont de classe $\mathcal{C}^k$, alors $\mathbf{f} + \mathbf{g}$ est de classe $\mathcal{C}^1$ et les dérivées partielles de $\mathbf{f}$ et $\mathbf{g}$ sont de classe $\mathcal{C}^{k-1}$. Leurs sommes, qui sont les dérivées partielles de $\mathbf{f} + \mathbf{g}$, sont donc de classe $\mathcal{C}^{k-1}$ par hypothèse de récurrence. Cela montre que $\mathbf{f} + \mathbf{g}$ est de classe $\mathcal{C}^k$.
2. Notons $(x_1, \ldots, x_n)$ les coordonnées des points de $\mathbb{R}^n$. Si f et $\mathbf{g}$ sont de classe $\mathcal{C}^k$ (en particulier de classe $\mathcal{C}^1$), alors les dérivées partielles $\dfrac{\partial f}{\partial x_j}$ et $\dfrac{\partial \mathbf{g}}{\partial x_j}$ sont de

classe $\mathcal{C}^{k-1}$. Par hypothèse de récurrence, les dérivées partielles :

$$\frac{\partial f\mathbf{g}}{\partial x_j} = \frac{\partial f}{\partial x_j} \cdot \mathbf{g} + f \cdot \frac{\partial \mathbf{g}}{\partial x_j}$$

sont de classe $\mathcal{C}^{k-1}$, et donc $f\mathbf{g}$ est de classe $\mathcal{C}^k$.

3. Notons $(x_1, \dots, x_n)$ les coordonnées des points de $\mathbb{R}^n$, $(y_1, \dots, y_m)$ les coordonnées des points de $\mathbb{R}^m$ et $\mathbf{f} = (f_1, \dots, f_m)$ les coordonnées de $\mathbf{f}$. Si les applications $\mathbf{f}$ et $\mathbf{g}$ sont de classe $\mathcal{C}^k$, alors elles sont de classe $\mathcal{C}^1$. Donc $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ est aussi de classe $\mathcal{C}^1$. La règle de la chaîne donne :

$$\forall j \in [\![1, n]\!] \quad \frac{\partial \mathbf{h}}{\partial x_j} = \sum_{\ell=1}^{m} \frac{\partial f_\ell}{\partial x_j} \cdot \left(\frac{\partial \mathbf{g}}{\partial y_\ell} \circ \mathbf{f} \right).$$

Or, l'application $\mathbf{f}$ et les fonctions $\dfrac{\partial \mathbf{g}}{\partial y_\ell}$ sont de classe $\mathcal{C}^{k-1}$. Donc les fonctions $\dfrac{\partial \mathbf{g}}{\partial y_\ell} \circ \mathbf{f}$ le sont aussi par l'hypothèse de récurrence. Les produits $\dfrac{\partial f_\ell}{\partial x_j} \cdot \left(\dfrac{\partial \mathbf{g}}{\partial y_\ell} \circ \mathbf{f} \right)$ sont donc encore de classe $\mathcal{C}^{k-1}$. Finalement, les dérivées partielles $\dfrac{\partial \mathbf{h}}{\partial x_j}$ sont de classe $\mathcal{C}^{k-1}$. Cela montre que $\mathbf{h}$ est de classe $\mathcal{C}^k$.

4. La composée d'une fonction $f : U \to \mathbb{R}^*$ de classe $\mathcal{C}^k$ et de la fonction $x \mapsto 1/x$ qui est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^*$ est de classe $\mathcal{C}^k$ sur U d'après le point précédent. ∎

Les deux exercices qui suivent montrent que si dans le théorème d'inversion locale et/ou dans le théorème des fonctions implicites, on suppose que les fonctions sont de classe $\mathcal{C}^k$, alors l'inverse local et/ou la fonction définie implicitement est de classe $\mathcal{C}^k$.

Exercice 10.

Soit $\mathbf{f} : U \subset \mathbb{R}^n \to V \subset \mathbb{R}^n$ un homéomorphisme et soit $k \geqslant 1$ un entier. Montrer que $\mathbf{f} : U \to V$ est un difféomorphisme de classe $\mathcal{C}^k$ si, et seulement si, $\mathbf{f}$ est une application de classe $\mathcal{C}^k$ dont la différentielle est inversible en tout point de U.

Solution. D'après le théorème d'inversion locale, la fonction $\mathbf{f}^{-1} : V \to U$ est de classe $\mathcal{C}^1$. Sa matrice Jacobienne est l'inverse de la matrice Jacobienne de $\mathbf{f}$. Les dérivées partielles de $\mathbf{f}^{-1}$ sont donc des fractions rationnelles en les dérivées partielles de $\mathbf{f}$, le dénominateur ne s'annulant pas. Ce sont donc des fonctions de classe $\mathcal{C}^{k-1}$. Cela montre que $\mathbf{f}^{-1}$ est de classe $\mathcal{C}^k$.

Exercice 11.

Supposons que $F : U \subset \mathbb{R}^{p+m} \to \mathbb{R}^m$ vérifie les hypothèses du théorème des fonctions implicites en un point $\mathbf{c} := (\mathbf{a}, \mathbf{b}) \in U$ et soit $\mathbf{f} : V \subset \mathbb{R}^p \to \mathbb{R}^m$ une fonction différentiable qui vérifie :

$$\forall \mathbf{x} \in V, \qquad \mathbf{F}\big(\mathbf{x}, \mathbf{f}(\mathbf{x})\big) = \mathbf{F}(\mathbf{a}, \mathbf{b}).$$

Montrer que $\mathbf{f}$ est de classe $\mathcal{C}^k$ au voisinage de $\mathbf{a}$.

Solution. Le théorème des fonctions implicites garantit que $\mathbf{f}$ est de classe $\mathcal{C}^1$. Comme dans l'exercice précédent, la proposition 5 de la page 708 implique que les dérivées partielles de $\mathbf{f}$ sont des fractions rationnelles en les dérivées partielles de $\mathbf{F}$, le dénominateur ne s'annulant pas. Ce sont donc des applications de classe $\mathcal{C}^{k-1}$, ce qui montre que $\mathbf{f}$ est de classe $\mathcal{C}^k$.

3.2 Les formules de Taylor

Venons-en maintenant à la généralisation de la notion de développement de Taylor. Le lecteur est fortement invité à consulter d'abord le module I.7 sur les polynômes à plusieurs variables. En particulier, nous reprenons les notations suivantes.
Si $\mathbf{x} := (x_1, \dots, x_n) \in \mathbb{R}^n$, et si $\alpha := (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$, alors :

$$\alpha! = \alpha_1! \cdots \alpha_n! \qquad \text{et} \qquad \mathbf{x}^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}.$$

Définition 13 (Polynômes de Taylor). Si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est une fonction de classe $\mathcal{C}^k$ sur un ouvert U de $\mathbb{R}^n$, le polynôme de degré inférieur à k :

$$P^k_{f,\mathbf{a}}(\mathbf{x}) = \sum_{m=0}^{k} \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| = m}} \frac{1}{\alpha!} \frac{\partial^{|\alpha|} f}{\partial \mathbf{x}^\alpha}(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a})^\alpha$$

est le $k^{\text{ème}}$ *polynôme de Taylor en* $\mathbf{a}$.

On parle aussi du polynôme de Taylor d'ordre k.

Exemple. Si f est une fonction de deux variables (x_1, x_2), si $\mathbf{a} := (a_1, a_2)$ et $\mathbf{h} := (h_1, h_2)$ on a :

$$\begin{aligned} P^2_{f,\mathbf{a}}(\mathbf{a}+\mathbf{h}) &= f(\mathbf{a}) + \frac{1}{1!0!}\frac{\partial f}{\partial x_1^1 \partial x_2^0}(\mathbf{a}) h_1^1 h_2^0 + \frac{1}{0!1!}\frac{\partial f}{\partial x_1^0 \partial x_2^1}(\mathbf{a}) h_1^0 h_2^1 + \\ &\quad + \frac{1}{2!0!}\frac{\partial^2 f}{\partial x_1^2 \partial x_2^0}(\mathbf{a}) h_1^2 h_2^0 + \frac{1}{1!1!}\frac{\partial^2 f}{\partial x_1^1 \partial x_2^1}(\mathbf{a}) h_1^1 h_2^1 + \frac{1}{0!2!}\frac{\partial^2 f}{\partial x_1^0 \partial x_2^2}(\mathbf{a}) h_1^0 h_2^2 \\ &= f(\mathbf{a}) + \frac{\partial f}{\partial x_1}(\mathbf{a}) h_1 + \frac{\partial f}{\partial x_2}(\mathbf{a}) h_2 \\ &\quad + \frac{1}{2}\frac{\partial^2 f}{\partial x_1^2}(\mathbf{a}) h_1^2 + \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{a}) h_1 h_2 + \frac{1}{2}\frac{\partial^2 f}{\partial x_2^2}(\mathbf{a}) h_2^2. \end{aligned}$$

Exercice 12.

Déterminer le deuxième polynôme de Taylor de la fonction $f : (x, y) \mapsto \exp(x + y^2)$ en $(0, 0)$.

Solution. La fonction f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^2$ comme produit de deux fonctions de classe $\mathcal{C}^\infty$: $(x, y) \mapsto \exp(x)$ et $(x, y) \mapsto \exp(y^2)$. On a $f(0, 0) = \exp(0 + 0^2) = 1$. De plus :

$$\frac{\partial f}{\partial x}(x, y) = \exp(x + y^2) \quad \text{et} \quad \frac{\partial f}{\partial y}(x, y) = 2y \exp(x + y^2).$$

Donc :

$$\frac{\partial f}{\partial x}(0,0) = 1 \quad \text{et} \quad \frac{\partial f}{\partial y}(0,0) = 0.$$

Enfin :

$$\frac{\partial^2 f}{\partial x^2}(x,y) = \exp(x+y^2), \quad \frac{\partial^2 f}{\partial x \partial y}(x,y) = 2y\exp(x+y^2)$$

et

$$\frac{\partial^2 f}{\partial y^2}(x,y) = 2\exp(x+y^2) + 4y^2\exp(x+y^2).$$

Par conséquent :

$$\frac{\partial^2 f}{\partial x^2}(0,0) = 1, \quad \frac{\partial^2 f}{\partial x \partial y}(0,0) = 0 \quad \text{et} \quad \frac{\partial^2 f}{\partial y^2}(0,0) = 2.$$

Le deuxième polynôme de Taylor de f en $(0,0)$ est donc :

$$P^2_{f,(0,0)}(x,y) = 1 + x + \frac{1}{2}x^2 + y^2.$$

Nous sommes maintenant en mesure d'énoncer et de démontrer les formules de Taylor pour les fonctions de plusieurs variables. Nous commençons par la formule de Taylor avec reste intégral.

Théorème 17 (Formule de Taylor avec reste intégral).
Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction de classe $\mathcal{C}^k$ sur U, $\mathbf{a}$ un point de U et $\mathbf{h}$ un vecteur de $\mathbb{R}^n$ tel que le segment $[\mathbf{a}, \mathbf{a}+\mathbf{h}]$ soit contenu dans U. Alors :

$$f(\mathbf{a}+\mathbf{h}) = P^{k-1}_{f,\mathbf{a}}(\mathbf{a}+\mathbf{h}) + \int_0^1 k(1-t)^{k-1} \left(\sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha|=k}} \frac{1}{\alpha!} \frac{\partial^k f}{\partial \mathbf{x}^\alpha}(\mathbf{a}+t\mathbf{h}) \cdot \mathbf{h}^\alpha \right) \mathrm{d}t.$$

Démonstration. Considérons la fonction d'une variable réelle g définie au voisinage de 0 par :

$$g(t) := f(\mathbf{a}+t\mathbf{h}).$$

Lemme 18. Pour tout $m \in [\![1,k]\!]$, la fonction g est de classe $\mathcal{C}^m$ au voisinage de 0 et :

$$\frac{g^{(m)}(t)}{m!} = \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha|=m}} \frac{1}{\alpha!} \frac{\partial^m f}{\partial \mathbf{x}^\alpha}(\mathbf{a}+t\mathbf{h}) \cdot \mathbf{h}^\alpha.$$

Démonstration. Nous allons montrer ce résultat par récurrence sur $m \in [\![1,k]\!]$.
Pour $m = 1$, le résultat est évident car lorsque t tend vers t_0, on a :

$$f(\mathbf{a}+t\mathbf{h}) = f(\mathbf{a}+t_0\mathbf{h}) + (t-t_0)D_{\mathbf{a}+t_0\mathbf{h}}f(\mathbf{h}) + \mathrm{o}(t-t_0)$$

et donc :

$$\lim_{t\to t_0}\frac{g(t)-g(t_0)}{t-t_0}=D_{\mathbf{a}+t_0\mathbf{h}}f(\mathbf{h}).$$

La fonction g est donc dérivable au voisinage de 0 et :

$$g'(t)=D_{\mathbf{a}+t\mathbf{h}}f(\mathbf{h})=\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(\mathbf{a}+t\mathbf{h})\cdot h_j.$$

Les dérivées partielles de f étant continues, la dérivée g' est continue.

Supposons maintenant que l'hypothèse de récurrence soit vraie pour un certain $m<k$. Alors $g^{(m)}$ est une combinaison linéaire de fonctions de classe $\mathcal{C}^1$ (les dérivées d'ordre m de f évaluées en $\mathbf{a}+t\mathbf{h}$). D'après ce que nous venons de voir, si l'on note $E_j=(0,\dots,0,1,0,\dots,0)$, le 1 étant en $j^{\text{ème}}$ position, on a :

$$\frac{g^{(m+1)}(t)}{(m+1)!}=\frac{1}{m+1}\sum_{j=1}^{n}\sum_{\substack{\alpha\in\mathbb{N}^n\\|\alpha|=m}}\frac{1}{\alpha!}\frac{\partial^{m+1}f}{\partial\mathbf{x}^{\alpha+E_j}}(\mathbf{a}+t\mathbf{h})\cdot\mathbf{h}^{\alpha}\cdot h_j.$$

Pour chaque $\beta\in\mathbb{N}^n$ de longueur $|\beta|=p+1$, le terme :

$$\frac{\partial^{m+1}f}{\partial\mathbf{x}^{\beta}}(\mathbf{a}+t\mathbf{h})\cdot\mathbf{h}^{\beta}$$

apparaît dans la somme chaque fois que $\beta=\alpha+E_j$ pour un certain $j\in[\![1,n]\!]$. Il apparaît alors avec un coefficient :

$$\frac{1}{\alpha!}=\frac{\beta_j}{\beta!}.$$

Comme la somme des β_j vaut $p+1$, on a donc :

$$\frac{g^{(m+1)}(t)}{(m+1)!}=\frac{1}{m+1}\sum_{\substack{\beta\in\mathbb{N}^n\\|\beta|=m+1}}(m+1)\frac{\partial^{m+1}f}{\partial\mathbf{x}^{\beta}}(\mathbf{a}+t\mathbf{h})\cdot\mathbf{h}^{\beta}=\sum_{\substack{\beta\in\mathbb{N}^n\\|\beta|=m+1}}\frac{\partial^{m+1}f}{\partial\mathbf{x}^{\beta}}(\mathbf{a}+t\mathbf{h})\cdot\mathbf{h}^{\beta},$$

ce qu'il fallait démontrer. ∎

D'après la formule de Taylor avec reste intégral pour les fonctions d'une variable réelle (module IV.8 de [L1]), on a :

$$f(\mathbf{a}+\mathbf{h})=g(1)=\sum_{m=0}^{k-1}\frac{g^{(m)}(0)}{m!}+\int_0^1 k(1-t)^{k-1}\frac{g^{(k)}(t)}{k!}\,\mathrm{d}t,$$

ce qui donne le résultat recherché en utilisant le lemme précédent. ∎

De même que dans le cas des fonctions d'une variable réelle, on peut déduire des variantes de la formule de Taylor avec reste intégral. En utilisant la formule de la moyenne (*cf.* proposition 29 de la page 569) dans le reste intégral, on obtient par exemple le résultat suivant.

Théorème 19 (Formule de Taylor-Lagrange). Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction de classe $\mathcal{C}^k$ sur U, $\mathbf{a}$ et $\mathbf{b}$ deux points de U tels que le segment $[\mathbf{a}, \mathbf{b}]$ soit contenu dans U. Alors :

$$\exists \mathbf{c} \in]\mathbf{a}, \mathbf{b}[\,, \quad f(\mathbf{b}) = P_{f,\mathbf{a}}^{k-1}(\mathbf{a}+\mathbf{h}) + \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha|=k}} \frac{1}{\alpha!} \frac{\partial^k f}{\partial \mathbf{x}^\alpha}(\mathbf{c}) \cdot \mathbf{h}^\alpha.$$

En utilisant la continuité des dérivées $k^{ème}$ de f, on obtient également la formule de Taylor-Young.

Théorème 20 (Formule de Taylor-Young). Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction de classe $\mathcal{C}^k$ sur U et $\mathbf{a}$ un point de U.

1. Le polynôme $P_{f,\mathbf{a}}^k$ est l'unique polynôme de degré inférieur à k qui a les mêmes dérivées partielles d'ordre inférieur à k en $\mathbf{a}$ que f.
2. Le polynôme $P_{f,\mathbf{a}}^k$ est l'unique polynôme de degré inférieur à k qui vérifie :
$$f(\mathbf{x}) = P_{f,\mathbf{a}}^k(\mathbf{x}) + \mathrm{o}\big(\|\mathbf{x}-\mathbf{a}\|^k\big)$$
quand $\mathbf{x}$ tend vers $\mathbf{a}$.

Démonstration.

1. D'après la formule de Taylor pour les polynômes de plusieurs variables (voir théorème 12 de la page 328) :
$$P_{f,\mathbf{a}}^k(\mathbf{a}+\mathbf{h}) = \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| \leqslant k}} \frac{1}{\alpha!} \frac{\partial^{|\alpha|} P_{f,\mathbf{a}}^k}{\partial \mathbf{x}^\alpha}(\mathbf{a}) \cdot \mathbf{h}^\alpha.$$
On en déduit que pour tout $\alpha \in \mathbb{N}^n$ tel que $|\alpha| \leqslant k$, on a :
$$\frac{\partial^{|\alpha|} P_{f,\mathbf{a}}^k}{\partial \mathbf{x}^\alpha}(\mathbf{a}) = \frac{\partial^{|\alpha|} f}{\partial \mathbf{x}^\alpha}(\mathbf{a}).$$
Si Q était un autre polynôme de degré inférieur à k ayant les mêmes dérivées partielles d'ordre inférieur à k en $\mathbf{a}$ que f, alors $Q - P_{f,\mathbf{a}}^k$ serait un polynôme de degré inférieur à k dont toutes les dérivées partielles d'ordre inférieur à k en $\mathbf{a}$ seraient nulles. En appliquant de nouveau la formule de Taylor pour les polynômes de plusieurs variables, on obtiendrait :
$$\forall \mathbf{h} \in \mathbb{R}^n, \quad P_{f,\mathbf{a}}^k(\mathbf{a}+\mathbf{h}) - Q(\mathbf{a}+\mathbf{h}) = 0.$$
Donc, $P_{f,\mathbf{a}}^k = Q$. Cela montre l'unicité.
2. La formule de Taylor-Young découle de la formule de Taylor-Lagrange comme dans le cas des fonctions d'une variable. L'unicité s'obtient alors comme suit. Si Q est un polynôme de degré inférieur à k tel que :
$$f(\mathbf{x}) = Q(\mathbf{x}) + \mathrm{o}\big(\|\mathbf{x}-\mathbf{a}\|^k\big)$$

quand $\mathbf{x}$ tend vers $\mathbf{a}$. Alors,

$$R(\mathbf{h}) := P^k_{f,\mathbf{a}}(\mathbf{a}+\mathbf{h}) - Q(\mathbf{a}+\mathbf{h}) = \mathrm{o}\big(\|\mathbf{h}\|^k\big).$$

Le polynôme en R est un polynôme de degré au plus k. Supposons que R ne soit pas le polynôme nul. Soit $\omega \leqslant k$ sa valuation, soit $S := \mathrm{In}(R)$ sa forme initiale, c'est-à-dire sa composante homogène de degré ω et soit $\mathbf{h}_0 \in \mathbb{R}^n$ un vecteur tel que $S(\mathbf{h}_0) \neq 0$. Alors, quand $t \in \mathbb{R}$ tend vers 0, on a :

$$R(t\mathbf{h}_0) = S(t\mathbf{h}_0) + (R-S)(t\mathbf{h}_0) = t^\omega S(\mathbf{h}_0) + \mathrm{O}(t^{\omega+1}).$$

En effet, $R-S$ est un polynôme de valuation supérieure à $\omega+1$ et $t^{\omega+1}$ se factorise dans $(R-S)(t\mathbf{h}_0)$. On obtient une contradiction :

$$\omega \leqslant k, \quad S(\mathbf{h}_0) \neq 0 \quad \text{et} \quad t^\omega S(\mathbf{h}_0) + \mathrm{O}(t^{\omega+1}) = \mathrm{o}(\|t^k \mathbf{h}_0\|) = \mathrm{o}(t^k). \qquad \blacksquare$$

3.3 Classification des points critiques d'une fonction

Nous allons maintenant utiliser le développement de Taylor à l'ordre 2 pour étudier les extrema locaux éventuels d'une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$. Nous avons déjà expliqué que si $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est une fonction de classe $\mathcal{C}^1$ et si f a un extremum local en $\mathbf{a} \in U$, alors $\mathbf{a}$ est un point critique de f (la différentielle $D_\mathbf{a} f$ est l'application nulle). Cela donne une condition nécessaire d'existence d'un extremum local. Nous souhaitons maintenant présenter des conditions suffisantes.

Pour cela, nous allons travailler sous l'hypothèse que $f : U \subset \mathbb{R}^n \to \mathbb{R}$ est une fonction de classe $\mathcal{C}^2$ sur U.

Proposition 21. Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une application de classe $\mathcal{C}^2$ sur un ouvert U de $\mathbb{R}^n$ et soit $\mathbf{a}$ un point de U. La partie homogène d'ordre 2 du polynôme de Taylor $P^2_{f,\mathbf{a}}$ est une forme quadratique sur $T_\mathbf{a}\mathbb{R}^n$. Plus précisément :

$$P^2_{f,\mathbf{a}}(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + D_\mathbf{a} f(\mathbf{h}) + \frac{1}{2} Q_\mathbf{a} f(\mathbf{h})$$

avec $D_\mathbf{a} f$ une forme linéaire sur $T_\mathbf{a}\mathbb{R}^n$ et $Q_\mathbf{a} f$ une forme quadratique sur $T_\mathbf{a}\mathbb{R}^n$. La matrice de $Q_\mathbf{a} f$ dans la base canonique de $T_\mathbf{a}\mathbb{R}^n$ est la matrice symétrique :

$$\begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(\mathbf{a}) & \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{a}) \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{a}) & \dfrac{\partial^2 f}{\partial x_2^2}(\mathbf{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{a}) \\ \vdots & & & \vdots \\ \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{a}) & \dfrac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(\mathbf{a}) \end{pmatrix}.$$

Démonstration. La partie homogène de degré 2 de $P^2_{f,\mathbf{a}}(\mathbf{a}+\mathbf{h})$ est :

$$Q_\mathbf{a} f(\mathbf{h}) := \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha|=2}} \frac{1}{\alpha!} \frac{\partial^{|\alpha|} f}{\partial \mathbf{x}^\alpha}(\mathbf{a}) \cdot \mathbf{h}^\alpha.$$

C'est un polynôme homogène de degré 2. On peut l'écrire sous la forme :

$$Q_{\mathbf{a}}f(\mathbf{h}) = \frac{1}{2}\left(\sum_{j=1}^{n}\frac{\partial^2 f}{\partial x_j^2}(\mathbf{a})\cdot h_j^2 + 2\sum_{1\leqslant i<j\leqslant n}\frac{\partial^2 f}{\partial x_i\partial x_j}(\mathbf{a})\cdot h_i h_j\right).$$

La forme polaire associée est :

$$\varphi(\mathbf{h},\mathbf{h}') := \frac{1}{2}\left(\sum_{j=1}^{n}\frac{\partial^2 f}{\partial x_j^2}(\mathbf{a})\cdot h_j h_j' + \sum_{1\leqslant i<j\leqslant n}\frac{\partial^2 f}{\partial x_i\partial x_j}(\mathbf{a})\cdot (h_i h_j' + h_j h_i')\right). \qquad \blacksquare$$

Définition 14. Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une application de classe $\mathcal{C}^2$ sur un ouvert U de $\mathbb{R}^n$ et soit $\mathbf{a}$ un point de U. La forme quadratique $Q_{\mathbf{a}}f$ est la *dérivée seconde* de f en $\mathbf{a}$. La matrice :

$$H_{\mathbf{a}}f = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(\mathbf{a}) & \dfrac{\partial^2 f}{\partial x_1\partial x_2}(\mathbf{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_1\partial x_n}(\mathbf{a}) \\ \dfrac{\partial^2 f}{\partial x_1\partial x_2}(\mathbf{a}) & \dfrac{\partial^2 f}{\partial x_2^2}(\mathbf{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_2\partial x_n}(\mathbf{a}) \\ \vdots & & & \vdots \\ \dfrac{\partial^2 f}{\partial x_1\partial x_n}(\mathbf{a}) & \dfrac{\partial^2 f}{\partial x_2\partial x_n}(\mathbf{a}) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(\mathbf{a}) \end{pmatrix}$$

est la Hessienne de f en $\mathbf{a}$.

La formule de Taylor-Young à l'ordre 2 s'écrit :

$$f(\mathbf{x}) = f(\mathbf{a}) + D_{\mathbf{a}}f(\mathbf{x}-\mathbf{a}) + \frac{1}{2}Q_{\mathbf{a}}f(\mathbf{x}-\mathbf{a}) + \mathrm{o}\big(\|\mathbf{x}-\mathbf{a}\|^2\big).$$

Si $\mathbf{a}\in U$ est un point critique de $\mathbf{f}$, le terme $D_{\mathbf{a}}f(\mathbf{x}-\mathbf{a})$ est nul, et :

$$f(\mathbf{x}) - f(\mathbf{a}) = \frac{1}{2}Q_{\mathbf{a}}f(\mathbf{x}-\mathbf{a}) + \mathrm{o}\big(\|\mathbf{x}-\mathbf{a}\|^2\big).$$

Pour savoir si f admet un extremum local en $\mathbf{a}$, on est donc amené à étudier le signe de $Q_{\mathbf{a}}f(\mathbf{x}-\mathbf{a})$, c'est-à-dire, le signe d'une forme quadratique.

Théorème 22 (Condition nécessaire d'existence d'un extremum local). Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une application de classe $\mathcal{C}^2$ sur un ouvert U de $\mathbb{R}^n$ et soit $\mathbf{a}\in U$ un point critique de f. Si f admet en $\mathbf{a}$ un minimum local, alors la forme quadratique $Q_{\mathbf{a}}f$ est positive. Si f admet en $\mathbf{a}$ un maximum local, alors la forme quadratique $Q_{\mathbf{a}}f$ est négative.

Démonstration. Supposons que f admette un minimum local en $\mathbf{a}$. Alors, pour tout $\mathbf{h}\in T_{\mathbf{a}}\mathbb{R}^n$, si $t\in\mathbb{R}$ est suffisamment petit, on a :

$$f(\mathbf{a}+t\mathbf{h}) - f(\mathbf{a}) \geqslant 0.$$

Or :

$$f(\mathbf{a}+t\mathbf{h}) - f(\mathbf{a}) = \frac{t^2}{2}Q_{\mathbf{a}}f(\mathbf{h}) + \mathrm{o}(t^2).$$

Cela montre que $Q_{\mathbf{a}}f(\mathbf{h}) \geqslant 0$, d'où la positivité de $Q_{\mathbf{a}}f$.
En appliquant le résultat à $-f$ on traite le cas d'un maximum local en $\mathbf{a}$.

Définition 15. Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ sur un ouvert U de $\mathbb{R}^n$ et $\mathbf{a} \in U$ un point critique de f. La fonction f a un *point selle* (ou un *col*) en $\mathbf{a}$ si la signature (définition 17 de la page 132) de la forme quadratique $Q_{\mathbf{a}}f$ est (p, q) avec $p > 0$ et $q > 0$.

Corollaire 23. Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une application de classe $\mathcal{C}^2$ sur un ouvert U de $\mathbb{R}^n$ qui admet un point selle en $\mathbf{a} \in U$. Alors, f n'admet pas d'extremum local en $\mathbf{a}$.

Le résultat suivant n'est pas tout à fait une réciproque du théorème 22 de la page précédente puisque l'on requiert que la forme quadratique $Q_{\mathbf{a}}f$ soit *définie* positive ou négative.

Théorème 24 (Existence d'un extremum local). Soit $f : U \subset \mathbb{R}^n \to \mathbb{R}$ une application de classe $\mathcal{C}^2$ sur un ouvert U de $\mathbb{R}^n$ et soit $\mathbf{a} \in U$ un point critique de f. Si la forme quadratique $Q_{\mathbf{a}}f$ est définie positive, alors f admet un minimum local en $\mathbf{a}$. Si la forme quadratique $Q_{\mathbf{a}}f$ est définie négative, alors f admet un maximum local en $\mathbf{a}$.

Démonstration. Nous allons traiter le cas où $Q_{\mathbf{a}}f$ est définie positive. L'autre cas se traite en appliquant le résultat à $-f$. La matrice Hessienne $H_{\mathbf{a}}f$ est symétrique et comme $Q_{\mathbf{a}}f$ est définie positive, ses valeurs propres sont toutes strictement positives. On note $c > 0$ la plus petite valeur propre de la matrice Hessienne. On a (corollaire 57 de la page 184) :

$$\inf_{\|h\|=1} Q_{\mathbf{a}}f(\mathbf{h}) = c.$$

Or :

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = \frac{1}{2}Q_{\mathbf{a}}f(\mathbf{h}) + \|h\|^2\varepsilon(\mathbf{h}) \quad \text{avec} \quad \varepsilon(\mathbf{h}) \underset{\mathbf{h}\to\mathbf{0}}{\longrightarrow} 0.$$

Par conséquent, il existe $r > 0$ tel que si $\|\mathbf{h}\| < r$, alors $\big|\varepsilon(\mathbf{h})\big| < c/2$. Dans ce cas :

$$\frac{1}{2}Q_{\mathbf{a}}f(\mathbf{h}) + \|h\|^2\varepsilon(\mathbf{h}) = \|h\|^2\left(\frac{1}{2}Q_{\mathbf{a}}f\left(\frac{\mathbf{h}}{\|\mathbf{h}\|}\right) + \varepsilon(\mathbf{h})\right) > 0.$$

Donc, si $0 < |\mathbf{x}-\mathbf{a}| < r$, on a $f(\mathbf{x}) > f(\mathbf{a})$ et f admet un minimum local en $\mathbf{a}$.

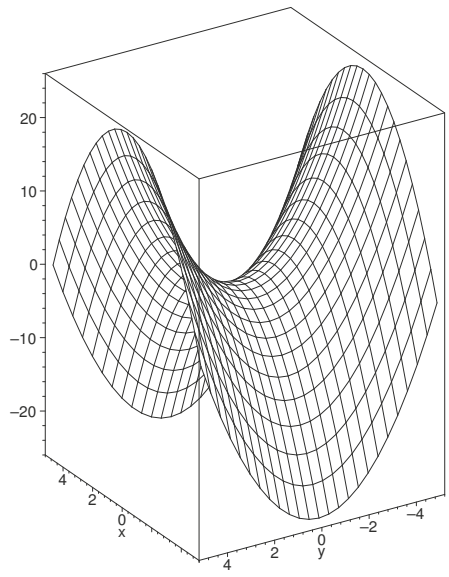
La difficulté principale consiste alors à rechercher si la forme quadratique $Q_{\mathbf{a}}f$ est ou n'est pas (définie) positive. En dimension 2, cela est relativement simple. La matrice Hessienne a deux valeurs propres. Le produit des valeurs propres est (strictement) positif si, et seulement si, les deux valeurs propres sont (non nulles et) de même signe. Dans ce cas, les valeurs propres sont positives si, et seulement si, la trace de la matrice est positive. Elles sont négatives si, et seulement si, la trace de la matrice est négative.
Dans le cas d'une fonction de deux variables (x, y), on utilise volontiers les notations de Monge :

$$p := \frac{\partial f}{\partial x},\ q := \frac{\partial f}{\partial y},\ r := \frac{\partial^2 f}{\partial x^2},\ s := \frac{\partial^2 f}{\partial x \partial y} \text{ et } t := \frac{\partial^2 f}{\partial y^2}.$$

Le théorème d'existence d'un extremum local s'écrit alors sous la forme suivante.

Théorème 25. Soit $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ sur un ouvert U de $\mathbb{R}^2$ et $(a,b) \in U$ un point où $p = q = 0$ (on utilise les notations de Monge). Alors :
- si $rs - t^2 < 0$, il y a un point selle en (a,b) ; en particulier, f n'admet pas d'extremum local en (a,b),
- si $rs - t^2 > 0$ et si $r + t > 0$, on a un minimum local en (a,b),
- si $rs - t^2 > 0$ et si $r + t < 0$, on a un maximum local en (a,b).

Exemple. La fonction $f : \mathbb{R}^2 \to \mathbb{R}^2$ définie par $f(x,y) := y^2 - x^2$ a un point selle en $(0,0)$. En effet, c'est une forme quadratique de signature $(1,1)$.



Pour la recherche pratique des extrema d'une fonction de deux variables, on procède de la manière suivante :

1. Sur un ouvert, on utilise le théorème 25.
2. Sur un fermé $\overline{U}$, on étudie ce qui se passe à l'intérieur et on fait une étude spéciale à la frontière ; les extrema se trouvent là où les conditions du théorème 25 sont satisfaites, et éventuellement sur la frontière de U.

Exercice 13.

Chercher les extrema de la fonction $f(x,y) := x^2 + xy + y^2$ sur le triangle fermé T de sommets $A = (0,0)$, $B = (1,0)$ et $C = (0,1)$.

Solution. Le triangle T est compact. La fonction f est continue sur T ; elle admet donc un minimum et un maximum. Supposons dans un premier temps qu'il y ait un extremum dans le triangle ouvert. Alors, la différentielle de f doit s'y annuler. On doit donc avoir :

$$0 = \frac{\partial f}{\partial x}(x,y) = 2x + y \quad \text{et} \quad 0 = \frac{\partial f}{\partial y}(x,y) = x + 2y.$$

Le déterminant de ce système système est $2 \cdot 2 - 1 \cdot 1 = 3 \neq 0$. Par conséquent, le système est inversible et il n'y a qu'une solution $(x,y) = (0,0)$. Ce point n'appartient pas au triangle ouvert. La fonction f atteint donc ses bornes sur la frontière du triangle.

Si l'on se restreint au segment $[AB]$ pour lequel $y = 0$, la fonction prend la forme $f(x,0) = x^2$ qui est croissante par rapport à x. Le minimum de f sur le segment $[AB]$ est atteint en A et vaut 0. Le maximum est atteint en B et vaut 1.

Une étude similaire menée sur le segment $[AC]$ montre que le minimum de f sur le segment $[AC]$ est atteint en A et vaut 0. Le maximum est atteint en C et vaut 1.

Enfin, si le point (x, y) appartient au segment $[BC]$, on a $y = 1 - x$ et :

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

Le maximum est atteint pour $x = \pm 1$ (c'est-à-dire en B et C) et vaut 1. Le minimum est atteint en $x = 1/2$ et vaut $3/4$.

En fin de compte, on voit que le minimum de f sur le triangle fermé est atteint en A et vaut 0. Le maximum est atteint en B et C et vaut 1.

4 Calcul différentiel

Nous allons maintenant étudier comment les notions de différentiation et d'intégration sont reliées. Pour les fonctions d'une variable réelle, ces deux notions sont intimement liées à travers le théorème fondamental de l'analyse qui dit que si f est une fonction de classe $\mathcal{C}^1$ sur un intervalle $[a, b]$, alors f' est intégrable au sens de Riemann sur $[a, b]$ (voir la section 6.4 du module II.3) et :

$$\int_a^b f'(x)\,\mathrm{d}x = f(b) - f(a).$$

Il existe une généralisation de ce résultat à des fonctions de n variables : le théorème de Stokes. L'énoncé général de ce théorème concerne les « formes différentielles ». Le théorème de Stokes est énoncé et démontré dans [MPA1].

Ici, nous allons étudier la version du théorème de Stokes pour les fonctions de 2 variables réelles. Dans ce contexte, le théorème de Stokes prend le nom de « formule de Green-Riemann ».

4.1 Intégrales curvilignes

Nous allons donner une première définition de ce qu'est une forme différentielle de degré 1 en deux variables. Nous reviendrons sur cette définition plus loin (section 4.4) quand nous nous serons habitués aux formes différentielles.

Définition 16 (1-forme). Une *forme différentielle* de degré 1 sur un ouvert $U \subset \mathbb{R}^2$ est une expression de la forme :

$$\omega := P(x, y)\,\mathrm{d}x + Q(x, y)\,\mathrm{d}y$$

avec $P : U \to \mathbb{R}$ et $Q : U \to \mathbb{R}$ des fonctions. La forme différentielle ω est continue (respectivement de classe $\mathcal{C}^1$) si P et Q sont continues (respectivement de classe $\mathcal{C}^1$).

Une forme différentielle de degré 1 sera également appelée une 1-forme. Nous renvoyons le lecteur à la partie 4.4 pour l'explication des notations $\mathrm{d}x$ et $\mathrm{d}y$.

Exemple. Si $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ est une fonction différentiable la *différentielle totale* de f est la 1-forme :

$$\mathrm{d}f := \frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y.$$

Si f est de classe $\mathcal{C}^1$, la 1-forme $\mathrm{d}f$ est continue. Les 1-formes qui sont des différentielles totales de fonctions joueront un rôle particulier tout au long de cette partie.

Notons la cohérence dans les notations : si f est la fonction définie par $f(x, y) := x$, alors $\mathrm{d}f = \mathrm{d}x$ et si g est la fonction définie par $g(x, y) := y$, alors $\mathrm{d}g = \mathrm{d}y$.

Définition 17. Une 1-forme ω est *exacte* sur un ouvert $U \subset \mathbb{R}^2$ si $\omega = \mathrm{d}f$ avec $f : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ sur U.

Exercice 14.

Montrer que la 1-forme $\omega := y\,\mathrm{d}x$ n'est exacte sur aucun ouvert de $\mathbb{R}^2$.

Démonstration. Supposons que $\omega = \mathrm{d}f$ avec $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ une fonction différentiable sur U. Alors :

$$\frac{\partial f}{\partial x} = y \quad \text{et} \quad \frac{\partial f}{\partial y} = 0.$$

Choisissons un point $(x_0, y_0) \in U$ et considérons r suffisamment petit pour que le carré $]x_0 - r, x_0 + r[\times]y_0 - r, y_0 + r[$ soit contenu dans U.

Étant donné que $\dfrac{\partial f}{\partial y}$ est identiquement nulle, la fonction f ne dépend pas de y. Plus précisément, si la fonction $g :]x_0 - r, x_0 + r[\to \mathbb{R}$ est définie par $g(x) := f(x, y_0)$, alors pour tout $(x, y) \in]x_0 - r, x_0 + r[\times]y_0 - r, y_0 + r[$, on a $f(x, y) = g(x)$. Alors, $\dfrac{\partial f}{\partial x}(x, y) = g'(x)$ ne dépend que de x et ne peut pas être égale à y. ∎

Remarque. Nous verrons plus loin un moyen plus court de conclure : la 1-forme ω n'est pas *fermée*.

Une 1-forme est un objet mathématique qu'il est possible d'intégrer le long de courbes paramétrées. Nous renvoyons le lecteur au module III.2 de [L1] pour la notion de courbe paramétrée.

Définition 18. Une *courbe paramétrée* continue (respectivement de classe $\mathcal{C}^1$) est une application $\gamma : I \to \mathbb{R}^2$ continue (respectivement de classe $\mathcal{C}^1$) avec $I \subset \mathbb{R}$ un intervalle d'intérieur non vide. On dit qu'une courbe continue $\gamma : [a, b] \to \mathbb{R}$ est $\mathcal{C}^1$ par morceaux s'il existe une subdivision $a = a_1 < a_2 < \cdots < a_N < a_{N+1} = b$ telle que pour tout $i \in [\![1, N]\!]$, la restriction de γ à l'intervalle $[a_i, a_{i+1}]$ soit de classe $\mathcal{C}^1$ (il y a une dérivée à droite en a_i et une dérivée à gauche en a_{i+1}). La courbe est *fermée* si $\gamma(a) = \gamma(b)$ et *simple* si la restriction de γ à $]a, b[$ est injective.

Définition 19 (Intégrale curviligne). Soit $\omega : U \subset \mathbb{R}^2$ une 1-forme continue sur un ouvert U de $\mathbb{R}^2$ et soit $\gamma : [a, b] \to U$ une courbe paramétrée de classe $\mathcal{C}^1$. Alors l'intégrale $\displaystyle\int_\gamma \omega$ de ω le long de γ est définie de la manière suivante. Si $\omega := P(x, y)\,\mathrm{d}x + Q(x, y)\,\mathrm{d}y$ et si $\gamma(t) := \big(x(t), y(t)\big)$, on a :

$$\int_\gamma \omega := \int_a^b \big(P\big(x(t), y(t)\big)x'(t) + Q\big(x(t), y(t)\big)y'(t)\big)\,\mathrm{d}t.$$

Exemples.

1. L'intégrale de la 1-forme différentielle $\omega := y\,\mathrm{d}x$ le long du graphe d'une fonction $f : [a,b] \to \mathbb{R}$ de classe $\mathcal{C}^1$ est égale à $\displaystyle\int_a^b f(t)\,\mathrm{d}t$.
2. L'intégrale de la 1-forme $x\,\mathrm{d}y - y\,\mathrm{d}x$ la long de la courbe $\gamma : [0, 2\pi] \to \mathbb{R}^2$ définie par $\gamma(t) = (\cos t, \sin t)$ est égale à :

$$\int_\gamma \omega = \int_0^{2\pi} \cos t \cos t - \sin t \cdot (-\sin t)\,\mathrm{d}t = \int_0^{2\pi} \mathrm{d}t = 2\pi.$$

Le résultat qui suit établit que l'intégrale d'une 1-forme le long d'une courbe paramétrée ne dépend pas véritablement du paramétrage. Il ne dépend que de la *courbe géométrique orientée* (nous renvoyons le lecteur au module III.2 de [L1] pour les notions de courbes géométriques et de courbes géométriques orientées).

Théorème 26. Soit ω une 1-forme continue sur un ouvert U de $\mathbb{R}^2$.
Soient $\gamma_1 : [a_1, b_1] \to U$ et $\gamma_2 : [a_2, b_2] \to U$ deux courbes paramétrées de classe $\mathcal{C}^1$.
Supposons que $\gamma_1 = \gamma_2 \circ h$ avec $h : [a_1, b_1] \to [a_2, b_2]$ un difféomorphisme croissant de classe $\mathcal{C}^1$. Alors :

$$\int_{\gamma_1} \omega = \int_{\gamma_2} \omega.$$

Démonstration.
Supposons que $\omega := P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y$, que $\gamma_1(u) := \big(x_1(u), y_1(u)\big)$ et que $\gamma_2(t) := \big(x_2(t), y_2(t)\big)$. Alors, $x_1 = x_2 \circ h$ et $y_1 = y_2 \circ h$. En faisant le changement de variable $t = h(u)$, on obtient :

$$\begin{aligned}\int_{a_2}^{b_2} P\big(x_2(t), y_2(t)\big) x_2'(t)\,\mathrm{d}t &= \int_{h^{-1}(a_2)}^{h^{-1}(b_2)} P\big(x_2 \circ h(u), y_2 \circ h(u)\big) \underbrace{x_2' \circ h(u) \cdot h'(u)}_{(x_2 \circ h)'(u)}\,\mathrm{d}u \\ &= \int_{a_1}^{b_1} P\big(x_1(u), y_1(u)\big) x_1'(u)\,\mathrm{d}u\end{aligned}$$

puisque $h^{-1}(a_2) = a_1$ et $h^{-1}(b_2) = b_1$. De même :

$$\begin{aligned}\int_{a_2}^{b_2} Q\big(x_2(t), y_2(t)\big) y_2'(t)\,\mathrm{d}t &= \int_{h^{-1}(a_2)}^{h^{-1}(b_2)} Q\big(x_2 \circ h(u), y_2 \circ h(u)\big) \underbrace{y_2' \circ h(u) \cdot h'(u)}_{(y_2 \circ h)'(u)}\,\mathrm{d}u \\ &= \int_{a_1}^{b_1} Q\big(x_1(u), y_1(u)\big) y_1'(u)\,\mathrm{d}u.\end{aligned}$$

On obtient le résultat requis en ajoutant les deux égalités. ∎

Si l'on change d'orientation, c'est-à-dire si h est décroissant, on change le signe de l'intégrale. En effet, on a alors $h^{-1}(a_2) = b_1$ et $h^{-1}(b_2) = a_1$. Autrement dit, le

signe de l'intégrale d'une 1-forme le long d'une courbe géométrique dépend du sens de parcours.

Si ω est une 1-forme continue sur un ouvert U de $\mathbb{R}^2$, si $\gamma : [a,b] \to U$ est une courbe paramétrée de classe $\mathcal{C}^1$ et si $c \in [a,b]$, la relation de Chasles montre que :

$$\int_\gamma \omega = \int_{\gamma_1} \omega + \int_{\gamma_2} \omega$$

avec $\gamma_1 : [a,c] \to U$ et $\gamma_2 : [c,b] \to U$ les restrictions de γ à $[a,c]$ et $[c,b]$.
Soit $\gamma : [a,b] \to U$ une courbe de classe $\mathcal{C}^1$ par morceaux.
Si $a = a_1 < a_2 < \cdots < a_N < a_{N+1} = b$ est une subdivision telle que la restriction γ_i de γ à chaque intervalle $[a_i, a_{i+1}]$, $i \in [\![1, N]\!]$, soit de classe $\mathcal{C}^1$, on peut donc définir l'intégrale de ω le long de γ, de façon indépendante de la subdivision, par :

$$\int_\gamma \omega := \sum_{i=1}^{N} \int_{\gamma_i} \omega.$$

Le résultat suivant est une généralisation du théorème fondamental du calcul différentiel et intégral. Il en est également une conséquence directe.

Théorème 27. Si $\omega := \mathrm{d}f$ est une 1-forme exacte sur un ouvert U de $\mathbb{R}^2$ et si $\gamma : [a,b] \to U$ est une courbe de classe $\mathcal{C}^1$ par morceaux, alors :

$$\int_\gamma \omega = f\big(\gamma(b)\big) - f\big(\gamma(a)\big).$$

Démonstration. Soit $a = a_1 < a_2 < \cdots < a_N < a_{N+1} = b$ une subdivision adaptée à γ telle que la restriction γ_i de γ à chaque intervalle $[a_i, a_{i+1}]$, $i \in [\![1, N]\!]$, soit de classe $\mathcal{C}^1$. Nous allons montrer que le théorème est vrai pour les courbes de classe $\mathcal{C}^1$. On en déduit alors que :

$$\begin{aligned}\int_\gamma \omega = \sum_{i=1}^{N} \int_{\gamma_i} \omega &= \sum_{i=1}^{N} f\big(\gamma(a_{i+1})\big) - f\big(\gamma(a_i)\big) \\ &= f\big(\gamma(a_{N+1})\big) - f\big(\gamma(a_0)\big) = f\big(\gamma(b)\big) - f\big(\gamma(a)\big).\end{aligned}$$

Supposons donc à partir de maintenant que γ est un chemin de classe $\mathcal{C}^1$ et démontrons le théorème dans ce cas. La fonction $g : [a,b] \to \mathbb{R}$ définie par $g(t) := f\big(\gamma(t)\big)$ est de classe $\mathcal{C}^1$ sur $[a,b]$. De plus :

$$g'(t) = \frac{\partial f}{\partial x}\big(x(t), y(t)\big)x'(t) + \frac{\partial f}{\partial y}\big(x(t), y(t)\big)y'(t).$$

Par conséquent :

$$\int_\gamma \omega = \int_a^b \frac{\partial f}{\partial x}\big(x(t), y(t)\big)x'(t) + \frac{\partial f}{\partial y}\big(x(t), y(t)\big)y'(t)\,\mathrm{d}t = \int_a^b g'(t)\,\mathrm{d}t = g(b) - g(a). \qquad \blacksquare$$

Corollaire 28. Si $\omega := \mathrm{d}f$ est une 1-forme exacte sur un ouvert U de $\mathbb{R}^2$ et si $\gamma : [a,b] \to U$ est une courbe fermée de classe $\mathcal{C}^1$ par morceaux, alors $\displaystyle\int_\gamma \omega = 0$.

Démonstration. La courbe $\gamma : [a,b] \to \mathbb{R}^2$ est fermée si, et seulement si, $\gamma(a) = \gamma(b)$. On a alors $f\big(\gamma(b)\big) - f\big(\gamma(a)\big) = 0$.

Exercice 15.

Montrer que la 1-forme ω définie sur $\mathbb{R}^2 \setminus \big\{(0,0)\big\}$ par $\omega := \dfrac{y\,\mathrm{d}x - x\,\mathrm{d}y}{x^2+y^2}$ n'est pas exacte sur $\mathbb{R}^2 \setminus \big\{(0,0)\big\}$.

Démonstration. L'intégrale de ω sur les cercles centrés en 0 n'est pas nulle.
Plus précisément, si $\gamma : [0,2\pi] \to \mathbb{R}^2 \setminus \big\{(0,0)\big\}$ est la courbe fermée définie par $\gamma(t) = \big(r\cos(t), r\sin(t)\big)$, on a :

$$\int_\gamma \omega = \int_0^{2\pi} \frac{r^2\cos^2(t) + r^2\sin^2(t)}{r^2}\,\mathrm{d}t = 2\pi.$$

Par conséquent, ω n'est pas exacte sur $\mathbb{R}^2 \setminus \big\{(0,0)\big\}$. En revanche, sa restriction à tout disque ouvert de $\mathbb{R}^2 \setminus \big\{(0,0)\big\}$, et même à tout ouvert où l'on peut définir $\arctan(x/y)$ est exacte.

4.2 Dérivée d'une forme différentielle de degré 1

À partir de maintenant, nous n'allons considérer que des 1-formes qui sont de classe $\mathcal{C}^1$.

Définition 20. Soit $\omega := P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y$ une 1-forme de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbb{R}^2$. La *dérivée* de ω est :

$$\mathrm{d}\omega := \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y)\right)\mathrm{d}x\,\mathrm{d}y.$$

Une 1-forme ω telle que $\mathrm{d}\omega = 0$ est dite *fermée*.

La nature du symbole $\mathrm{d}x\,\mathrm{d}y$ sera expliquée à la section 4.4.

Exercice 16.

Montrer que la 1-forme ω définie sur $\mathbb{R}^2 \setminus \big\{(0,0)\big\}$ par $\omega := \dfrac{y\,\mathrm{d}x - x\,\mathrm{d}y}{x^2+y^2}$ est fermée.

Solution. En effet :

$$\begin{aligned}\mathrm{d}\omega &= \left(\frac{\partial}{\partial x}\left(-\frac{x}{x^2+y^2}\right) - \frac{\partial}{\partial y}\left(\frac{y}{x^2+y^2}\right)\right)\mathrm{d}x\,\mathrm{d}y \\ &= \left(-\frac{x^2+y^2-2x^2}{(x^2+y^2)^2} - \frac{x^2+y^2-2y^2}{(x^2+y^2)^2}\right)\mathrm{d}x\,\mathrm{d}y = 0.\end{aligned}$$

Notons que l'application $\omega \mapsto \mathrm{d}\omega$ est une application linéaire : si ω_1 et ω_2 sont des 1-formes de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbb{R}^2$ et si $\lambda \in \mathbb{R}$, alors $\mathrm{d}(\omega_1 + \lambda\omega_2) = \mathrm{d}\omega_1 + \lambda\,\mathrm{d}\omega_2$. Le résultat suivant permet dans certains cas de détecter qu'une forme différentielle n'est pas exacte.

Théorème 29. Supposons qu'une 1-forme ω de classe $\mathcal{C}^1$ soit exacte sur un ouvert $U \subset \mathbb{R}^2$. Alors, ω est fermée.

Démonstration. Supposons que $\omega = \mathrm{d}f$ avec $f : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$. Alors :

$$\omega = P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y \quad \text{avec} \quad P := \frac{\partial f}{\partial x} \quad \text{et} \quad Q := \frac{\partial f}{\partial y}.$$

Comme ω est de classe $\mathcal{C}^1$, la fonction f est de classe $\mathcal{C}^2$. Le théorème de Schwarz (théorème 14 de la page 731) implique que :

$$\frac{\partial P}{\partial y} = \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial Q}{\partial x}.$$

Par conséquent :

$$\mathrm{d}\omega = \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y)\right) \mathrm{d}x\,\mathrm{d}y = 0. \qquad \blacksquare$$

Exemple. On peut en déduire que la 1-forme $\omega = y\,\mathrm{d}x$ n'est exacte sur aucun ouvert de $\mathbb{R}^2$. En effet, $\mathrm{d}\omega = -\,\mathrm{d}x\,\mathrm{d}y \neq 0$. Par conséquent, ω n'est fermée sur aucun ouvert de $\mathbb{R}^2$. Elle n'est donc exacte sur aucun ouvert de $\mathbb{R}^2$.

Le théorème précédent affirme qu'une 1-forme exacte est fermée. La réciproque n'est pas vraie. La 1-forme ω définie sur $\mathbb{R}^2 \setminus \{(0,0)\}$ par $\omega := \dfrac{y\,\mathrm{d}x - x\,\mathrm{d}y}{x^2+y^2}$ est fermée (exercice 16 de la page précédente) mais n'est pas exacte (exercice 15 de la page précédente). Le théorème suivant, dû au mathématicien Henri Poincaré, exprime que sur un ouvert étoilé, ce problème ne se présente pas.

Définition 21. Un ouvert $U \subset \mathbb{R}^2$ est *étoilé* s'il existe un point $\mathbf{a} \in U$ tel que pour tout point $\mathbf{z} \in U$, le segment $[\mathbf{a}, \mathbf{z}]$ soit contenu dans U. On dit alors que U est étoilé par rapport à $\mathbf{a}$.

Exemple. Un ouvert convexe $U \subset \mathbb{R}^2$ est étoilé par rapport à n'importe quel point $\mathbf{a} \in U$.

Théorème 30 (Poincaré). Si $U \subset \mathbb{R}^2$ est un ouvert étoilé et si ω est une 1-forme de classe $\mathcal{C}^1$ sur U, alors, ω est exacte si, et seulement si, ω est fermée.

Remarque. Ce théorème est en fait vrai sous l'hypothèses beaucoup plus générale que U est « simplement connexe ». Voir [MPA1].

Démonstration. Il nous faut montrer que si ω est fermée, alors elle est exacte (l'autre implication étant toujours vraie). Supposons donc que U soit étoilé par rapport à $(a,b) \in U$ et que ω soit une 1-forme fermée sur U. Définissons une fonction $f : U \to \mathbb{R}$ par :

$$f(x,y) := \int_\gamma \omega \quad \text{avec} \quad \gamma : t \in [0,1] \mapsto \big(a + t(x-a), b + t(y-b)\big)$$

un paramétrage du segment reliant (a,b) à (x,y). Nous allons montrer que $\mathrm{d}f = \omega$. D'abord :

$$f(x,y) = \int_0^1 P\big(\gamma(t)\big) \cdot (x-a) + Q\big(\gamma(t)\big) \cdot (y-b)\,\mathrm{d}t.$$

Ensuite, la fonction $(x,t) \mapsto P\big(\gamma(t)\big) \cdot (x-a) + Q\big(\gamma(t)\big) \cdot (y-b)$ est continue par rapport à x et t. Elle admet une dérivée partielle par rapport à x :

$$P\big(\gamma(t)\big) + \left(\frac{\partial P}{\partial x}\big(\gamma(t)\big) \cdot (x-a) + \frac{\partial Q}{\partial x}\big(\gamma(t)\big) \cdot (y-b)\right) \cdot t$$

qui est elle-même continue par rapport à x et t. On peut donc dériver sous le signe intégrale (corollaire 131 de la page 646), ce qui donne :

$$\frac{\partial f}{\partial x}(x,y) = \int_0^1 P\big(\gamma(t)\big)\,\mathrm{d}t + \int_0^1 \left(\frac{\partial P}{\partial x}\big(\gamma(t)\big) \cdot (x-a) + \frac{\partial Q}{\partial x}\big(\gamma(t)\big) \cdot (y-b)\right) \cdot t\,\mathrm{d}t.$$

Pour $t \in [0,1]$, posons $g(t) := P\big(\gamma(t)\big)$ et observons que :

$$g'(t) = \frac{\partial P}{\partial x}\big(\gamma(t)\big) \cdot (x-a) + \frac{\partial P}{\partial y}\big(\gamma(t)\big) \cdot (y-b) = \frac{\partial P}{\partial x}\big(\gamma(t)\big) \cdot (x-a) + \frac{\partial Q}{\partial x}\big(\gamma(t)\big) \cdot (y-b).$$

Nous venons d'utiliser l'hypothèse que ω est fermée. Par conséquent :

$$\frac{\partial f}{\partial x}(x,y) = \int_0^1 g(t)\,\mathrm{d}t + \int_0^1 t g'(t)\,\mathrm{d}t = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\big(t g(t)\big)\,\mathrm{d}t = g(1) = P(x,y).$$

Un argument similaire montre que :

$$\frac{\partial f}{\partial y}(x,y) = Q(x,y)$$

et donc $\mathrm{d}f = P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y = \omega$. ∎

Corollaire 31. Soit $U \subset \mathbb{R}^2$ un ouvert étoilé et ω une 1-forme de classe $\mathcal{C}^1$ sur U. Si ω est fermée, alors pour tout courbe fermée $\gamma : [a,b] \to U$, on a $\displaystyle\int_\gamma \omega = 0$.

Démonstration. Si ω est fermée sur U, elle y est exacte.

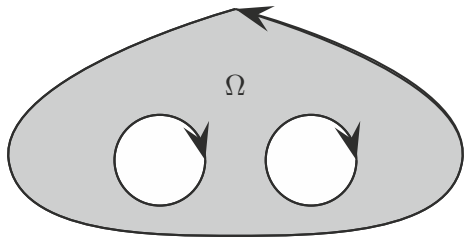
Le résultat se déduit du théorème 29 de la page précédente. ∎

4.3 Formule de Green-Riemann

Nous allons maintenant présenter la formule de Green-Riemann. Nous allons l'énoncer dans un cadre assez général, mais nous ne la démontrons que dans un cas particulier. La démonstration du cas général est donnée dans [MPA1], comme cas particulier du théorème de Stokes.

Pour énoncer la formule de Green-Riemann, commençons par définir quelques objets. Nous dirons qu'un ouvert $\Omega \subset \mathbb{R}^2$ a un bord $\mathcal{C}^1$ par morceaux si Ω est borné et si la frontière de Ω est une réunion finie d'ensembles Γ_i, $i \in [\![1, N]\!]$, deux à deux disjoints, chaque Γ_i étant le support d'une courbe fermée simple γ_i de classe $\mathcal{C}^1$ par morceaux. On note $\partial\Omega$ la frontière de Ω.

Quitte à remplacer $\gamma_i(t)$ par $\gamma_i(-t)$ si nécessaire, on peut supposer que lorsque t croît, $\gamma_i(t)$ se déplace le long de la courbe Γ_i en laissant la région Ω à gauche. On dit que le bord de Ω est orienté de sorte que Ω soit à sa gauche.



Si U est un ouvert contenant l'adhérence de Ω et si ω est une 1-forme continue sur $\cup$, on définit alors :

$$\int_{\partial\Omega} \omega := \sum_{i=1}^{N} \int_{\gamma_i} \omega.$$

Si de plus ω est de classe $\mathcal{C}^1$ sur U et si $\omega := P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y$, on définit :

$$\int_{\Omega} \mathrm{d}\omega := \int_{\Omega} \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y) \right) \mathrm{d}x\,\mathrm{d}y.$$

L'ouvert Ω étant borné et la fonction $\dfrac{\partial Q}{\partial x}(x,y) - \dfrac{\partial P}{\partial y}(x,y)$ étant continue sur U, l'intégrale précédente est bien définie.

Théorème 32 (Formule de Green-Riemann – admise dans ce tome).
Soit Ω un ouvert de $\mathbb{R}^2$ ayant un bord $\mathcal{C}^1$ par morceaux. Soit ω une 1-forme de classe $\mathcal{C}^1$ sur un ouvert U contenant $\overline{\Omega}$. Supposons que le bord de Ω soit orienté de sorte que Ω soit à sa gauche. Alors :

$$\int_{\Omega} \mathrm{d}\omega = \int_{\partial\Omega} \omega.$$

Nous allons démontrer cette formule dans un cas particulier. Auparavant, mentionnons l'une des nombreuses applications de la formule de Green-Riemann : le calcul de l'aire d'une région délimitée par une courbe de classe $\mathcal{C}^1$.

Proposition 33. Soit $\gamma : [a,b] \to \mathbb{R}^2$ une courbe fermée simple de classe $\mathcal{C}^1$ et soit Ω la région délimitée par γ. On suppose que lorsque t croît, le point $\gamma(t)$ parcourt le bord de Ω en laissant Ω à gauche. Alors, l'aire de Ω est donnée par :

$$\mathrm{Aire}(\Omega) = \int_a^b x(t)\cdot y'(t)\,\mathrm{d}t = -\int_a^b y(t)\cdot x'(t)\,\mathrm{d}t \quad \text{avec} \quad \gamma(t) := \big(x(t), y(t)\big).$$

Démonstration. L'aire de Ω est donnée par :

$$\mathrm{Aire}(\Omega) := \int_\Omega \mathrm{d}x\,\mathrm{d}y = \int_\Omega \mathrm{d}(x\,\mathrm{d}y) = \int_\Omega \mathrm{d}(-y\,\mathrm{d}x).$$

Il suffit donc d'appliquer la formule de Green-Riemann qui donne :

$$\mathrm{Aire}(\Omega) = \int_{\partial\Omega} x\,\mathrm{d}y = -\int_{\partial\Omega} y\,\mathrm{d}x.$$

Remarque. On a également la formule : $\mathrm{Aire}(\Omega) = \displaystyle\int_a^b \frac{xy' - yx'}{2}$, obtenue en prenant la moyenne des deux précédentes.

Nous allons maintenant démontrer la formule de Green-Riemann dans un cas particulier. Le cas que nous allons traiter est le cas d'un ouvert horizontalement et verticalement simple.

Définition 22. Un ouvert $\Omega \subset \mathbb{R}^2$ est *verticalement simple* si :

$$\Omega := \big\{(x,y) \in \mathbb{R}^2 \;\big|\; a < x < b,\ \varphi_1(x) < y < \varphi_2(x)\big\}$$

où φ_1 et φ_2 sont des fonctions $\mathcal{C}^1$ par morceaux sur $[a,b]$ vérifiant $\varphi_1 < \varphi_2$ sur $]a,b[$.
Un ouvert $\Omega \subset \mathbb{R}^2$ est *horizontalement simple* si :

$$\Omega := \big\{(x,y) \in \mathbb{R}^2 \;\big|\; c < y < d,\ \psi_1(y) < x < \psi_2(y)\big\}$$

où ψ_1 et ψ_2 sont des fonctions $\mathcal{C}^1$ par morceaux sur $[c,d]$ vérifiant $\psi_1 < \psi_2$ sur $]c,d[$.

Exemple. Un ouvert convexe Ω dont le bord est une courbe de classe $\mathcal{C}^1$ par morceaux est à la fois horizontalement et verticalement simple. En effet, la convexité de Ω implique que l'intersection de Ω avec la droite verticale d'abscisse x est soit vide, soit un intervalle $]\varphi_1(x), \varphi_2(x)[$, le second cas se produisant pour x appartenant à un intervalle $]a,b[$. Cela montre que Ω est verticalement simple. On montre de même que Ω est horizontalement simple.

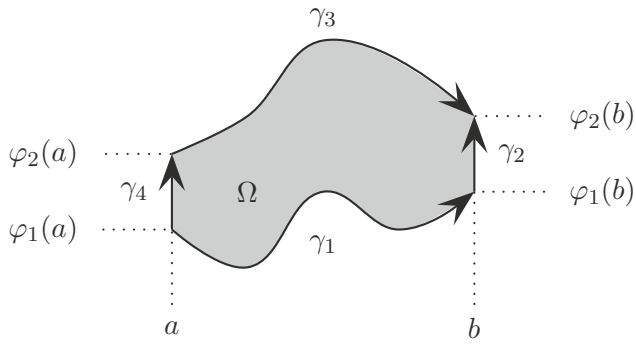
Lemme 34. Soit $\Omega := \big\{(x,y) \in \mathbb{R}^2 \;\big|\; a < x < b,\ \varphi_1(x) < y < \varphi_2(x)\big\}$ un ouvert verticalement simple et $P(x,y)\,\mathrm{d}x$ une 1-forme continue sur un ouvert U contenant $\overline{\Omega}$. Alors :

$$\int_{\partial\Omega} P(x,y)\,\mathrm{d}x = \int_a^b P\big(t,\varphi_1(t)\big) - P\big(t,\varphi_2(t)\big)\,\mathrm{d}t.$$

Démonstration. On peut paramétrer la frontière de Ω à l'aide des quatre courbes :

- $\gamma_1 : [a,b] \to \mathbb{R}^2$ définie par $\gamma_1(t) = \big(t, \varphi_1(t)\big)$,

- $\gamma_2 : \big[\varphi_1(b), \psi_2(b)\big] \to \mathbb{R}^2$ définie par $\gamma_2(t) = (b, t)$,
- $\gamma_3 : [a, b] \to \mathbb{R}^2$ définie par $\gamma_3(t) = \big(t, \varphi_2(t)\big)$ et
- $\gamma_4 : \big[\varphi_1(a), \varphi_2(a)\big] \to \mathbb{R}^2$ définie par $\gamma_4(t) = (a, t)$.



Si ω est une 1-forme continue sur un ouvert U contenant $\overline{\Omega}$, on a alors, en tenant compte de l'orientation sur le bord de Ω :

$$\int_{\partial\Omega} \omega := \int_{\gamma_1} \omega + \int_{\gamma_2} \omega - \int_{\gamma_3} \omega - \int_{\gamma_4} \omega.$$

Par conséquent :

$$\begin{aligned}\int_{\partial\Omega} P(x,y)\,\mathrm{d}x &= \int_a^b P\big(t, \varphi_1(t)\big) \cdot 1\,\mathrm{d}t + \int_{\varphi_1(b)}^{\varphi_2(b)} P\big(b, t\big) \cdot 0\,\mathrm{d}t \\ &\quad - \int_a^b P\big(t, \varphi_2(t)\big) \cdot 1\,\mathrm{d}t - \int_{\varphi_1(a)}^{\varphi_2(a)} P\big(a, t\big) \cdot 0\,\mathrm{d}t \\ &= \int_a^b P\big(t, \varphi_1(t)\big)\,\mathrm{d}t - \int_a^b P\big(t, \varphi_2(t)\big).\end{aligned}$$

Lemme 35. Soit $\Omega := \big\{(x,y) \in \mathbb{R}^2 \;\big|\; c < y < d,\ \psi_1(y) < x < \psi_2(y)\big\}$ un ouvert horizontalement simple et $Q(x,y)\,\mathrm{d}y$ une 1-forme continue sur un ouvert U contenant $\overline{\Omega}$. Alors :

$$\int_{\partial\Omega} Q(x,y)\,\mathrm{d}y = \int_c^d -Q\big(\psi_1(t), t\big) + Q\big(\psi_2(t), t\big)\,\mathrm{d}t.$$

Démonstration. La démonstration est similaire à celle du lemme précédent. ∎

Proposition 36.
Supposons que Ω soit un ouvert simultanément horizontalement et verticalement simple. Alors :

$$\int_\Omega d\omega = \int_{\partial\Omega} \omega.$$

Démonstration. La démonstration repose sur le théorème de Fubini (voir le théorème 133 de la page 648). Si $\omega := P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y$, alors :

$$\int_\Omega d\omega = \int_\Omega \frac{\partial Q}{\partial x}(x,y)\,\mathrm{d}x\,\mathrm{d}y - \int_\Omega \frac{\partial P}{\partial y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Comme Ω est verticalement simple, on a :

$$\Omega := \big\{(x,y) \in \mathbb{R}^2 \;\big|\; a < x < b,\ \varphi_1(x) < y < \varphi_2(x)\big\}$$

avec φ_1 et φ_2 des fonctions de classe $\mathcal{C}^1$ par morceaux telles que $\varphi_1 < \varphi_2$ sur $]a,b[$. D'après le théorème de Fubini :

$$\begin{aligned}\int_\Omega \frac{\partial P}{\partial y}(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_a^b \left(\int_{\varphi_1(x)}^{\varphi_2(x)} \frac{\partial P}{\partial y}(x,y)\,\mathrm{d}y\right)\mathrm{d}x \\ &= \int_a^b P\big(x,\varphi_2(x)\big) - P\big(x,\varphi_1(x)\big)\,\mathrm{d}x = -\int_{\partial\Omega} P(x,y)\,\mathrm{d}x.\end{aligned}$$

On a utilisé le lemme 34 de la page 751 pour la dernière égalité.

De même, on montre à l'aide du théorème de Fubini et du lemme 35 de la page ci-contre que :

$$\int_\Omega \frac{\partial P}{\partial y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\partial\Omega} Q(x,y)\,\mathrm{d}y.$$

On a donc :

$$\int_\Omega d\omega = -\int_{\partial\Omega} P(x,y)\,\mathrm{d}x + \int_{\partial\Omega} Q(x,y)\,\mathrm{d}y = \int_{\partial\Omega} \omega. \qquad \blacksquare$$

Notons qu'à l'aide de cette proposition, on peut traiter un grand nombre de cas en découpant l'ouvert Ω en morceaux qui sont à la fois horizontalement et verticalement simples.

Exemple. Considérons le cas où l'ouvert Ω et la couronne comprise entre les cercles centrés à l'origine de rayons 1 et 2 :

$$\omega := \big\{(x,y) \in \mathbb{R}^2 \;\big|\; 1 < x^2 + y^2 < 4\big\}.$$

Le bord de Ω est orienté dans le sens trigonométrique en ce qui concerne le cercle de rayon 2 et dans le sens inverse en ce qui concerne le cercle de rayon 1.

On peut découper la couronne Ω en quatre morceaux horizontalement et verticalement simples à l'aide de deux segments verticaux et de deux segments horizontaux.

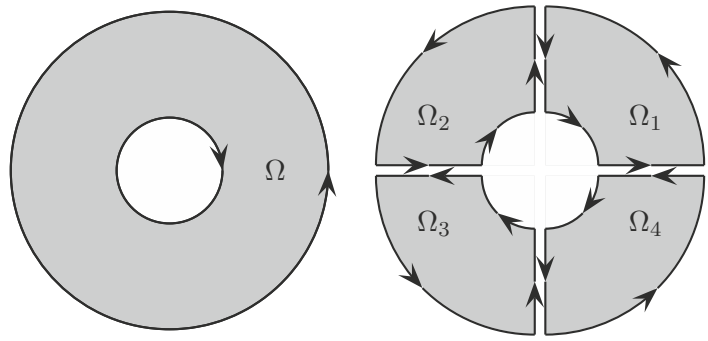


FIGURE II.5.1 – Découpage d'une couronne en quatre morceaux horizontalement et verticalement simples

On a alors :

$$\begin{aligned}\int_\Omega \mathrm{d}\omega &= \int_{\Omega_1} \mathrm{d}\omega + \int_{\Omega_2} \mathrm{d}\omega + \int_{\Omega_3} \mathrm{d}\omega + \int_{\Omega_4} \mathrm{d}\omega \\ &= \int_{\partial\Omega_1} \omega + \int_{\partial\Omega_2} \omega + \int_{\partial\Omega_3} \omega + \int_{\partial\Omega_4} \omega = \int_{\partial\Omega} \omega.\end{aligned}$$

Pour obtenir la dernière égalité, on utilise le fait que les intégrales sur les segments qui ont servi à découper Ω apparaissent chacune deux fois avec des signes opposés. En effet, chaque segment est dans le bord de deux ouvert Ω_i, chaque segment apparaît donc deux fois, mais avec des orientations opposées.

4.4 Quelques explications

Nous souhaitons apporter quelques éclaircissements sur les objets que nous venons de manipuler.

Une forme différentielle de degré 1 sur un ouvert $U \subset \mathbb{R}^2$ est une application de U à valeurs dans l'espace des formes linéaires sur $\mathbb{R}^2$. Plus précisément, c'est une application qui, à chaque point $\mathbf{a} \in U$, associe une forme linéaire $\omega(x,y) : T_{\mathbf{a}}\mathbb{R}^2 \to \mathbb{R}$, c'est-à-dire un élément de l'espace dual $(T_{\mathbf{a}}\mathbb{R}^2)^*$. On identifie $T_{\mathbf{a}}\mathbb{R}^2$ avec l'espace vectoriel $\mathbb{R}^2$ (en envoyant le couple $(a,b) \in T_{\mathbf{a}}\mathbb{R}^2$ sur le vecteur $\overrightarrow{ab} \in \mathbb{R}^2$). On identifie donc l'espace des formes linéaires sur $T_{\mathbf{a}}\mathbb{R}^2$ avec l'espace des formes linéaires sur $\mathbb{R}^2$.

On note $(\mathrm{d}x, \mathrm{d}y)$ la base canonique du dual de $\mathbb{R}^2$, habituellement notée (e_1^*, e_2^*). La forme différentielle $\mathrm{d}x$ est celle qui, à un vecteur de $\mathbb{R}^2$, associe sa première coordonnée (dans la base canonique) et la forme différentielle $\mathrm{d}y$ est celle qui lui associe sa seconde coordonnée. Une forme différentielle de degré 1 s'écrit alors $P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y$ avec $P(x,y)$ et $Q(x,y)$ les coordonnées de la forme linéaire $\omega(x,y)$ dans la base $(\mathrm{d}x, \mathrm{d}y)$.

Si $f : U \to \mathbb{R}$ est une application de classe $\mathcal{C}^1$, sa différentielle $D_{\mathbf{a}}f$ en un point $\mathbf{a} \in U$ est une forme linéaire sur $T_{\mathbf{a}}\mathbb{R}^2$. Ses coordonnées dans la base $(\mathrm{d}x, \mathrm{d}y)$ sont $\dfrac{\partial f}{\partial x}(\mathbf{a})$ et $\dfrac{\partial f}{\partial y}(\mathbf{a})$. La différentielle totale de f est donc l'application $\mathbf{a} \mapsto D_{\mathbf{a}}f$.

Une forme différentielle ω de degré 1 s'intègre le long de courbes paramétrées. Si $\gamma : [a,b] \to \mathbb{R}^2$ est une courbe paramétrée, ce que l'on note $\gamma'(t_0)$ est l'image de l'unique vecteur qui constitue la base canonique de $T_{t_0}\mathbb{R}$ par la différentielle $D_{t_0}\gamma$. C'est un vecteur de $T_{\gamma(t_0)}\mathbb{R}^2$ et on peut donc considérer son image par la forme linéaire $\omega\big(\gamma(t_0)\big)$. Si $\omega := P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y$ et si $\gamma(t) := \big(x(t), y(t)\big)$, cette image vaut $P\big(x(t),y(t)\big)x'(t) + Q\big(x(t),y(t)\big)y'(t)$. L'intégrale de ω le long de γ est donc égale à :

$$\int_\gamma \omega := \int_a^b \omega\big(\gamma(t)\big)\big(\gamma'(t)\big)\,\mathrm{d}t.$$

Décrivons maintenant ce que nous avons noté $R(x,y)\,\mathrm{d}x\,\mathrm{d}y$. Il s'agit en fait d'une forme de différentielle de degré 2 qu'il serait plus sage de noter $\mathrm{d}x \wedge \mathrm{d}y$ afin de distinguer $\mathrm{d}x \wedge \mathrm{d}y$ de $\mathrm{d}y \wedge \mathrm{d}x$, qui, comme nous le verrons, sont opposées. Mais qu'est-ce qu'une forme différentielle de degré 2 sur U ? Il s'agit d'une application de U à valeurs dans l'espace vectoriel des formes bilinéaires alternées sur $T_{\mathbf{a}}\mathbb{R}^2$ (que nous identifions avec l'espace vectoriel des formes bilinéaires alternées sur $\mathbb{R}^2$). Cet espace est un espace vectoriel de dimension 1. Une forme bilinéaire alternée sur $\mathbb{R}^2$ est un multiple de la forme déterminant dans la base canonique que l'on note $\mathrm{d}x \wedge \mathrm{d}y$.

En fait, la notation $\wedge$ a une autre raison d'être. Si l'on considère deux formes différentielles de degré 1 sur U, disons ω_1 et ω_2, on peut définir une forme bilinéaire alternée $\omega_1 \wedge \omega_2$, que l'on

appelle le produit extérieur de ω_1 et ω_2, par la relation :

$$(\omega_1 \wedge \omega_2)(\mathbf{h}, \mathbf{k}) = \omega_1(\mathbf{h})\omega_2(\mathbf{k}) - \omega_2(\mathbf{h})\omega_1(\mathbf{k}).$$

Le lecteur vérifiera facilement que le produit extérieur $\mathrm{d}x \wedge \mathrm{d}y$ de $\mathrm{d}x$ et $\mathrm{d}y$ donne bien la forme déterminant dans la base canonique. Notons également que $\mathrm{d}x \wedge \mathrm{d}x = \mathrm{d}y \wedge \mathrm{d}y = 0$.
On peut donc produire des différentielles de degré 2 en considérant le produit extérieur de deux différentielles de degré 1. On peut également le faire en dérivant une forme différentielle ω de degré 1. La forme $\mathrm{d}\omega$ s'appelle la dérivée extérieure de ω. Plutôt que de définir la dérivée extérieure de ω par une formule faisant intervenir les dérivées partielles des coordonnées de ω dans la base $(\mathrm{d}x, \mathrm{d}y)$, on peut en donner une définition qui est plus proche de la définition de la différentielle d'une fonction. Rappelons que si $f : U \subset \mathbb{R}^2 \to \mathbb{R}$ est une fonction et si $\mathbf{a}$ est un point de U, la valeur de la différentielle $\mathrm{d}f(\mathbf{a}) := D_{\mathbf{a}}f$ évaluée sur un vecteur $\mathbf{h} \in T_{\mathbf{a}}\mathbb{R}^2$ est donnée par la limite, si elle existe :

$$\mathrm{d}f(\mathbf{a})(\mathbf{h}) := D_{\mathbf{a}}(\mathbf{h}) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\big(f(\mathbf{a} + \varepsilon\mathbf{h}) - f(\mathbf{a})\big).$$

Pour définir la valeur de la dérivée extérieure d'une forme ω en un point $\mathbf{a}$, évaluée sur deux vecteurs $\mathbf{h}, \mathbf{k} \in T_{\mathbf{a}}\mathbb{R}^2$, on définit le parallélogramme $P_{\mathbf{a}}(\mathbf{h}, \mathbf{k})$ de la manière suivante :

$$P_{\mathbf{a}}(\mathbf{h}, \mathbf{k}) = \big\{\mathbf{a} + r\mathbf{h} + s\mathbf{k} \;\big|\; 0 < r, s < 1\big\}.$$

Notons que le bord du parallélogramme $P_{\mathbf{a}}(\mathbf{h}, \mathbf{k})$ est une courbe de classe $\mathcal{C}^1$ par morceaux et qu'il est possible d'intégrer ω le long de ce bord (en l'orientant de sorte que le parallélogramme soit à gauche de la courbe). La valeur de la forme bilinéaire symétrique $\mathrm{d}\omega(\mathbf{a})$ évaluée sur le couple de vecteurs $(\mathbf{h}, \mathbf{k})$ par la limite, si elle existe :

$$\mathrm{d}\omega(\mathbf{a})(\mathbf{h}, \mathbf{k})) := \lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} \int_{\partial P_{\mathbf{a}}(\mathbf{h},\mathbf{k})} \omega.$$

Même dans le cas où cette limite existe pour tous couples de vecteurs $(\mathbf{h}, \mathbf{k}) \in T_{\mathbf{a}}\mathbb{R}^2$, il n'est pas du tout évident que l'on définit ainsi une forme bilinéaire alternée. Il est cependant possible de montrer que c'est le cas et que :

$$\begin{aligned}
\mathrm{d}\big(P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y\big) &= \mathrm{d}P \wedge \mathrm{d}x + \mathrm{d}Q \wedge \mathrm{d}y \\
&= \left(\frac{\partial P}{\partial x}(x,y)\,\mathrm{d}x + \frac{\partial P}{\partial y}(x,y)\,\mathrm{d}y\right) \wedge \mathrm{d}x \\
&\quad + \left(\frac{\partial Q}{\partial x}(x,y)\,\mathrm{d}x + \frac{\partial Q}{\partial y}(x,y)\,\mathrm{d}y\right) \wedge \mathrm{d}y \\
&= \frac{\partial P}{\partial y}(x,y)\,\mathrm{d}y \wedge \mathrm{d}x + \frac{\partial Q}{\partial x}(x,y)\,\mathrm{d}x \wedge \mathrm{d}y \\
&= \left(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y)\right) \mathrm{d}x \wedge \mathrm{d}y.
\end{aligned}$$

Ce genre de calcul formel est justifié dans [MPA1]. Il devrait aider le lecteur à comprendre pourquoi il y a un signe $-$ devant $\dfrac{\partial P}{\partial y}$ et pourquoi les dérivées partielles de P et Q respectivement par rapport à x et à y n'apparaissent pas.

Exercice type corrigé

II.5.0 Soit $f : \mathbb{R}^2 \to \mathbb{R}$ la fonction définie par $f(x, y) := x + x^2 + xy + y^3$.

1. Déterminer les points critiques de f.
2. La fonction f admet-elle des exterma locaux dans $\mathbb{R}^2$?

Solution.

1. Les dérivées partielles de f sont :

$$\frac{\partial f}{\partial x}(x, y) = 1 + 2x + y \quad \text{et} \quad \frac{\partial f}{\partial y}(x, y) = x + 3y^2.$$

Le point (x, y) est un point critique de f si, et seulement si :

$$\begin{cases} 1 + 2x + y = 0 \\ x + 3y^2 = 0. \end{cases}$$

On voit donc que si (x, y) est un point critique de f, alors $x = -3y^2$, et donc, $1 - 6y^2 + y = 0$. Les racines de ce polynôme de degré 2 sont :

$$y_1 := \frac{1 + \sqrt{1 + 24}}{12} = \frac{1}{2} \quad \text{et} \quad y_2 := \frac{1 - \sqrt{1 + 24}}{12} = -\frac{1}{3}.$$

On a alors $x_1 = -3/4$ ou $x_2 = -1/3$. Les points critiques de f sont donc les points :

$$(-3/4, 1/2) \quad \text{et} \quad (-1/3, -1/3).$$

En effet, on vérifie facilement que ces points sont solutions du système.

2. Notons que f est un polynôme à deux variables, donc une fonction de classe $\mathcal{C}^2$. On a :

$$r := \frac{\partial^2 f}{\partial x^2}(x, y) = 2, \quad s := \frac{\partial^2 f}{\partial x \partial y}(x, y) = 1 \quad \text{et} \quad t := \frac{\partial^2 f}{\partial y^2}(x, y) = 6y.$$

On a $rt - s^2 = 12y - 1$.

Au point $(-3/4, 1/2)$, $rt - s^2 = 5 > 0$ et le point est un extremum local.

Comme $r + t = 5 > 0$, le point est un minimum local. La valeur de f en ce point est $-7/16$.

Notons que ce n'est pas un minimum global car $f(0, y) = y^3 \underset{y \to -\infty}{\longrightarrow} -\infty$.

Au point $(-1/3, -1/3)$, on a $rt - s^2 = -5 < 0$. Ce point est un col.

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 Parmi les fonctions suivantes, lesquelles ne sont pas localement inversibles avec un inverse différentiable ?

1. $\mathbf{f}(x,y,z) := (xyz, x^2)$ en $(0,0,0)$,
2. $\mathbf{f}(x,y) := (x^2y, -2x)$ en $(0,1)$
3. $\mathbf{f}(x,y) := (x^2+1-y, x)$ en $(2,1)$.

II.5.2
Montrer que l'application $f : \mathbb{R}^2 \to \mathbb{R}^2$ définie par $f(x,y) = \big(\sin(y/2) - x, \sin(x/2) - y\big)$ est localement inversible en tout point de $\mathbb{R}^2$.

II.5.3 En quels points le théorème d'inversion locale garantit-il que l'application :

$$f : (x,y) \mapsto \big(xye^x, ye^x\big)$$

admet un inverse local ?

II.5.4 Le système d'équations :

$$\begin{cases} x + y + \sin(xy) = 2a \\ \sin(x^2+y) = a^2 \end{cases}$$

admet-il une solution pour a suffisamment proche de 0 ?

II.5.5 Soit $F : \mathbb{R}^2 \to \mathbb{R}$ la fonction définie par $F(x,y) := x^3 + y^3 - 3xy$.
Montrer qu'au voisinage du point $(0,1)$, l'équation $F(x,y) = 1$ définit une fonction implicite $y = \varphi(x)$.
Calculer $\varphi'(0)$.

II.5.6 Soit $F(x,y) = x - y + 2\ln(x+y)$. Montrer que l'équation $F\big(x, g(x)\big) = 0$ et la condition $g(1/2) = 1/2$ définissent une fonction g dérivable au voisinage de $x = 1/2$.
Donner la dérivée de g en $1/2$.

II.5.7 Pour $c \in \mathbb{R}$, on considère le polynôme $P_c : x \mapsto x^2 + c$. Supposons que $\alpha_0 \in \mathbb{R}$ soit un point périodique de P_{c_0} avec $c_0 \in \mathbb{R}$, c'est-à-dire tel que :

$$\alpha_0 = P_{c_0}^{\circ k}(\alpha_0) := \underbrace{P_{c_0} \circ \cdots \circ P_{c_0}}_{k \text{ fois}}(\alpha_0).$$

Supposons de plus que $\big(P_{c_0}^{\circ k}\big)'(\alpha_0) \neq 1$. Montrer qu'il existe un intervalle ouvert I contenant c_0 et une fonction dérivable $\alpha : I \to \mathbb{R}$ telle que $\alpha(c_0) = \alpha_0$ et pour tout $c \in I$, $\alpha(c) = P_c^{\circ k}\big(\alpha(c)\big)$.

II.5.8 Montrer qu'il existe un voisinage $U \subset M_2(\mathbb{R})$ de la matrice $\begin{pmatrix} -3 & 0 \\ 0 & -3 \end{pmatrix}$ et une application $F : U \to M_2(\mathbb{R})$ de classe $\mathcal{C}^1$ telle que :

$$F\begin{pmatrix} -3 & 0 \\ 0 & -3 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 2 & -1 \end{pmatrix} \quad \text{et} \quad \forall M \in U,\ \big(F(M)\big)^2 = M.$$

II.5.9 Soit $I \subset \mathbb{R}$ un intervalle et $A : I \to M_n(\mathbb{C})$ une fonction de classe $\mathcal{C}^1$. On suppose que pour un $t_0 \in I$, la matrice $A(t_0)$ n'a que des valeurs propres simples. Montrer qu'il existe un voisinage $J \subset I$ de t_0 et des fonctions de classe $\mathcal{C}^1$, $(\lambda_k : J \to \mathbb{C})_{k\in[\![1,n]\!]}$, tels que pour tout $t \in J$, la matrice $A(t)$ n'ait que des valeurs propres simples $\lambda_1(t), \dots, \lambda_n(t)$.

Exercices sur la section 2

II.5.10 Parmi les ensembles suivants, lesquels sont des sous-variétés lisses de $\mathbb{R}^2$?

1. $X := \big\{(x,y) \in \mathbb{R}^2 \;\big|\; x^3 + 3xy = 1\big\}$,
2. $Y_c := \big\{(x,y) \in \mathbb{R}^2 \;\big|\; x^2 - y^3 = c\big\}$, $c \in \mathbb{R}$,
3. $Z := \left\{(x,y) \in \mathbb{R}^2 \;\middle|\; \frac{x^2 y}{2} + x^2 + \frac{y^3}{3} - 4y = -3\right\}$,

II.5.11 Parmi les ensembles suivants, lesquels sont des sous-variétés lisses de $\mathbb{R}^3$?

1. $X_c := \big\{(x,y,z) \in \mathbb{R}^3 \;\big|\; x^2 + y^3 + z^4 = c\big\}$, $c \in \mathbb{R}$,
2. $Y := \big\{(x,y,z) \in \mathbb{R}^3 \;\big|\; x^2 + y^2 + z^2 = 1 \text{ et } x^2 + y^2 - x = 0\big\}$,
3. $Z := \big\{(x,y,z) \in \mathbb{R}^3 \;\big|\; y = |x| \text{ et } x^2 - z^2 = 0\big\}$

II.5.12 Le cône $\big\{(x,y,z) \in \mathbb{R}^3 \;\big|\; x^2 + y^2 = z^2\big\}$ est-il une sous-variété lisse de $\mathbb{R}^3$?

II.5.13 Soit Q une forme quadratique non dégénérée de $\mathbb{R}^n$ de signature (p,q).

1. (a) Montrer que l'ensemble $\big\{\mathbf{x} \in \mathbb{R}^n \;\big|\; Q(\mathbf{x}) = 1\big\}$ est une sous-variété lisse de $\mathbb{R}^n$ de dimension $n-1$.
   (b) Déterminer en tout point son espace tangent.
2. On suppose $pq \neq 0$.
   (a) Montrer que $\big\{\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\} \;\big|\; Q(\mathbf{x}) = 0\big\}$ est une sous-variété lisse de $\mathbb{R}^n$ de dimension $n-1$.
   (b) Montrer que le cône $\big\{\mathbf{x} \in \mathbb{R}^n \;\big|\; Q(\mathbf{x}) = 0\big\}$ n'est pas une sous-variété lisse de $\mathbb{R}^n$.

II.5.14 L'ensemble $\big\{(t^2, t^3) \mid t \in \mathbb{R}\big\}$ est-il une sous-variété lisse de $\mathbb{R}^2$?

II.5.15 Soit $\varphi : \mathbb{R}^2 \to \mathbb{R}^3$ l'application définie par :

$$\varphi(r,s) = \big((2+\cos r)\cos s, (2+\cos r)\sin s, \sin r\big).$$

1. Montrer que $\varphi(\mathbb{R}^2)$ est une sous-variété lisse de $\mathbb{R}^3$.
2. Déterminer son espace tangent au point $\varphi(r,s)$.

II.5.16 Soient $X \subset \mathbb{R}^m$ et $Y \subset \mathbb{R}^p$ deux sous-variétés lisses. Montrer que $X \times Y$ est une sous-variété lisse de $\mathbb{R}^{m+p}$. En déduire que le tore $\mathbb{U} \times \mathbb{U}$ est une sous-variété lisse de $\mathbb{C}^2$.

II.5.17 On identifie $M_n(\mathbb{R})$ avec $\mathbb{R}^{n^2}$. On veut montrer que l'ensemble des matrices orthogonales de $M_n(\mathbb{R})$ est une sous-variété lisse de $M_n(\mathbb{R})$. On note $E_s \subset M_n(\mathbb{R})$ le sous-espace vectoriel des matrices symétriques et $E_a \subset M_n(\mathbb{R})$ le sous-espace vectoriel des matrices antisymétriques.

1. Montrer que l'application $F : M_n(\mathbb{R}) \to E_s$ définie par $F(M) = {}^tM \cdot M$ est de classe $\mathcal{C}^1$ et déterminer sa différentielle $D_M F$.
2. On suppose que M est inversible. Déterminer $\operatorname{Ker} D_M F$. En déduire le rang de $D_M F$.
3. Conclure.

Exercices sur la section 3

II.5.18 Soit $f : U \subset \mathbb{C} \to \mathbb{C}$ une fonction holomorphe de classe $\mathcal{C}^2$. Montrer que f' est holomorphe.

II.5.19 Écrire la formule de Taylor à l'ordre 2 pour chacune des fonctions suivantes au point (x_0, y_0) donné.

1. $f_1(x, y) := \sin(x + 2y),\ (x_0, y_0) = (0, 0)$.
2. $f_2(x, y) := \dfrac{1}{x^2 + y^2 + 1},\ (x_0, y_0) = (0, 0)$.
3. $f_3(x, y) := e^{-x^2-y^2} \cos xy,\ (x_0, y_0) = (0, 0)$.
4. $f_4(x, y) := \sin(xy) + \cos(xy),\ (x_0, y_0) = (0, 0)$.
5. $f_5(x, y) := e^{(x-1)^2} \cos y,\ (x_0, y_0) = (1, 0)$.

II.5.20 Trouver les points critiques des fonctions suivantes et déterminer leur nature (minima locaux, maxima locaux, points selle ou autres).

1. $f_1(x, y) := x^3 + 6x^2 + 3y^2 - 12xy + 9x$.
2. $f_2(x, y) := \sin x + y^2 - 2y + 1$.
3. $f_3(x, y, z) := \cos 2x \cdot \sin y + z^2$.
4. $f_4(x, y, z) := (x + y + z)^2$.

II.5.21 Étudier les points critiques des fonctions suivantes :

1. $f_1(x, y) := x^2 - y^2$,
2. $f_2(x, y) := x^4 - y^4$,
3. $f_3(x, y) := x^3 + y^3 - 3xy$,
4. $f_4(x, y) := (x^2 + y^2)e^{-(2x^2+3y^2)}$.

Déterminer leur nature.

II.5.22 Montrer que l'équation $x^2 + y^3 + xyz^3 = 3$ définit implicitement $y = \varphi(x, z)$ au voisinage de $(1, 1, 1)$. Déterminer le polynôme de Taylor d'ordre 2 de φ au point $(1, 1)$.

Exercices sur la section 4

II.5.23 Les formes différentielles suivantes sont-elles fermées ? Sont-elles exactes ?

1. $e^y\,\mathrm{d}x + (x\mathrm{e}^y - 2y)\,\mathrm{d}y$ sur $\mathbb{R}^2$.
2. $(y + 1/x)\,\mathrm{d}x + (x + 1/y)\,\mathrm{d}y$ sur $\{(x, y) \mid x > 0 \text{ et } y > 0\}$.
3. $x\,\mathrm{d}x + y\,\mathrm{d}y$ sur $\mathbb{R}^2$.
4. $y\,\mathrm{d}x - x\,\mathrm{d}y$ sur $\mathbb{R}^2$.

II.5.24 Posons :

$$\omega := \frac{(2x+1)y\,\mathrm{d}x - (x^2 - y^2 + x)\,\mathrm{d}y}{(x^2 + y^2 + x)^2 + y^2}.$$

Quel est le plus grand ouvert $U \subset \mathbb{R}^2$ sur lequel ω est définie ? Montrer que ω est fermée sur U, mais n'y est pas exacte.

II.5.25 Soit $f : U \to \mathbb{C}$ une fonction holomorphe sur un ouvert étoilé $U \subset \mathbb{C}$.
On écrit $f(x + iy) = P(x, y) + iQ(x, y)$ et on pose $\mathrm{d}f = \mathrm{d}P + i\,\mathrm{d}Q$.

1. Montrer que $\mathrm{d}f = f'(z)(\mathrm{d}x + i\,\mathrm{d}y)$.
2. Montrer que la forme différentielle $f(z)(\mathrm{d}x + i\,\mathrm{d}y)$ est exacte.
3. Montrer qu'il existe une fonction holomorphe $F : U \to \mathbb{C}$ telle que $F' = f$.

II.5.26 Calculer l'aire de la région délimitée par :

1. La cardioïde paramétrée en polaire par $\rho = 1 + \cos\theta$,
2. La boucle du folium de Descartes $x^3 + y^3 - 3axy = 0$ (dont on cherchera un paramétrage en posant $y = tx$) et
3. L'astroïde d'équation paramétrique $x(t) = \cos^3(t)$ et $y(t) = \sin^3(t)$.

Fonctions analytiques

II.6

Dans ce module, nous allons étudier les fonctions analytiques d'une variable complexe, c'est-à-dire les fonctions $f : U \to \mathbb{C}$ définies sur un ouvert $U \subset \mathbb{C}$ qui sont développables en séries entières au voisinage de tout point de U.
Nous montrerons dans un premier temps qu'une telle fonction est holomorphe sur U, c'est-à-dire $\mathbb{C}$-dérivable en tout point de U (voir le module IV.7 de [L1] pour une introduction aux fonctions holomorphes). La dérivée d'une fonction analytique est elle-même analytique. Une fonction analytique dans un ouvert $U \subset \mathbb{C}$ est donc indéfiniment $\mathbb{C}$-dérivable dans U. En fait, la réciproque est vraie et, aussi surprenant que cela puisse paraître, toute fonction $\mathbb{C}$-dérivable dans un ouvert $U \subset \mathbb{C}$ est automatiquement développable en série entière au voisinage de tout point de U. En particulier, toute fonction $\mathbb{C}$-dérivable dans U est indéfiniment $\mathbb{C}$-dérivable dans U. Cela est bien entendu faux pour les fonctions d'une variable réelle qui peuvent être dérivables une fois sans pour autant l'être deux fois. À la fin du module (section 4.2), nous démontrerons un résultat légèrement plus faible, mais qui est tout aussi surprenant. Nous montrerons (corollaire 29 de la page 787) qu'une fonction holomorphe dans un ouvert $U \subset \mathbb{C}$ dont la dérivée est continue sur U est une fonction analytique sur U. Le lecteur pourra voir dans [MPA] que la condition de continuité de la dérivée est automatiquement satisfaite lorsque f est $\mathbb{C}$-dérivable.
La plupart des démonstrations présentées dans ce module reposent sur les propriétés des séries entières. Le lecteur est donc invité à lire les parties des modules II.2 et I.1 qui concernent les séries entières et les séries formelles avant d'aborder ce module. Dans la dernière partie, nous montrerons qu'il existe un lien étroit entre les séries entières et les séries de Fourier. Le lecteur est donc invité à d'abord lire le module II.4 sur les séries de Fourier avant d'aborder la section 4.

1 Définition et premières propriétés

Dans tout ce module, si $z_0 \in \mathbb{C}$ et $R > 0$, nous noterons $D(z_0, R) \subset \mathbb{C}$ le disque ouvert de centre z_0 et de rayon R et $\overline{D}(z_0, R)$ son adhérence, le disque fermé de centre z_0 et de rayon R. En particulier, nous notons $\mathbb{D} := D(0, 1)$ le disque unité ouvert et $\mathbb{U}$ la frontière de ce dernier (cercle unité). Dans ce qui suit, nous serons fréquemment amenés à considérer des séries entières de rayon de convergence égal à $+\infty$. Lorsque $R = +\infty$, $D(z_0, R) := \mathbb{C}$.

1.1 Retour sur les séries entières

Définition 1. Une fonction f définie au voisinage de z_0 est *développable en série entière en* z_0 s'il existe une série entière $F := \sum_{n\geqslant 0} a_n z^n$ dont le rayon de convergence n'est pas nul, telle que $f(z) = \sum_{n=0}^{+\infty} a_n (z-z_0)^n$ pour $|z-z_0|$ assez petit.

Cette définition s'applique aussi bien dans le cas où z est une variable réelle que dans le cas où z est une variable complexe. Dans le module II.2, nous avons traité le cas où $z_0 = 0$. On s'y ramène aisément en faisant le changement de variable $w = z - z_0$. Ici, nous considérerons principalement le cas de fonctions d'une variable complexe.
Commençons par rappeler des résultats démontrés dans le module II.2 en les adaptant à notre situation. Auparavant, rappelons que l'ordre $\omega(F)$ d'une série entière $F := \sum_{n\geqslant 0} a_n z^n$ non nulle est le plus petit entier n tel que $a_n \neq 0$. Par convention, $\omega(0) = +\infty$.

Théorème 1. Soient f et g deux fonctions développables en série entière en z_0. Alors $f+g$ et fg sont développables en série entière en z_0.

Démonstration.
Par hypothèse, il existe $r > 0$ et deux séries entières $\sum_{n\geqslant 0} a_n z^n$ et $\sum_{n\geqslant 0} b_n z^n$ de rayons de convergence non nuls, tels que si $|z-z_0| < r$, alors :

$$f(z) = \sum_{n=0}^{+\infty} a_n (z-z_0)^n \quad \text{et} \quad g(z) = \sum_{n=0}^{+\infty} b_n (z-z_0)^n.$$

D'après le théorème 35 de la page 509, pour tout $z \in D(z_0, r)$, on a :

$$(f+g)(z) = \sum_{n=0}^{+\infty} (a_n + b_n)(z-z_0)^n.$$

D'après le théorème 36 de la page 510, pour tout $z \in D(z_0, r)$, on a :

$$(fg)(z) = \sum_{n=0}^{+\infty} c_n (z-z_0)^n \quad \text{avec} \quad c_n := \sum_{k=0}^{n} a_k b_{n-k}.$$

Cela montre que $f+g$ et fg sont développables en série entière en z_0. ∎

Théorème 2.
Soit f une fonction développable en série entière en z_0 et g une fonction développable en série entière en $f(z_0)$. Alors $g \circ f$ est développable en série entière en z_0.

Démonstration. Posons $w_0 := f(z_0)$. Par hypothèse, pour $|z-z_0|$ assez petit, on a :

$$f(z) = w_0 + \sum_{n=1}^{+\infty} a_n (z-z_0)^n.$$

De plus, pour $|w-w_0|$ suffisamment petit, on a :

$$g(w)=\sum_{n=0}^{+\infty} b_n(w-w_0)^n.$$

Posons $u:=z-z_0$ et $v:=w-w_0$. Alors :

$$f(z_0+u)-w_0=\sum_{n=1}^{+\infty} a_n u^n \quad \text{et} \quad g(w_0+v)=\sum_{n=0}^{+\infty} b_n v^n.$$

Soient F et G les séries entières formelles $F:=\sum\limits_{n\geqslant 1} a_n u^n$ et $G:=\sum\limits_{n\geqslant 0} b_n v^n$. Étant donné que l'ordre $\omega(F)$ de la série formelle F est non nul, il existe une série formelle $H=G\circ F$ (voir le théorème 66 de la page 45). On la note $H:=\sum\limits_{p\geqslant 0} c_p z^p$. D'après la proposition 38 de la page 511, si $|u|$ est assez petit :

$$g\Big(w_0+\big(f(z_0+u)-w_0\big)\Big)=\sum_{p=0}^{+\infty} c_p u^p.$$

Cela montre que si $|z-z_0|$ est suffisamment petit :

$$g\big(f(z)\big)=\sum_{p=0}^{+\infty} c_p (z-z_0)^p.$$

Remarque. Notons que l'application $g\mapsto g\circ f$ est une application linéaire tandis que l'application $f\mapsto g\circ f$ n'est pas linéaire, sauf cas très particulier.

Corollaire 3. Soit f une fonction développable en série entière en z_0 avec $f(z_0)\neq 0$. Alors $1/f$ est développable en série entière en z_0.

Démonstration. Posons $w_0:=f(z_0)\neq 0$. La fonction $g: w\mapsto \dfrac{1}{w}$ est développable en série entière en w_0. En effet, pour tout $w\in D\big(w_0,|w_0|\big)$, on a $\left|\dfrac{w-w_0}{w_0}\right|<1$ et donc :

$$\frac{1}{w}=\frac{1}{w_0+(w-w_0)}=\frac{1}{w_0}\cdot\frac{1}{1+\dfrac{w-w_0}{w_0}}=\sum_{n=0}^{+\infty}\frac{(-1)^n}{w_0^{n+1}}(w-w_0)^n.$$

Il suffit maintenant d'appliquer le théorème précédent pour en déduire que $1/f=g\circ f$ est développable en série entière en z_0.

Le résultat qui suit n'a été démontré dans le module II.2 que dans le cadre des fonctions d'une variable réelle. Nous le généralisons ici au cas des fonctions d'une variable complexe.

Rappelons auparavant qu'une fonction f définie au voisinage de $z_0 \in \mathbb{C}$ est $\mathbb{C}$-dérivable en z_0 (voir le module IV.7 de [L1]) si, et seulement si, la limite suivante existe :

$$\lim_{z\to z_0} \frac{f(z)-f(z_0)}{z-z_0}.$$

On la note alors $f'(z_0)$. Si $f : U \to \mathbb{C}$ est définie sur un ouvert $U \subset \mathbb{C}$ et si f est $\mathbb{C}$-dérivable en tout point de U, on dit que f est holomorphe sur U.

Théorème 4. Soit $F := \sum_{n\geqslant 0} a_n z^n$ une série entière de rayon de convergence $R > 0$.

Sa somme $f : z \mapsto \sum_{n=0}^{+\infty} a_n z^n$ est holomorphe sur $D(0,R)$ et :

$$f'(z) = \sum_{n=1}^{+\infty} n a_n z^{n-1}.$$

Démonstration. D'après la proposition 40 de la page 515, la série dérivée $\sum_{n\geqslant 1} n a_n z^{n-1}$ a même rayon de convergence R que f. Nous devons montrer que pour tout $z_0 \in D(0,R)$:

$$\lim_{z\to z_0} \frac{f(z)-f(z_0)}{z-z_0} = \sum_{n=1}^{+\infty} n a_n z_0^{n-1}.$$

Pour $z \in D(0,R)\setminus\{z_0\}$, on a :

$$\frac{f(z)-f(z_0)}{z-z_0} = \sum_{n=1}^{+\infty} a_n \frac{z^n - z_0^n}{z-z_0} = \sum_{n=1}^{+\infty} g_n(z)$$

avec :

$$g_n(z) = a_n\big(z^{n-1} + z_0 z^{n-2} + \cdots + z_0^{n-2} z + z_0^{n-1}\big).$$

Choisissons r tel que $|z_0| < r < R$ et observons que pour tout $z \in D(0,r)$, on a :

$$\big|g_n(z)\big| \leqslant n|a_n| r^{n-1}.$$

La série $\sum_{n\geqslant 1} g_n$ est donc normalement convergente sur $D(0,r)$. Les fonctions g_n sont continues sur $\mathbb{C}$. La somme $G : z \mapsto \sum_{n=1}^{+\infty} g_n(z)$ est donc continue sur $D(0,r)$. En particulier, elle est continue en z_0. On a donc :

$$\lim_{z\to z_0} \frac{f(z)-f(z_0)}{z-z_0} = \lim_{z\to z_0} G(z) = G(z_0) = \sum_{n=1}^{+\infty} n a_n z_0^{n-1}.$$

■

Corollaire 5. Soit $F := \sum_{n\geqslant 0} a_n z^n$ une série entière de rayon de convergence $R > 0$ et soit $f : D(0,R) \to \mathbb{C}$ sa somme. Alors :

1. La fonction f est indéfiniment $\mathbb{C}$-dérivable sur $D(0,R)$. En particulier, f est holomorphe sur $D(0,R)$.
2. Pour tout $k \geqslant 0$:
$$f^{(k)}(z) = \sum_{n=k}^{+\infty} \frac{n!}{(n-k)!} a_n z^{n-k}.$$
3. Le développement en série entière de f en 0 est la série de Taylor en 0 :
$$\forall k \geqslant 0, \quad a_k = \frac{1}{k!} f^{(k)}(0).$$

Démonstration. Les points (1) et (2) se démontrent sans aucune difficulté par récurrence. Les détails sont laissés au lecteur. Pour le point (3), il suffit de remplacer z par 0 dans la formule de la dérivée $k^{ème}$. Le seul terme non nul est alors donné par $n = k$. ∎

Remarque. Rappelons (module IV.7 de [L1]) qu'une fonction d'une variable complexe peut-être assimilée à une fonction de deux variables réelles après identification de $\mathbb{C}$ avec $\mathbb{R}^2$. Dans l'énoncé précédent, nous avons écrit « indéfiniment $\mathbb{C}$-dérivable » plutôt que « de classe $\mathcal{C}^\infty$ ». C'est un énoncé plus fort étant donné qu'une fonction de classe $\mathcal{C}^1$ n'est pas forcément $\mathbb{C}$-dérivable (il faut que la différentielle soit une similitude).

Exemple. La fonction $\exp : \mathbb{C} \to \mathbb{C}$ est définie par la série entière :
$$\exp(z) := \sum_{n=0}^{+\infty} \frac{z^n}{n!}.$$
Le rayon de convergence est infini. Le disque de convergence est égal à $\mathbb{C}$. La fonction $\exp$ est donc $\mathbb{C}$-dérivable sur $\mathbb{C}$, *i.e.* holomorphe sur $\mathbb{C}$. Sa dérivée est :
$$\exp'(z) = \sum_{n=1}^{+\infty} \frac{n z^{n-1}}{n!} = \exp(z).$$

Exercice 1.

Montrer que pour tout $z \in D(0,1)$ et tout $k \geqslant 0$,
$$\frac{k!}{(1-z)^{k+1}} = \sum_{n=k}^{+\infty} \frac{n!}{(n-k)!} z^{n-k}.$$

Solution. Une solution a été proposée dans la proposition 37 de la page 510 en utilisant la formule du produit de séries entières. Une alternative consiste à dériver k fois terme à

terme la série :

$$\frac{1}{1-z} = \sum_{n=0}^{+\infty} z^n.$$

On démontre alors la formule par récurrence. Elle est vraie pour $k = 0$ et :

$$\frac{k!}{(1-z)^{k+1}} = \left(\frac{(k-1)!}{(1-z)^k}\right)' = \sum_{n=k-1}^{+\infty} \left(\frac{n!}{(n-k+1)!} z^{n-k+1}\right)' = \sum_{n=k}^{+\infty} \frac{n!}{(n-k)!} z^{n-k}.$$

1.2 Fonctions analytiques

Ces rappels sur les séries entières ayant été faits, venons-en maintenant à la notion de fonction analytique sur un ouvert de $\mathbb{C}$.

Définition 2. Une fonction $f : U \to \mathbb{C}$ définie sur un ouvert $U \subset \mathbb{C}$ est *analytique* dans U si elle est développable en série entière en tout point de U.

Exercice 2.

Montrer que tout polynôme $P(z) := \sum_{n=0}^{d} a_n z^n$ définit une fonction analytique sur $\mathbb{C}$ et que :

$$P(z) = \sum_{k=0}^{d} \frac{1}{k!} \left(\sum_{n=k}^{d} \frac{n!}{(n-k)!} a_n z_0^{n-k}\right) (z - z_0)^k.$$

Solution. Étant donné $z_0 \in \mathbb{C}$, la formule de Taylor à l'ordre k donne :

$$P(z) = \sum_{k=0}^{d} \frac{1}{k!} P^{(k)}(z_0) \cdot (z - z_0)^k.$$

Le polynôme P est donc développable en série entière en z_0. Le corollaire 5 de la page précédente implique que :

$$P^{(k)}(z_0) = \sum_{n=k}^{d} \frac{n!}{(n-k)!} a_n z_0^{n-k}.$$

Le résultat suit immédiatement.

Les trois énoncés qui suivent découlent directement des propriétés des séries entières que nous avons rappelées plus haut.

Proposition 6. Soient $f : U \to \mathbb{C}$ et $g : U \to \mathbb{C}$ deux fonctions analytiques sur un ouvert U de $\mathbb{C}$. Alors $f + g$ et fg sont analytiques sur U. Si de plus g ne s'annule pas sur U, alors f/g est analytique sur U.

Exemple. Soient $P, Q \in \mathbb{C}[X]$ deux polynômes (premiers entre eux). Soit $\mathcal{R}$ l'ensemble des zéros de Q. Soit $f : \mathbb{C} \setminus \mathcal{R} \to \mathbb{C}$ la fraction rationnelle définie par $f(z) = P(z)/Q(z)$. Alors, f est analytique sur $\mathbb{C} \setminus \mathcal{R}$.

Proposition 7. Soient U et V deux ouverts de $\mathbb{C}$. Soient $f : U \to V$ et $g : V \to \mathbb{C}$ deux fonctions analytiques. Alors, $g \circ f$ est analytique sur U.

Proposition 8. Si $f : U \to \mathbb{C}$ est analytique sur un ouvert U de $\mathbb{C}$, alors f est indéfiniment $\mathbb{C}$-dérivable sur U et ses dérivées successives $f^{(k)}$ sont analytiques sur U. En particulier, f est holomorphe sur U.

> ***Remarque***. Nous venons d'énoncer un résultat fondamental. Une fonction analytique sur un ouvert $U \subset \mathbb{C}$ y est holomorphe. Dans [MPA], nous montrerons que la réciproque est vraie. Toute fonction holomorphe sur un ouvert $U \subset \mathbb{C}$ y est analytique. À la fin de ce module, nous montrerons cette réciproque dans le cas où f est holomorphe sur U et f' est continue sur U.

Il est généralement fastidieux de vérifier directement qu'une fonction $f : U \to \mathbb{C}$ définie sur un ouvert U de $\mathbb{C}$ est développable en série entière au voisinage de chacun de ses points. Le travail est largement simplifié par l'énoncé suivant.

Théorème 9. Si une série entière $\sum_{n \geqslant 0} a_n z^n$ a un rayon de convergence R non nul, sa somme f est analytique sur le disque de convergence $D(0, R)$. Plus précisément, si $D(z_0, r) \subset D(0, R)$, alors f est développable en série entière en z_0 avec un rayon de convergence supérieur ou égal à r.

Démonstration. Soit $z_0 \in D(0, R)$. Posons $r_0 := |z_0|$. Nous venons de voir que f est indéfiniment $\mathbb{C}$-dérivable sur $D(0, R)$ et que pour tout $k \geqslant 0$:

$$f^{(k)}(z_0) = \sum_{n=k}^{+\infty} \frac{n!}{(n-k)!} a_n z_0^{n-k}.$$

Nous allons montrer que pour tout $z \in D(z_0, R - r_0)$, on a :

$$f(z) = \sum_{k=0}^{+\infty} \frac{1}{k!} f^{(k)}(z_0) \cdot (z - z_0)^k.$$

Cela complétera la démonstration du théorème.

D'après l'exercice 2 de la page précédente, pour tout entier $d \geqslant 0$, on a :

$$\sum_{n=0}^{d} a_n z^n = \sum_{k=0}^{d} \frac{1}{k!} \left(\sum_{n=k}^{d} \frac{n!}{(n-k)!} a_n z_0^{n-k} \right) (z - z_0)^k.$$

Par conséquent :

$$f(z) - \sum_{k=0}^{d} \frac{1}{k!} f^{(k)}(z_0) \cdot (z - z_0)^k = A_d + B_d$$

avec :

$$A_d := f(z) - \sum_{n=0}^{d} a_n z^n \underset{d \to +\infty}{\longrightarrow} 0$$

et :

$$B_d := \sum_{k=0}^{d} \frac{1}{k!} \left(\sum_{n=d+1}^{+\infty} \frac{n!}{(n-k)!} a_n z_0^{n-k} \right) (z - z_0)^k = \sum_{n=d+1}^{+\infty} \sum_{k=0}^{d} a_n \binom{n}{k} z_0^{n-k} (z - z_0)^k.$$

Notons que :

$$|B_d| \leqslant \sum_{n=d+1}^{+\infty} |a_n| \sum_{k=0}^{n} \binom{n}{k} |z_0|^{n-k} |z - z_0|^k = \sum_{n=d+1}^{+\infty} |a_n| \big(|z_0| + |z - z_0|\big)^n.$$

Comme $r := |z_0| + |z - z_0| < R$, on a :

$$\lim_{d \to +\infty} \sum_{n=d+1}^{+\infty} |a_n| r^n = 0 \quad \text{et} \quad \lim_{d \to +\infty} B_d = 0.$$

Exemples.

1. La fonction $\exp : \mathbb{C} \to \mathbb{C}$ définie par la série entière $\exp(z) := \sum_{n=0}^{+\infty} \frac{z^n}{n!}$ est analytique sur $\mathbb{C}$. En effet, son rayon de convergence est infini et le disque de convergence est égal à $\mathbb{C}$.
2. Les fonctions $\sin : \mathbb{C} \to \mathbb{C}$ et $\cos : \mathbb{C} \to \mathbb{C}$ définies par :
$$\sin(z) := \frac{\exp(iz) - \exp(-iz)}{2i} \quad \text{et} \quad \cos(z) := \frac{\exp(iz) + \exp(-iz)}{2}$$
sont analytiques sur $\mathbb{C}$.
3. Les fonctions $\sinh : \mathbb{C} \to \mathbb{C}$ et $\cosh : \mathbb{C} \to \mathbb{C}$ définies par :
$$\sinh(z) := \frac{\exp(z) - \exp(-z)}{2} \quad \text{et} \quad \cosh(z) := \frac{\exp(z) + \exp(-z)}{2}$$
sont analytiques sur $\mathbb{C}$. On remarquera que, si $z \in \mathbb{C}$:
$$\cosh(iz) = \cos z \quad \text{et} \quad \sinh(iz) = i \sin z.$$

2 Principe du prolongement analytique Principe des zéros isolés

À partir de maintenant, la plupart des fonctions que nous allons considérer seront analytiques sur un ouvert connexe.

Définition 3. Une partie $U \subset \mathbb{C}$ est un *domaine* si U est ouverte et connexe.

2.1 Principe du prolongement analytique

Théorème 10. Soient U un domaine de $\mathbb{C}$, z_0 un point de U et f une fonction analytique sur U. Les conditions suivantes sont équivalentes :

1. La fonction f est identiquement nulle dans U.
2. La fonction f est identiquement nulle dans un voisinage de z_0.
3. Pour tout $n \in \mathbb{N}$, on a $f^{(n)}(z_0) = 0$.

Remarque. Ce théorème est faux si l'on ne fait pas l'hypothèse que U est connexe. En effet, dans ce cas, il se peut que f soit identiquement nulle sur une des composantes connexes de U, sans pour autant être identiquement nulle sur toutes les composantes connexes de U.

Démonstration. L'implication (1) $\Rightarrow$ (3) est évidente.

(3) $\Rightarrow$ (2). Comme f est analytique sur U, elle est développable en série entière au voisinage de z_0. La série entière de f en z_0 est la série de Taylor de f en z_0. Par conséquent, au voisinage de z_0 on a :

$$f(z) = \sum_{n=0}^{+\infty} \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n = 0.$$

(2) $\Rightarrow$ (1). Soit $V \subset U$ l'ensemble des points $z \in U$ tels que f soit identiquement nulle au voisinage de z. Nous allons montrer que V est non vide, qu'il est à la fois ouvert et fermé dans U. Comme U est connexe, nous aurons alors établi que $V = U$.

Par hypothèse, V est non vide puisqu'il contient z_0.

Par définition V est ouvert puisqu'il contient un voisinage de chacun de ses points.

Montrons maintenant que V est fermé dans U et pour cela, considérons une suite de points $z_n \in V$ qui converge vers un point $a \in U$. La fonction f et toutes ses dérivées $f^{(k)}$ sont analytiques, donc continues sur U. Comme f est identiquement nulle au voisinage de chaque point z_n, pour tout n et pour tout $k \geqslant 0$, on a $f^{(k)}(z_n) = 0$. Par continuité de $f^{(k)}$ au point $a \in U$, on en déduit que pour tout $k \geqslant 0$, $f^{(k)}(a) = 0$. Or au voisinage de a, f est la somme de sa série de Taylor en a. Par conséquent, f est identiquement nulle au voisinage de a. Donc, $a \in V$. ∎

Corollaire 11. Soit U un domaine de $\mathbb{C}$. Soient f et g deux fonctions analytiques sur U. Si f et g coïncident au voisinage d'un point $z_0 \in U$, alors f est égale à g sur U.

Démonstration. Appliquer le théorème précédent à la fonction analytique $f - g : U \to \mathbb{C}$. ∎

Exercice 3.

Soit $f : U \to \mathbb{C}$ une fonction analytique dans un domaine $U \subset \mathbb{C}$. Montrer que f est constante dans U si, et seulement si, la dérivée f' est identiquement nulle.

Solution. Si f est constante, alors sa dérivée est identiquement nulle car les taux d'accroissement sont tous nuls.

Si la dérivée de f est nulle ne serait-ce qu'au voisinage d'un point $z_0 \in U$, tous les termes du développement en série entière de f en z_0 sont nuls sauf éventuellement le terme constant a_0. La fonction f et la fonction constante égale à a_0 coïncident donc au voisinage de z_0. Elles coïncident donc partout.

Définition 4. Soient $U \subset V$ deux ouverts de $\mathbb{C}$ avec V connexe. Soient $f : U \to \mathbb{C}$ et $g : V \to \mathbb{C}$ deux fonctions analytiques. L'application g est un *prolongement analytique* de f si f est la restriction de g à U : $f = g_{|U}$.

Le corollaire précédent peut être reformulé de la manière suivante.

Corollaire 12. Soit $f : U \to \mathbb{C}$ une fonction analytique sur U et soit $V \supset U$ un domaine contenant U. Alors, f admet au plus un prolongement analytique $g : V \to \mathbb{C}$.

Étant donnée une fonction analytique $f : U \to \mathbb{C}$, il est alors naturel de se demander si f admet un prolongement analytique $g : V \to \mathbb{C}$ avec $U \subsetneq V$. Le résultat précédent affirme que si $g_1 : V \to \mathbb{C}$ et $g_2 : V \to \mathbb{C}$ sont deux tels prolongements, alors ils sont égaux.

Il existe des fonctions analytiques $f : U \to \mathbb{C}$ avec U borné, qui n'admettent pas de prolongement analytique à des domaines plus gros que U. C'est l'objet de l'exercice suivant qui traite le cas où $U = \mathbb{D} := \{z \in \mathbb{C} \mid |z| < 1\}$.

Exercice 4.

Montrer que la somme de la série entière $\sum_{n\geqslant 0} z^{2^n}$ est analytique sur $\mathbb{D}$, mais qu'il n'existe aucun prolongement analytique $g : V \to \mathbb{C}$ avec $\mathbb{D} \subset V$, V connexe et $V \neq \mathbb{D}$.

Solution. Supposons par l'absurde que $g : V \to \mathbb{D}$ soit un tel prolongement. Alors V intersecte le cercle unité, et comme V est ouvert, il existe un point $z_0 \in V$ et un entier n_0 tels que $z_0^{2^{n_0}} = 1$ (voir l'exercice II.6.10 de la page 795). Alors, pour tout $n \geqslant n_0$, on a $z_0^{2^n} = 1$ et :

$$g(z_0) = \lim_{\substack{t\to 1\\ t<1}} g(tz_0) = \lim_{\substack{t\to 1\\ t<1}} f(tz_0).$$

Or, $f(tz_0) = \sum_{n=0}^{n_0-1} t^{2^n} z_0^{2^n} + \sum_{n=n_0}^{+\infty} t^{2^n}$ et $\lim\limits_{\substack{t\to 1\\ t<1}} \sum_{n=0}^{n_0-1} t^{2^n} z_0^{2^n} = \sum_{n=0}^{n_0-1} z_0^{2^n}$.

De plus, pour tout $n_1 \geqslant n_0$:

$$\sum_{n=n_0}^{+\infty} t^{2^n} \geqslant \sum_{n=n_0}^{n_1} t^{2^n} \underset{\substack{t\to 1\\ t<1}}{\longrightarrow} n_1 - n_0 + 1 \underset{n_1\to+\infty}{\longrightarrow} +\infty.$$

Donc :

$$\lim_{\substack{t\to 1\\ t<1}} \sum_{n=n_0}^{+\infty} t^{2^n} = +\infty \quad \text{et} \quad \lim_{\substack{t\to 1\\ t<1}} f(tz_0) = \infty.$$

On a donc une contradiction.

Il n'existe pas nécessairement un prolongement analytique *maximal* $\tilde{f} : \widetilde{U} \to \mathbb{C}$ au sens où tout autre prolongement analytique $g : V \to \mathbb{C}$ avec $U \subset V$ en serait une restriction.

Exemple. Nous verrons plus loin (section 3) que la fonction $\mathrm{Log} : D(1,1) \to \mathbb{C}$ définie par :

$$\mathrm{Log}(z) = \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n}(z-1)^n$$

n'admet pas de prolongement analytique maximal $\tilde{f} : \widetilde{U} \to \mathbb{C}$ avec $\widetilde{U} \subset \mathbb{C}$.

2.2 Principe des zéros isolés

Définition 5. Si $f : U \to \mathbb{C}$ est une fonction analytique sur un ouvert $U \subset \mathbb{C}$, on note $Z(f)$ l'ensemble des zéros de f dans U, *i.e.* :

$$Z(f) = \big\{ z \in U \;\big|\; f(z) = 0 \big\}.$$

Proposition et définition 13. Si $U \subset \mathbb{C}$ est un domaine et si $f : U \to \mathbb{C}$ est analytique et non identiquement nulle, à chaque zéro $z_0 \in Z(f)$ correspond un unique entier positif m et une unique fonction analytique $g : U \to \mathbb{C}$ tels que :

$$\forall z \in U, \quad f(z) = (z - z_0)^m g(z) \quad \text{avec} \quad g(z_0) \neq 0.$$

L'entier m est la *multiplicité* de z_0 comme zéro de f, ou l'*ordre* de f en z_0.

Démonstration. Fixons $z_0 \in Z(f)$. Soit $r > 0$ et $F := \sum_{n\geqslant 0} a_n z^n$ tels que :

$$|z - z_0| < r \quad \Longrightarrow \quad f(z) = \sum_{n=0}^{+\infty} a_n (z - z_0)^n.$$

Comme U est connexe et f n'est pas identiquement nulle, le principe du prolongement analytique (théorème 10 de la page 769) montre qu'il existe un plus petit entier m tel que $a_m \neq 0$. L'entier m est l'ordre $\omega(F)$ de la série entière F.

Soit $g : U \to \mathbb{C}$ la fonction définie par :

$$g(z) = \begin{cases} \dfrac{f(z)}{(z-z_0)^m} & \text{si } z \neq z_0 \\ a_m & \text{si } z = z_0. \end{cases}$$

La fonction g est analytique sur $U \setminus \{z_0\}$ car $(z-z_0)^m$ ne s'y annule pas. De plus, pour tout $z \in D(z_0, r)$, on a :

$$(z-z_0)^m g(z) = f(z) = (z-z_0)^m \sum_{n=m}^{+\infty} a_n (z-z_0)^{n-m} = (z-z_0)^m \sum_{k=0}^{+\infty} a_{m+k} (z-z_0)^k.$$

On en déduit que pour tout $z \in D(z_0, r) \setminus \{z_0\}$, on a :

$$g(z) = \sum_{k=0}^{+\infty} a_{m+k} (z-z_0)^k.$$

Cette égalité reste vraie pour $z = z_0$ puisque les deux membres sont alors égaux à a_m. Nous venons de montrer que g est développable en série entière en z_0.
Pour l'unicité, il suffit d'observer que si $(z-z_0)^{m_1} g_1(z) = (z-z_0)^{m_2} g_2(z)$ dans U avec g_1 et g_2 continues en z_0, $g_1(z_0) \neq 0$ et $g_2(z_0) \neq 0$, alors $\dfrac{(z-z_0)^{m_1}}{(z-z_0)^{m_2}}$ admet une limite non nulle quand z tend vers z_0. Cela n'est possible que si $m_1 = m_2$. Dans ce cas, $g_1 = g_2$ sur $U \setminus \{z_0\}$. Comme g_1 et g_2 sont continues en z_0, on a $g_1 = g_2$ sur U. ∎

Exemple. La fonction $\sin : \mathbb{C} \to \mathbb{C}$ s'annule en 0 et sa dérivée en 0 est $\cos(0) = 1$. La fonction sinus a donc un zéro d'ordre 1 à l'origine et la fonction *sinus cardinal* définie par :

$$\operatorname{sinc}(z) := \begin{cases} \dfrac{\sin z}{z} & \text{si } z \neq 0 \\ 1 & \text{si } z = 0 \end{cases}$$

est analytique sur $\mathbb{C}$.

Définition 6. Soit X un sous-ensemble de $\mathbb{C}$.
- Un point $z_0 \in X$ est un *point isolé* de X s'il existe $r > 0$ tel que $D(z_0, r) \cap X = \{z_0\}$.
- Un point $z_0 \in \mathbb{C}$ est un *point d'accumulation* de X si pour tout $r > 0$, $D(z_0, r) \cap X$ contient une infinité de points.
- L'ensemble X est *discret* si tous ses points sont isolés, c'est-à-dire si pour tout $z_0 \in X$, il existe $r > 0$ tel que $D(z_0, r) \cap X = \{z_0\}$.

Un ensemble $X \subset \mathbb{C}$ est discret si, et seulement si, X ne contient pas de point d'accumulation de X.
Plus généralement, un point de X est soit isolé soit un point d'accumulation de X.

Exemples.
- L'ensemble $\{1/n \mid n \in \mathbb{N}^*\}$ est discret tandis que l'ensemble $\{0\} \cup \{1/n \mid n \in \mathbb{N}^*\}$ n'est pas discret.
- Un ensemble X est discret s'il n'a pas de point d'accumulation dans X. On a alors $X \subset \bigcup_{z \in X} D(z, r_z)$ avec $r_z > 0$ et $D(z, r_z) \cap X = \{z\}$. Si X est un ensemble compact et discret, la propriété de Borel-Lebesgue implique que l'on peut recouvrir X par un nombre fini de disque $D(z, r_z)$ (théorème 50 de la page 435), et donc que X est un ensemble fini.

Théorème 14 (Principe des zéros isolés).
Soit $U \subset \mathbb{C}$ un domaine et $f : U \to \mathbb{C}$ une fonction analytique non identiquement nulle. Soit $z_0 \in Z(f)$ un zéro de f. Alors z_0 est un point isolé de $Z(f)$.

Démonstration.
Soit $z_0 \in Z(f)$, soit m sa multiplicité comme zéro de f et soit $g : U \to \mathbb{C}$ la fonction analytique vérifiant :

$$f(z) = (z - z_0)^m g(z) \quad \text{avec} \quad g(z_0) \neq 0.$$

Par continuité de g en z_0, il existe $r > 0$ tel que g ne s'annule pas sur $D(z_0, r)$. L'identité $f(z) = (z - a)^m g(z)$ montre que f ne s'annule pas sur $D(z_0, r) \setminus \{z_0\}$. Le point z_0 est donc un point isolé de $Z(f)$ car $Z(f) \cap D(z_0, r) = \{z_0\}$. ∎

Nous venons de voir que $Z(f)$ est un ensemble discret. En utilisant la continuité de $f : U \to \mathbb{C}$, on peut renforcer cette assertion.

Corollaire 15. Soit $U \subset \mathbb{C}$ un domaine et $f : U \to \mathbb{C}$ une fonction analytique non identiquement nulle. L'ensemble $Z(f)$ est un fermé discret de U. En particulier, $Z(f)$ n'a pas de points d'accumulation dans U.

Démonstration. Comme f est continue $Z(f) := f^{-1}\big(\{z_0\}\big)$ est un fermé de U.
Les points d'accumulation de $Z(f)$ dans U sont donc exactement les points non isolés de $Z(f)$. Le principe des zéros isolés donne le résultat. ∎

Exemple. Le corollaire précédent dit qu'une fonction analytique sur un ouvert $U \subset \mathbb{C}$ ne possède pas de point d'accumulation dans U. Il se peut cependant qu'il y ait un point d'accumulation dans $\mathbb{C} \setminus U$. Par exemple, si $U = \mathbb{C}^*$, l'application $f : U \to \mathbb{C}$ définie par :

$$f(z) = \sin \frac{\pi}{z}$$

s'annule en tous les points $1/k$ pour $k \in \mathbb{Z} \setminus \{0\}$. Le point $0 \in \mathbb{C} \setminus U$ est un point d'accumulation de l'ensemble $Z(f)$.

Corollaire 16. Soit $U \subset \mathbb{C}$ un domaine et $f : U \to \mathbb{C}$ une fonction analytique non identiquement nulle.
- Pour tout compact $K \subset U$, f n'a qu'un nombre fini de zéros dans K.
- L'ensemble $Z(f)$ des zéros de f est au plus dénombrable.

Démonstration.
- D'après l'exemple de la page ci-contre, si f possédait une infinité de zéros dans un compact $K \subset U$, alors $Z(f)$ aurait un point d'accumulation dans $K \subset U$.
- Tout ouvert $U \subset \mathbb{C}$ est une réunion dénombrable de compacts.
  Par exemple, $U = \bigcup_{n \geqslant 1} \big\{ z \in \mathbb{C} \;\big|\; |z| \leqslant n \text{ et } D(z, 1/n) \subset U \big\}$ (voir l'exercice II.6.11 de la page 795). L'ensemble $Z(f)$ est donc une réunion dénombrable d'ensembles finis. ∎

Exercice 5.
Soit $f : \mathbb{C} \to \mathbb{C}$ une fonction analytique telle que $f(\mathbb{R}) \subset \mathbb{R}$. Montrer que pour tout $z \in \mathbb{C}$, on a :

$$\overline{f(z)} = f(\bar{z}).$$

Solution. Soit $g : \mathbb{C} \to \mathbb{C}$ la fonction définie par $g(z) := \overline{f(\bar{z})}$.
Notons que g est analytique sur $\mathbb{C}$. En effet, pour tout $z_0 \in \mathbb{C}$, la fonction f est développable en série entière au voisinage de $\bar{z}_0$: $f(w) = \sum_{n=0}^{+\infty} a_n(w - \bar{z}_0)^n$ pour $|w - \bar{z}_0|$ suffisamment petit. Alors, si $|z - z_0|$ est suffisamment petit, en posant $w = \bar{z}$, on a :

$$g(z) = \overline{f(w)} = \sum_{n=0}^{+\infty} \overline{a_n(w - \bar{z}_0)^n} = \sum_{n=0}^{+\infty} \bar{a}_n(z - z_0)^n.$$

Pour tout $z \in \mathbb{R}$, on a $z = \bar{z}$ et $f(z) = \overline{f(z)}$ car $f(z) \in \mathbb{R}$. Donc, pour tout $z \in \mathbb{R}$:

$$g(z) = g(\bar{z}) = \overline{f(z)} = f(z).$$

La fonction analytique $f - g$ s'annule donc sur $\mathbb{R}$. Ses zéros ne sont pas isolés. Comme $\mathbb{C}$ est connexe, $f - g$ est identiquement nulle, ce qui démontre le résultat demandé.

3 Logarithme complexe et racines

3.1 Logarithme complexe

Rappelons d'abord brièvement les propriétés de la fonction $\exp : \mathbb{C} \to \mathbb{C}$. Ces propriétés ont été, pour l'essentiel, établies dans les modules II.5 et IV.3 de [L1] et ont été précisées ici dans la section 3.3 du module II.2.
La fonction $\exp : \mathbb{C} \to \mathbb{C}$ est définie comme la somme d'une série entière de rayon de convergence infini :

$$\exp(z) := \sum_{n=0}^{+\infty} \frac{z^n}{n!}.$$

Elle est donc analytique sur $\mathbb{C}$. Elle y est donc $\mathbb{C}$-dérivable, et en dérivant terme à terme la série entière, on obtient :

$$\exp'(z) = \exp(z).$$

En utilisant le produit de Cauchy de deux séries entières, on montre que :

$$\exp(a + b) = (\exp a)(\exp b).$$

En particulier, pour tout $z \in \mathbb{C}$,

$$\exp(z) \neq 0 \quad \text{et} \quad \exp(-z) = \frac{1}{\exp(z)}.$$

Les coefficients de la série entière étant réels, on voit que :

$$\exp(\bar{z}) = \overline{\exp z}$$

où $\bar{z}$ est le conjugué complexe de z et $\overline{\exp z}$ celui de $\exp z$. De là, on déduit que pour tout réel x, le nombre $\exp(ix)$ est de module 1 :

$$\exp(ix)\overline{\exp(ix)} = \exp(ix)\exp(-ix) = \exp(0) = 1.$$

Cela permet également de montrer que les fonctions $\cos : \mathbb{C} \to \mathbb{C}$ et $\sin : \mathbb{C} \to \mathbb{C}$ définies par :

$$\cos(z) := \frac{\exp(iz) + \exp(-iz)}{2} \quad \text{et} \quad \sin(z) := \frac{\exp(iz) - \exp(-iz)}{2i}$$

prennent des valeurs réelles quand $z \in \mathbb{R}$. On montre alors qu'il existe un plus petit réel, noté $\pi/2$, vérifiant :

$$\cos\frac{\pi}{2} = 0.$$

C'est cet argument (ou un argument similaire) qui permet de définir π.
La fonction $x \mapsto \exp(ix)$ est un morphisme surjectif du groupe additif $\mathbb{R}$ dans le groupe multiplicatif $\mathbb{U}$, de noyau $2\pi\mathbb{Z}$. La fonction $\exp : \mathbb{C} \to \mathbb{C}^*$ est un morphisme surjectif du groupe additif $\mathbb{C}$ dans le groupe multiplicatif $\mathbb{C}^*$, de noyau $i2\pi\mathbb{Z}$. En particulier, pour tout $w_0 \in \mathbb{C}^*$, l'équation $\exp(z) = w_0$ a une infinité de solutions, et deux telles solutions diffèrent d'un multiple entier de $i2\pi$.

Définition 7. Étant donné $z \in \mathbb{C}^*$, un nombre complexe w est un *logarithme* de z si $\exp(w) = z$. Un nombre $\theta \in \mathbb{R}$ est un *argument* de z si $\exp(i\theta) = \dfrac{z}{|z|} \in \mathbb{U}$.

Tout nombre complexe non nul admet donc une infinité de logarithmes et d'arguments. Deux logartihmes de z diffèrent d'un multiple entier de $i2\pi$ et deux arguments de z diffèrent d'un multiple entier de 2π.

Définition 8 (Détermination du logarithme). Si $X \subset \mathbb{C}^*$, une *détermination du logarithme* sur X est une fonction $f : X \to \mathbb{C}$ vérifiant $\exp\big(f(z)\big) = z$ pour tout $z \in X$. Une *détermination de l'argument* est une fonction $\theta : X \to \mathbb{R}$ telle que $\exp\big(i\theta(z)\big) = \dfrac{z}{|z|}$ pour tout $z \in X$.

Les déterminations du logarithme et de l'argument sont intimement liées.
En effet, si $\theta : X \to \mathbb{R}$ est une détermination de l'argument, alors $f : z \mapsto \ln\big(|z|\big) + i\theta(z)$ est une détermination du logarithme et si $f : X \to \mathbb{C}$ est une détermination du logarithme, alors $z \mapsto \operatorname{Im}\big(f(z)\big)$ est une détermination de l'argument. De plus, pour tout $z \in X$, $\operatorname{Re}\big(f(z)\big) = \ln\big(|z|\big)$.

Définition 9. Les *déterminations principales du logarithme et de l'argument* sont les fonctions Log et Arg définies sur $\mathbb{C}^*$ de la manière suivante :

- $\operatorname{Arg}(z)$ est l'unique argument de z appartenant à l'intervalle $]-\pi, \pi]$
- $\operatorname{Log}(z) := \ln|z| + i\operatorname{Arg}(z)$.

Les déterminations principales du logarithme et de l'argument sont discontinues le long de la demi-droite $]-\infty,0[$. En effet, si $z_0 \in]-\infty,0[$, l'argument de z_0 (ou la partie imaginaire de $\mathrm{Log}(z_0)$) est égal à π, tandis que lorsque z tend vers z_0 avec une partie imaginaire strictement négative, $\mathrm{Arg}(z) = \mathrm{Im}\big(\mathrm{Log}(z)\big)$ tend vers $-\pi$.
Il y a des parties de $\mathbb{C}^*$ sur lesquelles on peut définir une détermination continue du logarithme ou de l'argument et d'autres sur lesquelles il n'existe pas de détermination continue.

Exemple. Les restrictions de Log et Arg à $X := \mathbb{C}\setminus\mathbb{R}_-$ sont des déterminations continues du logarithme et de l'argument.

Proposition 17. Il n'existe pas de détermination continue de l'argument sur $\mathbb{U}$.

Démonstration. Soit $\theta : \mathbb{U} \to \mathbb{C}$ une détermination de l'argument sur $\mathbb{U}$. Soit $f : \mathbb{U} \to \mathbb{C}$ la fonction définie par $f(z) = \theta(z) + \theta(\bar{z})$. Comme $\theta(z)$ est un argument de z et $\theta(\bar{z})$ est un argument de $\bar{z}$, pour tout $z \in \mathbb{U}$, $f(z)$ est un multiple entier de 2π. Comme $f(\mathbb{U})$ est un connexe de $\mathbb{R}$, c'est un intervalle. Comme il est contenu dans $2\pi\mathbb{Z}$, c'est un singleton. Donc, la fonction f serait constante sur $\mathbb{U}$.
Cependant, $f(1) = 2\theta(1) = 4k\pi$ avec $k \in \mathbb{Z}$, tandis que $f(-1) = 2\theta(-1) = 2\pi + 4k'\pi$ avec $k' \in \mathbb{Z}$. On a donc $f(1) \neq f(-1)$, ce qui montre que θ ne peut pas être continue. ∎

Corollaire 18. Il n'existe pas de détermination continue du logarithme sur $\mathbb{U}$.

Il n'existe *a fortiori* pas de détermination continue du logarithme sur $\mathbb{C}^*$. Nous n'étudierons pas ici les conditions auxquelles doit satisfaire X pour que l'on puisse définir une détermination continue du logarithme sur X.

Proposition 19. Soit $U \subset \mathbb{C}$ un domaine de $\mathbb{C}^*$. Supposons qu'il existe une détermination continue du logarithme $f : U \to \mathbb{C}$. Alors les déterminations du logarithmes sur U sont exactement les fonctions de la forme $f_k := f + i2k\pi$ avec $k \in \mathbb{Z}$.

Démonstration. Supposons que f et g soient deux déterminations continues du logarithme sur U. Alors, $f-g$ est une fonction continue sur U, à valeurs dans $i2\pi\mathbb{Z}$. Comme U est connexe, une telle fonction est nécessairement constante sur U. On a donc $f-g = i2\pi k$ avec $k \in \mathbb{Z}$.
Le fait que $f + i2\pi k$ soit une détermination continue du logarithme est évident. ∎

Proposition 20. Pour tout $z \in D(0,1)$:

$$\mathrm{Log}(1+z) = \sum_{n=1}^{+\infty} (-1)^{n-1} \frac{z^n}{n}\cdot$$

En particulier, la fonction Log est analytique sur $D(1,1)$.

Démonstration. Le rayon de convergence de la série entière $\displaystyle\sum_{n\geqslant 1} (-1)^{n-1}\frac{z^n}{n}$ est égal à 1.

La somme f de la série est donc analytique sur $D(0,1)$. Elle y est $\mathbb{C}$-dérivable et :

$$f'(z) = \sum_{n=1}^{+\infty} (-1)^{n-1} z^{n-1} = \frac{1}{1+z}\cdot$$

Par conséquent, pour tout $x \in]-1,1[$, on a :

$$f(x) = f(x) - f(0) = \int_0^x \frac{1}{1+t}\,\mathrm{d}t = \ln(1+x).$$

On voit donc que :

$$\forall x \in]-1,1[\quad \exp\circ f(x) = 1+x.$$

La fonction $\exp\circ f : D(0,1) \to \mathbb{C}$ est analytique sur $D(0,1)$. Elle coïncide sur $]-1,1[$ avec la fonction $z \mapsto 1+z$ qui est également analytique sur $D(0,1)$. D'après le théorème du prolongement analytique, ces deux fonctions sont égales sur $D(0,1)$. On a donc :

$$\forall z \in D(0,1) \quad \exp\circ f(z) = 1+z.$$

La fonction $u \mapsto f(u-1)$ est donc une détermination continue du logarithme sur $D(1,1)$. La fonction $u \mapsto \mathrm{Log}(u)$ en est une autre. D'après la proposition 19 de la page ci-contre, ces deux déterminations continues du logarithme diffèrent d'un multiple entier de $i2\pi$. Comme elles prennent la même valeur en $u=1$, elles sont égales :

$$\forall u \in D(1,1) \quad \mathrm{Log}(u) = f(u-1) = \sum_{n=1}^{+\infty} (-1)^{n-1} \frac{(u-1)^n}{n}\cdot$$

Exercice 6.

Montrer que $z \mapsto \mathrm{Log}\dfrac{z+1}{z-1}$ définit une fonction analytique sur $\mathbb{C} \setminus [-1,1]$.

Solution. Observons que l'application $z \mapsto (z+1)/(z-1)$ est une homographie qui envoie $\mathbb{C} \setminus [-1,1]$ dans $\mathbb{C} \setminus]-\infty,0]$. En effet :

$$\frac{z+1}{z-1} = -t \in]-\infty,0] \quad \Longleftrightarrow \quad z = \frac{t-1}{t+1} \in [-1,1[\,.$$

Par conséquent, la fonction $z \mapsto \mathrm{Log}\dfrac{z+1}{z-1}$ est bien définie et analytique sur $\mathbb{C} \setminus [-1,1]$, comme composée de fonctions analytiques.

3.2 Racines $k^{\text{èmes}}$

Dans toute cette partie, on suppose que k est un entier supérieur ou égal à 2.

Définition 10.

On appelle *racine* $k^{\text{ème}}$ de $z \in \mathbb{C}^*$, tout nombre complexe w tel que $w^k = z$.

On a vu dans le module II.5 de [L1] que tout nombre complexe $z \in \mathbb{C}^*$ a exactement k racines $k^{\text{èmes}}$. Si w_0 est l'une d'entre elles, les autres s'obtiennent en multipliant w_0 par les racines $k^{\text{èmes}}$ de 1, *i.e.* les racines $k^{\text{èmes}}$ de z sont :

$w_0,\ w_0{\cdot}\exp(i2\pi{\cdot}1/k),\ w_0{\cdot}\exp(i2\pi{\cdot}2/k),\ w_0{\cdot}\exp(i2\pi{\cdot}3/k),\ \ldots,\ w_0{\cdot}\exp(i2\pi{\cdot}(k-1)/k).$

Définition 11. On appelle *détermination* de $z^{1/k}$ sur un ensemble $X \subset \mathbb{C}^*$ tout application $f : X \to \mathbb{C}$ vérifiant $\big(f(z)\big)^k = z$ pour tout $z \in X$.

Exercice 7.
Montrer qu'il n'existe pas de détermination continue de $z^{1/k}$ sur $\mathbb{C}^*$.

Solution. Comme dans le cas de l'argument ou du logarithme, il suffit de montrer qu'il n'y pas de détermination continue sur $\mathbb{U}$. Supposons par l'absurde que $f : \mathbb{U} \to \mathbb{C}$ soit une détermination continue de $z^{1/k}$. Définissons $g : U \to \mathbb{C}$ par $g(z) = f(z) \cdot f(\bar{z})$. Notons que $\big(g(z)\big)^k = z \cdot \bar{z} = 1$. Par conséquent g prend ses valeurs dans l'ensemble discret des racines $k^{\,ème}$ de 1. Comme $\mathbb{U}$ est connexe, la fonction g est donc constante. Or :

$$g(1) = f(1) \cdot f(1) \quad \text{et} \quad g(-1) = f(-1) \cdot f(-1).$$

Étant donné que $k \geqslant 2$, on a, modulo $4\pi/k$:

$$\operatorname{Arg}\big(g(1)\big) \equiv 2 \operatorname{Arg}\big(f(1)\big) \equiv 0$$

et

$$\operatorname{Arg}\big(g(-1)\big) \equiv 2 \operatorname{Arg}\big(f(-1)\big) \equiv \frac{2\pi}{k}\cdot$$

Les deux nombres ne peuvent pas être égaux. D'où une contradiction.

Si $U \subset \mathbb{C}^*$ est un domaine et si $f : U \to \mathbb{C}$ est une détermination continue du logarithme sur U, on voit que : $z \mapsto \exp\big(f(z)/k\big)$ est une détermination continue de $z^{1/k}$ sur U. Les autres détermination continues s'obtiennent à partir de celle-ci en les multipliant par les racines $k^{\,èmes}$ de 1.

Définition 12. On note $\sqrt{\cdot} : \mathbb{C} \setminus \mathbb{R}_- \to \mathbb{C}$ la détermination de $z^{1/2}$ sur $\mathbb{C} \setminus \mathbb{R}_-$ définie par :

$$\sqrt{z} = \exp\left(\frac{1}{2}\operatorname{Log}(z)\right).$$

Comme $\sqrt{\cdot} : D(1,1) \to \mathbb{C}$ est la composée de deux fonctions analytiques, elle est elle-même analytique. Nous allons maintenant déterminer son développement en série entière en 1.

Exercice 8.
Déterminer le développement en série entière en 0 de la fonction $z \mapsto \sqrt{1+z}$.

Solution. La fonction $f : z \mapsto \sqrt{1+z}$ est analytique sur $D(0,1)$. Son développement en 0 est donné par sa série de Taylor en 0. On montre par récurrence que pour tout entier $n \geqslant 1$ et tout $z \in D(0,1)$, on a :

$$f^{(n)}(z) = \frac{1}{2} \cdot \left(\frac{1}{2} - 1\right) \cdots \left(\frac{1}{2} - n + 1\right) \exp\left(\left(\frac{1}{2} - n\right) \operatorname{Log}(1+z)\right).$$

Par conséquent :

$$\sqrt{1+z} = \sum_{n=0}^{+\infty} \binom{1/2}{n} z^n \quad \text{avec} \quad \binom{1/2}{n} = \frac{1/2 \cdot (1/2-1) \cdots (1/2-n+1)}{n!}.$$

3.3 Théorème d'inversion locale

Un résultat important qui n'a pas été abordé dans le module II.2, même dans le cadre des séries entières d'une variable réelle, est le théorème d'inversion locale.

Nous avons vu que toute fonction f analytique sur un ouvert U y est holomorphe. De plus, f' est continue. Le théorème d'inversion locale pour les fonctions holomorphes (proposition 6 de la page 709) affirme que si $z_0 \in U$ est un point tel que $f'(z_0) \neq 0$, il existe une fonction holomorphe g définie au voisinage de $f(z_0)$, telle que $g \circ f = \mathrm{Id}$ et $f \circ g = \mathrm{Id}$. De plus g' est continue. Comme signalé plus haut, nous verrons qu'il s'ensuit que g est une fonction analytique (corollaire 29 de la page 787).

Nous allons dès à présent présenter une preuve uniquement basée sur le développement en série entière.

Théorème 21 (Inversion locale). Soit f une fonction développable en série entière en z_0. Supposons que $f'(z_0) \neq 0$. Alors, il existe un voisinage U de z_0, un voisinage V de $f(z_0)$ et une bijection $g : V \to U$ telle que :

- $g\big(f(z_0)\big) = z_0$,
- g est une fonction analytique,
- pour tout $z \in U$, $g \circ f(z) = z$ et
- pour tout $w \in V$, $f \circ g(w) = w$.

La démonstration que nous allons présenter est d'un niveau soutenu et le lecteur est invité à la laisser de côté dans un premier temps.

Démonstration. Par hypothèse il existe une série entière $F := \sum_{n \geqslant 0} a_n z^n$ avec $a_1 \neq 0$, telle que pour $|z - z_0|$ suffisamment petit, on ait :

$$f(z) = \sum_{n=0}^{+\infty} a_n (z - z_0)^n.$$

Lemme 22. On peut se ramener au cas $z_0 = f(z_0) = 0$ et $f(z) = z + \sum_{n=2}^{+\infty} a_n z^n$ avec pour tout $n \geqslant 2$, $|a_n| \leqslant 1$.

Démonstration. Étant donné $\lambda > 0$, posons :

$$f_\lambda(z) := \frac{f(z_0 + \lambda z) - a_0}{\lambda a_1}.$$

D'une part, pour $|z|$ suffisamment petit, on a :

$$f_\lambda(z) = z + \sum_{n=2}^{+\infty} \frac{a_n \lambda^{n-1}}{a_1} z^n.$$

D'autre part, si g_λ est un inverse local de f_λ envoyant 0 sur 0, alors

$$g : w \mapsto z_0 + g_\lambda\left(\frac{w - a_0}{\lambda a_1}\right)$$

est un inverse local de f envoyant $f(z_0)$ sur z_0. Bien entendu, si g_λ est analytique, g l'est également.

Enfin, le rayon de convergence de la série entière F étant non nul, si $r > 0$ est suffisamment petit, la suite $\big(|a_n| r^n\big)_{n \geqslant 1}$ tend vers 0, et est donc bornée, disons par M. Si l'on pose $\lambda := |a_1| r^2 / M \leqslant r$, alors, pour tout $n \geqslant 2$:

$$\left|\frac{a_n \lambda^{n-1}}{a_1}\right| = \frac{|a_n| r^n}{|a_1| r} \cdot \left(\frac{\lambda}{r}\right)^{n-1} \leqslant \frac{M}{|a_1| r} \cdot \frac{\lambda}{r} = 1.$$

Nous supposons donc que $F = z + \sum\limits_{n \geqslant 2} a_n z^n$ avec $|a_n| \leqslant 1$. Étant donné que $F'(0) \neq 0$, il existe une série formelle réciproque $G := w + \sum\limits_{n \geqslant 2} b_n w^n$ telle que $F \circ G = \mathrm{I}$ avec I la série entière définie par $I(z) = z$ (proposition 67 de la page 45).

Lemme 23. Le rayon de convergence de G est non nul.

Démonstration. Pour tout entier $p \geqslant 2$, on note $Q_p \in \mathbb{C}[a_2, \ldots, a_p, b_2, \ldots, b_{p-1}]$ le polynôme à coefficients entiers positifs tel que $Q_p(a_2, \ldots, a_p, b_2, \ldots, b_{p-1})$ soit le coefficient de w^p dans l'expression :

$$a_2 \cdot (w + b_2 w^2 + \cdots + b_{p-1} w^{p-1})^2 + \cdots + a_p \cdot (w + b_2 w^2 + \cdots + b_{p-1} w^{p-1})^p.$$

Ainsi :

$$Q_2(a_2) = a_2, \quad Q_3(a_2, a_3, b_2) = 2a_2 b_2 + a_3, \quad \ldots.$$

On voit alors que pour $p \geqslant 2$, le coefficient de w^p dans :

$$G(w) + a_2 \cdot \big(G(w)\big)^2 + \cdots + a_p \cdot \big(G(w)\big)^p$$

est nul si, et seulement si :

$$b_p + Q_p(a_2, \ldots, a_p, b_2, \ldots, b_{p-1}) = 0.$$

Cela permet de définir les coefficients b_p par récurrence :

$$b_2 := -Q_2(a_2) = -a_2, \quad b_3 := -Q_3(a_2, a_3, b_2) = 2a_2^2 - a_3, \quad \ldots$$

$$b_p := -Q_p(a_2, \ldots, a_p, b_2, \ldots, b_{p-1}).$$

Il nous faut maintenant montrer que le rayon de convergence de la série entière G ainsi définie est non nul. Définissons par récurrence :

$$B_2 := Q_2(1), \quad B_3 := Q_3(1,1,B_2), \quad \dots, \quad B_p := Q_p(1,\dots,1,B_2,\dots,B_{p-1}).$$

Étant donné que les polynômes Q_p sont à coefficients entiers et positifs et que $|a_n| \leqslant 1$ pour tout $n \geqslant 2$, on obtient par récurrence que $|b_p| \leqslant B_p$ pour tout $p \geqslant 2$. Il ne reste plus qu'à montrer que le rayon de convergence de la série $\sum_{p\geqslant 2} B_p w^p$ est non nul. Posons :

$$\overline{F} := z + \sum_{n\geqslant 2} A_n z^n \quad \text{avec} \quad A_2 = A_3 = A_4 = \cdots = -1.$$

Le rayon de convergence de cette série est égal à 1, et la somme $\overline{f} : D(0,1) \to \mathbb{C}$ vaut :

$$\overline{f}(z) = z - \frac{z^2}{1-z} = \frac{z-2z^2}{1-z}.$$

Cette fonction est localement inversible et son inverse est donné en résolvant une équation de degré 2 :

$$\frac{z-2z^2}{1-z} = w \quad \Longleftrightarrow \quad 2z^2 - (w+1)z + w = 0.$$

Pour $|w|$ suffisamment petit pour que $w(w-6) \in D(0,1)$, la fonction réciproque est donnée par :

$$\overline{g} : w \mapsto z = \frac{1+w-\sqrt{(1+w)^2-8w}}{4} = \frac{1+w-\sqrt{1+w(w-6)}}{4}.$$

Cette dernière fonction est développable en série entière en $w = 0$. Comme $\overline{f} \circ \overline{g} = \mathrm{Id}$, le développement en série entière $\overline{G}$ satisfait $\overline{F} \circ \overline{G} = \mathrm{I}$. Or, pour tout $p \geqslant 2$:

$$B_p + Q_p(A_2,\dots,A_p,B_2,\dots,B_{p-1}) = 0.$$

Par conséquent :

$$\overline{G} = w + \sum_{p\geqslant 2} B_p w^p,$$

ce qui montre que le rayon de convergence de la série $\sum_{p\geqslant 2} B_p w^p$ est non nul. ∎

Nous pouvons maintenant conclure la démonstration du théorème d'inversion locale.
Comme $F \circ G = \mathrm{I}$ et comme les rayons de convergences des séries f et g sont non nuls, la proposition 39 de la page 513 montre que, pour w de module suffisamment petit, on a :

$$f \circ g(w) = w \quad \text{avec} \quad g(w) := w + \sum_{n=2}^{+\infty} b_n w^n.$$

En appliquant le même raisonnement à la fonction g que celui appliqué à la fonction f, on voit qu'il existe une fonction h analytique au voisinage de 0 telle que, pour tout z de module suffisamment petit, on a :

$$g \circ h(z) = z.$$

Or, lorsque le module de z est petit, celui de $h(z)$ est petit et on peut écrire :

$$h(z) = \mathrm{Id} \circ h(z) = (f \circ g) \circ h(z) = f \circ (g \circ h)(z) = f(z).$$

Le théorème d'inversion locale en découle en choisissant par exemple $U := D(0,r)$ et $V := f\big(D(0,r)\big)$ avec r suffisamment petit. ∎

3.4 Analyticité du logarithme complexe et des racines $k^{èmes}$

Proposition 24. Soit U un domaine de $\mathbb{C}^*$. Supposons qu'il existe une détermination continue du logarithme $f : U \to \mathbb{C}$. Alors :

1. La fonction f est analytique sur U, en particulier $\mathbb{C}$-dérivable sur U.
2. La dérivée de f sur U est $f' : z \mapsto 1/z$.

Démonstration.

1. Soit f une détermination continue du logarithme sur U. Soit $z_0 \in U$ et $w_0 := f(z_0)$. Alors, $\exp(w_0) = z_0$ et $\exp\big(f(z)\big) = z$. Comme $\exp'(w_0) = \exp(w_0) \neq 0$, le théorème d'inversion locale de la page 779 affirme qu'il existe une fonction analytique g définie dans un voisinage de z_0 telle que $g(z_0) = w_0$ et $g\big(\exp(w)\big) = w$ au voisinage de w_0. Comme f est continue en z_0, pour $|z - z_0|$ suffisamment petit, $\big|f(z) - w_0\big|$ est petit et $g(z) = g\Big(\exp\big(f(z)\big)\Big) = (g \circ \exp)\big(f(z)\big) = f(z)$.

   La fonction f est donc analytique au voisinage de z_0. Comme cela est vrai pour tout $z_0 \in U$, la fonction f est analytique sur U.

2. Étant donné $z_0 \in U$, si h est suffisamment petit, on a :
$$\frac{f(z_0+h) - f(z_0)}{h} = \frac{f(z_0+h) - f(z_0)}{\exp\big(f(z_0+h)\big) - \exp\big(f(z_0)\big)}.$$
   Lorsque h tend vers 0, cette quantité a la même limite que :
$$\lim_{w \to w_0} \frac{w - w_0}{\exp(w) - \exp(w_0)}.$$
   Cette limite est égale à $\dfrac{1}{\exp'(w_0)} = \dfrac{1}{\exp(w_0)} = \dfrac{1}{z_0}$. ∎

Définition 13. Étant donné $\theta \in \mathbb{R}$, on appelle L_θ la demi-droite
$$L_\theta := \big\{ r \exp(i\theta) \;\big|\; r \in [0, +\infty[\big\}.$$
Pour $z \in \mathbb{C} \setminus L_\theta$, on note $\mathrm{Arg}_\theta(z)$ l'unique argument de z appartenant à l'intervalle $]\theta, 2\pi + \theta[$. On pose :
$$\mathrm{Log}_\theta(z) = \log|z| + i\,\mathrm{Arg}_\theta(z).$$

Il est facile de voir que pour tout $\theta \in \mathbb{R}$, la fonction $\mathrm{Log}_\theta : \mathbb{C} \setminus L_\theta \to \mathbb{C}$ est une détermination continue, donc analytique, du logarithme sur $\mathbb{C} \setminus L_\theta$.

Notons que la fonction $\mathrm{Log} : D(1,1) \to \mathbb{C}$ n'a pas de prolongement analytique maximal $\tilde{f} : \widetilde{U} \to \mathbb{C}$ au sens où tout prolongement $g : U \to \mathbb{C}$ avec $D(1,1) \subset U$ en serait une restriction. En effet, pour tout $\theta \in [-3\pi/2, -\pi/2]$, la fonction $\mathrm{Log}_\theta : \mathbb{C} \setminus L_\theta \to \mathbb{C}$ est un prolongement de la fonction $\mathrm{Log} : D(1,1) \to \mathbb{C}$. Mais si $\theta \neq \theta'$, les fonctions Log_θ et $\mathrm{Log}_{\theta'}$ ne sont pas restriction d'un prolongement commun puisqu'on peut toujours trouver des points auxquels les valeurs des deux fonctions diffèrent.

Proposition 25. Si $U \subset \mathbb{C}^*$ est un domaine, toute détermination continue de $z^{1/k}$ sur U est une fonction analytique sur U.

Démonstration. Comme dans le cas des déterminations continues du logarithme complexe, c'est une conséquence directe du théorème d'inversion locale.

Si $\theta \in \mathbb{R}$, on peut définir une détermination analytique de $z^{1/k}$ sur $\mathbb{C} \setminus L_\theta$ en posant :

$$z^{1/k} := \exp\bigl(\mathrm{Log}_\theta(z)/k\bigr).$$

En multipliant par les racines $k^{\text{èmes}}$ de 1, on montre qu'il y a exactement k déterminations $z^{1/k}$ analytiques sur $\mathbb{C} \setminus L_\theta$. En effet, le rapport de deux déterminations continues est une fonction continue sur $\mathbb{C} \setminus L_\theta$ qui prend ses valeurs dans les racines $k^{\text{èmes}}$ de 1. Comme $\mathbb{C} \setminus L_\theta$ est connexe, une telle fonction est nécessairement constante.

4 Séries entières et séries de Fourier

Il existe un lien étroit entre la théorie des fonctions analytiques et la théorie des séries de Fourier. Nous allons exploiter ce lien pour montrer, entre autres choses, qu'une fonction holomorphe sur un disque $D(z_0, R)$, ayant une dérivée continue sur $D(z_0, R)$, est développable en série entière en z_0 avec rayon de convergence supérieur ou égal à R.

4.1 Formule de Cauchy pour un cercle

Observons tout d'abord que si $f : D(0, R) \to \mathbb{C}$ définie par :

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

est la somme d'une série entière de rayon de convergence R non nul, alors pour $r \in [0, R[$, la fonction $F : \mathbb{R} \to \mathbb{C}$ définie par :

$$F(t) := f(re^{i2\pi t}) = \sum_{n=0}^{+\infty} a_n r^n e^{i2\pi nt}$$

est une fonction 1-périodique dont la série de Fourier converge normalement vers F. Si l'on appelle $c_n(F)$ les coefficients de Fourier de F, les résultats établis dans le module II.4 montrent que l'on a :

$$\int_0^1 f(re^{i2\pi t}) \cdot e^{-i2\pi nt}\,\mathrm{d}t = c_n(F) = \begin{cases} a_n r^n & \text{si } n \geqslant 0 \\ 0 & \text{si } n < 0. \end{cases}$$

Théorème 26.
Si $f : D(0, R) \to \mathbb{C}$ est définie par $f(z) := \sum_{n=0}^{+\infty} a_n z^n$, alors pour $r \in [0, R[$:

$$\int_0^1 \frac{re^{i2\pi t}}{re^{i2\pi t} - z} f(re^{i2\pi t})\,\mathrm{d}t = \begin{cases} f(z) & \text{si } |z| < r \\ 0 & \text{si } |z| > r. \end{cases}$$

Démonstration. Commençons par traiter le cas $|z| < r$. Dans ce cas, $\left|\dfrac{z}{re^{i2\pi t}}\right| < 1$ et :

$$\frac{re^{i2\pi t}}{re^{i2\pi t} - z} = \frac{1}{1 - \dfrac{z}{re^{i2\pi t}}} = \sum_{n=0}^{+\infty} \frac{z^n}{r^n} e^{-i2\pi nt}.$$

La fonction $t \mapsto f(re^{i2\pi t})$ est continue, donc bornée sur le compact $[0,1]$. La série :

$$\sum_{n=0}^{+\infty} \frac{z^n}{r^n} e^{-i2\pi nt} f(re^{i2\pi t})$$

converge donc normalement sur $[0,1]$. Par conséquent, on peut échanger série et intégrale :

$$\begin{aligned}\int_0^1 \frac{re^{i2\pi t}}{re^{i2\pi t} - z} f(re^{i2\pi t})\,\mathrm{d}t &= \sum_{n=0}^{+\infty} \frac{z^n}{r^n} \left(\int_0^1 f(re^{i2\pi t}) \cdot e^{-i2\pi nt}\,\mathrm{d}t \right) \\ &= \sum_{n=0}^{+\infty} \frac{c_n(F)}{r^n} z^n = \sum_{n=0}^{+\infty} a_n z^n f(z).\end{aligned}$$

Si $|z| > r$, on écrit :

$$\frac{re^{i2\pi t}}{re^{i2\pi t} - z} = -\frac{re^{i2\pi t}}{z} \cdot \sum_{n=0}^{+\infty} \frac{r^n}{z^n} e^{i2\pi nt}.$$

Cette fois, on obtient :

$$\begin{aligned}\int_0^1 \frac{re^{i2\pi t}}{re^{i2\pi t} - z} f(re^{i2\pi t})\,\mathrm{d}t &= \sum_{n=0}^{+\infty} -\frac{r^n}{z^n} \left(\int_0^1 f(re^{i2\pi t}) \cdot e^{i2\pi nt}\,\mathrm{d}t \right) \\ &= -\sum_{n=0}^{+\infty} c_{-n}(F) \frac{r^n}{z^n} = 0.\end{aligned}$$

Il est possible de justifier que l'on peut dériver la relation :

$$f(z) = \int_0^1 \frac{re^{i2\pi t}}{re^{i2\pi t} - z} f(re^{i2\pi t})\,\mathrm{d}t$$

sous le signe intégral pour obtenir une formule intégrale pour les dérivées de f. Nous proposons l'alternative suivante.

Théorème 27 (Formules de Cauchy). Si $f : D(z_0, R) \to \mathbb{C}$ est analytique, alors pour $r \in [0, R[$, tout $z \in D(z_0, r)$ et tout entier $k \geqslant 0$:

$$\frac{f^{(k)}(z)}{k!} = \int_0^1 \frac{re^{i2\pi t}}{\left(z_0 + re^{i2\pi t} - z\right)^{k+1}} f(z_0 + re^{i2\pi t})\,\mathrm{d}t.$$

Démonstration. Quitte à considérer la fonction $w \mapsto f(z_0 + w)$, on peut supposer sans perte de généralité que $z_0 = 0$.

Pour $\zeta \in D(0,1)$, on a la relation (voir l'exercice 1 de la page 765) :

$$\frac{k!}{(1-\zeta)^{k+1}} = \sum_{n=k}^{+\infty} \frac{n!}{(n-k)!}\zeta^{n-k}.$$

De même, on peut dériver k fois terme à terme la série entière définissant f, ce qui donne, comme dans la preuve précédente :

$$\begin{aligned} f^{(k)}(z) = \sum_{n=k}^{+\infty} \frac{n!}{(n-k)!} a_n z^{n-k} &= \sum_{n=k}^{+\infty} \frac{n!}{(n-k)!} \frac{z^{n-k}}{r^n} \left(\int_0^1 f(re^{i2\pi t}) \cdot e^{-i2\pi nt} \, \mathrm{d}t \right) \\ &= \int_0^1 \left(\sum_{n=k}^{+\infty} \frac{n!}{(n-k)!} \left(\frac{z}{re^{i2\pi t}} \right)^{n-k} \right) \frac{f(re^{i2\pi t})}{\left(re^{i2\pi t}\right)^k} \, \mathrm{d}t \\ &= \int_0^1 \frac{k!}{\left(1 - \frac{z}{re^{i2\pi t}}\right)^{k+1}} \frac{f(re^{i2\pi t})}{\left(re^{i2\pi t}\right)^k} \, \mathrm{d}t \\ &= k! \int_0^1 \frac{re^{i2\pi t}}{(re^{i2\pi t} - z)^{k+1}} f(re^{i2\pi t}) \, \mathrm{d}t. \end{aligned}$$

4.2 Analyticité des fonctions holomorphes

Pour établir les formules de Cauchy, nous avons fait l'hypothèse que f est développable en série entière sur le disque $D(0,R)$. Nous allons maintenant renverser la procédure pour montrer que si f est holomorphe sur le disque $D(0,R)$ et si la dérivée est continue, alors f est automatiquement développable en série entière en 0 avec rayon de convergence supérieur ou égal à R. En particulier, si f est analytique sur $D(0,R)$, elle est somme de sa série de Taylor en 0 sur $D(0,R)$.

Théorème 28. Soit $f : D(0,R)$ une fonction holomorphe ayant une dérivée continue sur $D(0,R)$. Pour $r \in [0,R[$, posons :

$$c_n(r) := \int_0^1 f(re^{i2\pi t}) e^{-i2\pi nt} \, \mathrm{d}t.$$

Alors, il existe $a_n \in \mathbb{C}$ tel que pour tout $r \in [0,R[$, $c_n(r) = a_n r^n$. De plus, pour tout $z \in D(0,R)$:

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

Démonstration. Soit $f : D(0,R) \to \mathbb{C}$ une fonction holomorphe sur $D(0,R)$ et soit $g : [0,R[\times \mathbb{R} \to \mathbb{C}$ la fonction définie par :

$$g(r,t) := f(re^{i2\pi t}).$$

Pour tout $r \in [0,R[$, la fonction $g_r : t \mapsto g(r,t)$ est continue et 1-périodique. Pour $n \in \mathbb{Z}$,

on note $c_n(r)$ le $n^{ème}$ coefficient de Fourier de la fonction g_r :

$$c_n(r) := \int_0^1 g(r,t)e^{-i2\pi nt}\,\mathrm{d}t = \int_0^1 f(re^{i2\pi t})e^{-i2\pi nt}\,\mathrm{d}t.$$

Nous allons montrer que lorsque la dérivée de f est continue sur $D(0,R)$:
- pour $n<0$ et pour $r\in[0,R[$, $c_n(r)=0$ et
- pour $n\geqslant 0$, il existe $a_n\in\mathbb{C}$ tel que pour tout $r\in[0,R[$, $c_n(r)=a_nr^n$.

Lorsque la dérivée de f est continue sur $D(0,R)$, pour tout $r\in[0,r[$, la fonction g_r est de classe $\mathcal{C}^1$. D'après le théorème de Dirichlet (théorème 18 de la page 675), g_r est la somme de sa série de Fourier :

$$f(re^{i2\pi t}) = g_r(t) = \sum_{n=-\infty}^{+\infty} c_n(r)e^{i2\pi nt} = \sum_{n=0}^{+\infty} a_n(re^{i2\pi t})^n.$$

Cela conclura la démonstration.

Montrons donc que pour tout $n\in\mathbb{Z}$, $c_n(r)=a_nr^n$ avec $a_n\in\mathbb{C}$ qui ne dépend pas de r. Pour cela, montrons que c_n est solution d'une équation différentielle. L'hypothèse que f' est continue garantit que les dérivées partielles :

$$\frac{\partial g}{\partial r}(r,t) = e^{i2\pi t}f'(re^{i2\pi t}) \quad\text{et}\quad \frac{\partial g}{\partial t}(r,t) = i2\pi re^{i2\pi t}f'(re^{i2\pi t})$$

sont continues sur $]0,R[\times\mathbb{R}$. La fonction c_n est donc dérivable et l'on a :

$$\begin{aligned} c_n'(r) &= \int_0^1 \frac{\partial g}{\partial r}(r,t)e^{-i2\pi nt}\,\mathrm{d}t \\ &= \int_0^1 e^{i2\pi t}f'(re^{i2\pi t})e^{-i2\pi nt}\,\mathrm{d}t \\ &= \int_0^1 \frac{1}{i2\pi r}\frac{\partial g}{\partial t}(r,t)e^{-i2\pi nt}\,\mathrm{d}t. \end{aligned}$$

Une intégration par parties donne alors :

$$c_n'(r) = \frac{1}{i2\pi r}\left(\underbrace{\big[g(r,t)e^{-i2\pi nt}\big]_0^1}_{=0} + i2\pi n\int_0^1 g(r,t)e^{-i2\pi nt}\,\mathrm{d}t\right) = \frac{n}{r}c_n(r).$$

Si l'on pose $a_n(r) := c_n(r)/r^n$ pour $r>0$, on voit alors que $a_n'(r)=0$ et donc, que la fonction a_n est constante sur $]0,R[$.

Si l'on fixe $r'\in]0,R[$, pour tout $r\in]0,r'[$, on a :

$$\big|c_n(r)\big| \leqslant \int_0^1 \big|g(r,t)\big|\,\mathrm{d}t \leqslant \sup_{|z|\leqslant r'}\big|f(z)\big|.$$

Lorsque r tend vers 0, le coefficient $c_n(r)$ reste borné, ce qui montre que $a_n=0$ pour tout $n<0$.

Corollaire 29. Soit f une fonction holomorphe dans un ouvert U de $\mathbb{C}$ dont la dérivée f' est continue dans U. Alors f est analytique dans U et, pour tout $z_0 \in U$, la série de Taylor de f en z_0 converge et représente f dans le plus grand disque ouvert de centre z_0 contenu dans U.

Nous verrons dans [MPA] que la condition que f' est continue dans U est superflue : elle est automatiquement réalisée si f est holomorphe dans U. Autrement dit, les fonctions holomorphes sur U sont exactement les fonctions analytiques sur U.

Démonstration. Soit $z_0 \in U$ et $R > 0$ tels que $D(z_0, R) \subset U$. La fonction $w \mapsto f(z_0 + w)$ est holomorphe sur $D(0, R)$ et sa dérivée $w \mapsto f'(z_0 + w)$ est continue sur $D(0, R)$. D'après le théorème précédent, il existe une série entière $F := \sum_{n \geqslant 0} a_n w^n$ telle que pour tout $w \in D(0, R)$, on ait $f(z_0 + w) = \sum_{n=0}^{+\infty} a_n w^n$. ∎

Corollaire 30. Soit f une fonction analytique dans un ouvert U de $\mathbb{C}$. Alors, pour tout $z_0 \in U$, la série de Taylor de f en z_0 converge et représente f dans le plus grand disque ouvert de centre z_0 contenu dans U.

4.3 Inégalités de Cauchy

Les formules de Cauchy (théorème 27 de la page 784) ont une multitude d'applications. Nous allons maintenant en présenter quelques-unes.

Proposition 31 (Inégalités de Cauchy). Soit $f : D(z_0, R) \to \mathbb{C}$ une fonction analytique telle que pour tout $z \in D(z_0, R)$, $f(z) = \sum_{n=0}^{+\infty} a_n (z - z_0)^n$.

Alors, pour tout $r \in]0, R[$ et tout $n \geqslant 0$, on a :

$$|a_n| = \left| \frac{f^{(n)}(z_0)}{n!} \right| \leqslant \frac{1}{r^n} \sup_{|z - z_0| = r} \big|f(z)\big|.$$

Démonstration. D'après les formules de Cauchy, on a :

$$\begin{aligned} |a_n| = \left| \frac{f^{(n)}(z_0)}{n!} \right| &= \left| \int_0^1 \frac{1}{(re^{i2\pi t})^n} f(z_0 + re^{i2\pi t})\, \mathrm{d}t \right| \\ &\leqslant \frac{1}{r^n} \int_0^1 \big|f(z_0 + re^{i2\pi t})\big|\, \mathrm{d}t \leqslant \frac{1}{r^n} \sup_{|z - z_0| = r} \big|f(z)\big|. \end{aligned}$$

∎

Exercice 9.
Soit f une fonction holomorphe dans $D(0,1)$ telle que :

$$\forall z \in D(0,1), \quad \big|f(z)\big| \leqslant 1.$$

Montrer que pour tout $z_0 \in D(0,1)$ et tout entier $n \geqslant 0$:

$$\big|f^{(n)}(z_0)\big| \leqslant \frac{n!}{\big(1-|z_0|\big)^n}.$$

Solution. Fixons $z_0 \in D(0,1)$ et choisissons $r < 1 - |z_0|$. La fonction f est analytique sur $D\big(z_0, 1-|z_0|\big)$. D'après les inégalités de Cauchy, on a :

$$\left|\frac{f^{(n)}(z_0)}{n!}\right| \leqslant \frac{1}{r^n} \sup_{|z-z_0|=r} \big|f(z)\big| \leqslant \frac{1}{r^n}.$$

On peut alors laisser r tendre vers $1-|z_0|$, ce qui donne le résultat attendu.

Une célèbre application des inégalités de Cauchy est le théorème de Liouville.

Définition 14.
Une fonction $f : \mathbb{C} \to \mathbb{C}$ est une *fonction entière* si f est analytique sur $\mathbb{C}$.

Théorème 32 (Liouville). Toute fonction entière bornée est constante.

Ce théorème est un corollaire du théorème plus général suivant.

Théorème 33. Soit α un réel strictement positif et $f : \mathbb{C} \to \mathbb{C}$ une fonction entière. Si $\big|f(z)\big| = \mathrm{O}\big(|z|^\alpha\big)$, alors f est un polynôme de degré inférieur ou égal à α.

Démonstration. Une fonction entière étant analytique sur $\mathbb{C}$, elle est développable en série entière en 0 avec rayon de convergence infini. On a donc :

$$\forall z \in \mathbb{C}, \quad f(z) := \sum_{n=0}^{+\infty} a_n z^n.$$

Les inégalités de Cauchy montrent que pour tout $n \geqslant 0$ et tout $r > 0$, on a :

$$|a_n| \leqslant \frac{1}{r^n} \sup_{|z|=r} \big|f(z)\big| = \mathrm{O}(r^{\alpha-n}).$$

En faisant tendre r vers $+\infty$, on voit que $a_n = 0$ dès que $\alpha - n < 0$, *i.e.* dès que $n > \alpha$. On a donc :

$$f(z) = \sum_{n=0}^{\lfloor\alpha\rfloor} a_n z^n$$

où $\lfloor\alpha\rfloor$ désigne la partie entière de α. Comme requis, f est un polynôme de degré inférieur ou égal à α. ∎

Une application classique du théorème de Liouville est le théorème de d'Alembert-Gauß, également appelé théorème fondamental de l'algèbre, qui a déjà été démontré dans le module II.5 de [L1].

Théorème 34. Tout polynôme $P : \mathbb{C} \to \mathbb{C}$ non constant admet au moins une racine complexe.

Démonstration. Soit $P : \mathbb{C} \to \mathbb{C}$ un polynôme ne s'annule pas dans $\mathbb{C}$. Alors $f := 1/P : \mathbb{C} \to \mathbb{C}$ est une fonction entière. De plus :

$$\lim_{|z|\to+\infty} \big|f(z)\big| = 0.$$

D'après le théorème de Liouville, f (et donc P) est une fonction constante. ∎

Une autre application du théorème de Liouville est le résultat suivant.

Exercice 10.

Montrer que si f est une fonction entière non constante, alors $f(\mathbb{C})$ est dense dans $\mathbb{C}$.

Solution. Supposons que $f : \mathbb{C} \to \mathbb{C}$ soit une fonction entière mais que $f(\mathbb{C})$ ne soit pas dense dans $\mathbb{C}$. Il existe un ouvert $U \subset \mathbb{C}$ qui n'intersecte pas $f(\mathbb{C})$. Choisissons $w_0 \in U$ et définissons $g : \mathbb{C} \to \mathbb{C}$ par :

$$g(z) = \frac{1}{f(z) - w_0}.$$

La fonction g est une fonction analytique sur $\mathbb{C}$, car c'est la composée des deux fonctions analytiques $f : \mathbb{C} \to \mathbb{C}\setminus\{w_0\}$ et $w \mapsto 1/(w - w_0) : \mathbb{C}\setminus\{w_0\} \to \mathbb{C}$. Comme la fonction f évite un voisinage de w_0, la fonction g évite un voisinage de l'infini. Autrement dit, la fonction g est une fonction entière bornée. D'après le théorème de Liouville, la fonction g est constante. La fonction $f = w_0 + 1/g$ est également constante.

4.4 Principe du maximum

Nous allons maintenant montrer le principe du module maximum, couramment appelé principe du maximum. La démonstration que nous présenterons repose sur la formule de Parseval pour les séries de Fourier (voir le théorème 15 de la page 673 du module II.4).

Théorème 35 (Principe du module maximum). Soit $U \subset \mathbb{C}$ un domaine et $f : U \to \mathbb{C}$ une fonction analytique. Si $|f|$ admet un maximum local, alors f est constante.

Démonstration. Soit $z_0 \in U$ un point tel que $|f|$ admette un maximum local en z_0. Soit $r > 0$ suffisamment petit pour que :

$$\overline{D}(z_0, r) \in U \quad \text{et} \quad \forall z \in \overline{D}(z_0, r), \quad \big|f(z)\big| \leqslant \big|f(z_0)\big|.$$

En appliquant l'égalité de Parseval à la série de Fourier :

$$f(z_0 + re^{i2\pi t}) = \sum_{n=0}^{+\infty} a_n r^n e^{i2\pi nt},$$

on obtient :

$$\sum_{n=0}^{+\infty} |a_n|^2 r^{2n} = \int_0^1 \left|f(z_0 + re^{i2\pi t})\right|^2 \mathrm{d}t.$$

Par hypothèse, on a $\left|f(z_0 + re^{i2\pi t})\right| \leqslant \left|f(z_0)\right|$ et donc :

$$\sum_{n=0}^{+\infty} |a_n|^2 r^{2n} \leqslant \left|f(z_0)\right|^2 = |a_0|^2.$$

On en déduit que $a_n = 0$ pour tout $n \geqslant 1$. La série entière de f au point z_0 se réduit donc à son terme constant, de sorte qu'au voisinage de z_0, f est constante.

D'après le principe du prolongement analytique, U étant connexe, la fonction f est constante dans U. ∎

Corollaire 36. Soit $U \subset \mathbb{C}$ un domaine et $f : U \to \mathbb{C}$ une fonction analytique. Si $|f|$ admet un minimum local non nul, alors f est constante.

Démonstration. Le cas d'un minimum local se ramène au cas du maximum local en considérant la fonction $1/f$. Elle est définie et analytique au voisinage du minimum local de $|f|$. De plus $|1/f|$ y possède un maximum local. Les fonctions $1/f$ et f sont donc localement constantes par le principe du module maximum. Elle sont donc globalement constantes par le principe du prolongement analytique. ∎

Notons que si U est un domaine borné, alors $\overline{U}$ est compact. Si $f : \overline{U} \to \mathbb{C}$ est une fonction continue, alors $|f|$ atteint un maximum sur $\overline{U}$. Le corollaire suivant affirme que ce maximum est atteint au bord de U.

Corollaire 37. Soit $U \subset \mathbb{C}$ un domaine borné. Soit $f : \overline{U} \to \mathbb{C}$ une fonction continue sur $\overline{U}$ et analytique sur U. Alors :

$$\forall z \in U \quad \left|f(z)\right| \leqslant \max_{w \in \partial U} \left|f(w)\right|$$

avec inégalité stricte si f n'est pas constante.

Comme application, nous proposons l'exercice suivant qui sera complété par l'exercice II.6.21 de la page 797.

Exercice 11.

Soit $f : \overline{D}(0,1) \to \mathbb{C}$ une fonction continue sur $\overline{D}(0,1)$ et analytique sur $D(0,1)$ telle que $\left|f(z)\right| = 1$ dès que $|z| = 1$. Montrer que soit f s'annule dans $D(0,1)$, soit f est constante sur $D(0,1)$.

Solution. D'après le principe du maximum, on sait que $\left|f(z)\right| \leqslant 1$ pour tout $z \in D(0,1)$. De plus, si f ne s'annule pas dans $D(0,1)$, la fonction $1/f$ est analytique dans $D(0,1)$. En lui appliquant le principe du maximum, on voit que $\left|1/f(z)\right| \leqslant 1$ pour tout $z \in D(0,1)$. Autrement dit, $\left|f(z)\right| \geqslant 1$ pour tout $z \in D(0,1)$.

On voit donc que $\big|f(z)\big| = 1$ pour tout $z \in D(0,1)$. Le module de f étant constant, $|f|$ admet des maxima locaux en tout point de $D(0,1)$ et toujours d'après le principe du maximum, f est constante sur $D(0,1)$.

4.5 Suites de fonctions analytiques

Rappelons que dans le cas des fonctions d'une variable réelle, la limite uniforme d'une suite de fonctions dérivables n'est pas nécessairement dérivable. Pour vérifier que la limite est dérivable, il est habituellement nécessaire de vérifier que la suite des dérivées converge uniformément. Nous allons maintenant montrer, grâce à la formule de Cauchy, que la limite uniforme d'une suite de fonctions analytiques est automatiquement analytique. L'analyticité étant une propriété locale, il est d'ailleurs suffisant de ne requérir que la convergence uniforme au voisinage de tout point.

Définition 15. Soit $U \subset \mathbb{C}$ un ouvert. Une suite de fonctions $(f_n : U \to \mathbb{C})_{n\geqslant 0}$ *converge uniformément sur tout compact* de U vers $f : U \to \mathbb{C}$ si pour tout compact $K \subset U$, la suite (f_n) converge uniformément vers f sur K.

Exemple. La suite de fonctions $z \mapsto z^n$ converge uniformément vers la fonction nulle sur tout compact du disque unité ouvert $\mathbb{D}$. En revanche, elle ne converge pas uniformément sur $\mathbb{D}$.

Le lecteur vérifiera aisément qu'une suite de fonctions $(f_n : U \to \mathbb{C})_{n\geqslant 0}$ converge uniformément sur tout compact de U vers $f : U \to \mathbb{C}$ si, et seulement si, pour tout $z \in U$, il existe $r > 0$ tel que la suite (f_n) converge uniformément sur $D(z,r)$.

Théorème 38. Soit $U \subset \mathbb{C}$ un ouvert et $(f_n : U \to \mathbb{C})_{n\geqslant 0}$ une suite de fonctions analytiques sur U qui converge uniformément sur tout compact de U vers $f : U \to \mathbb{C}$. Alors f est analytique sur U. De plus, pour tout $k \geqslant 1$, la suite des dérivées $k^{\text{èmes}}$ $f_n^{(k)} : U \to \mathbb{C}$ converge uniformément sur tout compact de U vers $f^{(k)} : U \to \mathbb{C}$.

Démonstration. Soit $z_0 \in U$ et $R > r > 0$ tels que $D(z_0, R) \subset U$. Les formules de Cauchy affirment que pour tout $z \in D(z_0, r)$, on a :

$$f_n(z) = \int_0^1 \frac{re^{i2\pi t}}{z_0 + re^{i2\pi t} - z} f_n(z_0 + re^{i2\pi t})\,\mathrm{d}t.$$

Comme la suite (f_n) converge uniformément vers f sur le cercle de centre z_0 et de rayon r et comme les fonctions intégrées sont continues, on obtient par passage à la limite :

$$f(z) = \int_0^1 \frac{re^{i2\pi t}}{z_0 + re^{i2\pi t} - z} f(z_0 + re^{i2\pi t})\,\mathrm{d}t.$$

Il suffit alors de développer $z \mapsto \dfrac{1}{z_0 + re^{i2\pi t} - z}$ en série entière en z_0 et d'intervenir la somme et l'intégrale puisque la convergence de la série entière est uniforme sur le cercle de

centre z_0 et de rayon r :

$$f(z) = \sum_{k=0}^{+\infty} \left(\int_0^1 \frac{f(z_0 + re^{i2\pi t})}{r^k e^{i2\pi kt}} \, \mathrm{d}t \right) \cdot (z - z_0)^k.$$

Cela montre que f est analytique sur $D(z_0, r)$.

Étant donné $k \geqslant 1$, les formules de Cauchy montrent alors que pour tout $z \in D(z_0, r/2)$, on a :

$$\begin{aligned} \frac{f_n^{(k)}(z) - f^{(k)}(z)}{k!} &= \int_0^1 \frac{re^{i2\pi t}}{\left(z_0 + re^{i2\pi t} - z\right)^{k+1}} (f_n - f)(z_0 + re^{i2\pi t}) \, \mathrm{d}t \\ &\leqslant \sup_{|\zeta - z_0| = r} \frac{r \cdot \left| f_n(\zeta) - f(\zeta) \right|}{|\zeta - z|^{k+1}} \leqslant \frac{2^{k+1}}{r^k} \sup_{|\zeta - z_0| = r} \left| f_n(\zeta) - f(\zeta) \right|. \end{aligned}$$

On voit donc que $f_n^{(k)}$ converge uniformément vers $f^{(k)}$ sur $D(z_0, r/2)$. ∎

Notons que si $f : U \to \mathbb{C}$ est définie comme la somme d'une série de fonctions analytiques sur U, alors f est analytique dès que la convergence est uniforme sur tout compact de U. C'est par exemple le cas pour la fonction zeta de Riemann.

Exercice 12.

Montrer que la fonction zeta de Riemann $\zeta : s \mapsto \sum_{n=1}^{+\infty} \frac{1}{n^s}$ est définie et analytique sur $\mathbb{H} := \left\{ z \in \mathbb{C} \;\middle|\; \mathrm{Re}(z) > 1 \right\}$.

Solution. La fonction $s \mapsto \frac{1}{n^s} = \exp(-s \ln n)$ est analytique sur $\mathbb{H}$ (et même sur $\mathbb{C}$) comme composée de fonctions analytiques. Pour tout $\varepsilon > 0$, si $\mathrm{Re}(s) > \eta$, on a la majoration :

$$\left| \frac{1}{n^s} \right| = \exp\bigl(-\mathrm{Re}(s) \ln n \bigr) \leqslant e^{-n \ln \eta} = \frac{1}{n^\eta}.$$

La série $\sum_{n \geqslant 1} n^{-s}$ converge donc normalement (en particulier uniformément) sur tout compact de $\mathbb{H}$. Cela montre que $\zeta(s)$ est définie et analytique sur $\mathbb{H}$.

Nous concluons ce module avec la notion de famille normale. Nous reviendrons plus longuement sur cette notion dans [MPA] sur les fonctions holomorphes.

Définition 16. Soit $U \subset \mathbb{C}$ un ouvert, $\mathcal{A}$ un ensemble et $(f_\alpha : U \to \mathbb{C})_{\alpha \in \mathcal{A}}$ une famille de fonctions analytiques sur U, indexée par $\mathcal{A}$. La famille (f_α) est *normale* si de toute suite $(\alpha_n \in \mathcal{A})_{n \geqslant 0}$ on peut extraire une sous-suite $(\alpha_{n_k})_{k \geqslant 0}$ telle que la suite $f_{\alpha_{n_k}}$ converge uniformément sur tout compact de U vers une limite $f : U \to \mathbb{C}$.

Exemple. Soit $\mathbb{D} := \{z \in \mathbb{C} \mid |z| < 1\}$.
La famille de similitudes $(f_\lambda : \mathbb{C} \to \mathbb{C})_{\lambda\in\mathbb{D}}$ définies par $f_\lambda(z) = \lambda z$ est normale. En effet, $\overline{\mathbb{D}}$ est compact. Donc, de toute suite $(\lambda_n \in \mathbb{D})_{n\geqslant 0}$, on peut extraire une sous-suite λ_{n_k} convergeant vers $\lambda \in \overline{\mathbb{D}}$. Alors, la suite f_{λ_n} converge uniformément sur tout compact de $\mathbb{C}$ vers la fonction $f : z \mapsto \lambda z$. Notons qu'il se peut que $|\lambda| = 1$ auquel cas la limite f n'appartient pas à la famille $(f_\lambda)_{\lambda\in\mathbb{D}}$.

Un des résultats majeurs que nous établirons dans [MPA] est le théorème suivant dû à Montel.

Théorème 39 (Montel). Soit $U \subset \mathbb{C}$ un ouvert et soit $(f_\alpha : U \to \mathbb{C})_{\alpha\in\mathcal{A}}$ une famille de fonctions holomorphes. Si la famille est uniformément bornée sur tout compact de U, alors elle est normale.

Exercice type corrigé

II.6.0 On rappelle que la fonction $\Gamma :]0, +\infty[\to \mathbb{R}$ est définie par $\Gamma(x) = \displaystyle\int_0^{+\infty} t^{x-1}e^{-t}\,\mathrm{d}t$ (voir le module II.3). On rappelle que pour tout $x > 0$, $\Gamma(x+1) = x\Gamma(x)$ et que :

$$\Gamma(x) = \lim_{n\to+\infty} \frac{n^x n!}{(x+1)\cdots(x+n)}.$$

1. Montrer que la formule intégrale définit une fonction analytique dans le demi plan :
$$\mathbb{H} := \{z \in \mathbb{C} \mid \mathrm{Re}(z) > 0\}.$$
2. Montrer que pour tout $z \in \mathbb{H}$, on a $\Gamma(z+1) = z\Gamma(z)$.
3. Montrer qu'il existe une fonction analytique sur $\mathbb{C}\backslash\{0, -1, -2, \ldots\}$ qui prolonge Γ. Montrer que ce prolongement peut être défini par la limite :
$$\Gamma(z) = \lim_{n\to+\infty} \frac{n^z n!}{(z+1)\cdots(z+n)}.$$

Solution.

1. Rappelons que, pour $t > 0$ et $z \in \mathbb{C}$, on pose $t^z := e^{z\ln t}$. Notons que la fonction $t \mapsto t^{z-1}e^{-t}$ est continue (donc localement intégrable) sur $]0, +\infty[$. De plus, pour tout $z \in \mathbb{C}$, on a :
$$|t^{z-1}e^{-t}| = t^{\mathrm{Re}(z)-1}e^{-t}.$$
On peut donc définir $\Gamma(z) := \displaystyle\int_0^{+\infty} t^{z-1}e^{-t}\,\mathrm{d}t$ dès que $\mathrm{Re}(z) > 0$.
De plus, pour tout $t \in]0, +\infty[$, la fonction $z \mapsto t^{z-1}e^{-t}$ est holomorphe de dérivée $t^{z-1}e^{-t}\ln t$. Et pour tout $z \in \mathbb{C}$, on a :
$$|t^{z-1}e^{-t}\ln t| = t^{\mathrm{Re}(z)-1}e^{-t}\ln t.$$
On peut alors procéder comme dans les exercices 32 de la page 645 et 33 de la page 647 sur la dérivabilité des fonctions définies par une intégrale. On montre ainsi que Γ est holomorphe sur $\mathbb{H}$, que la dérivée et continue et vaut :
$$\Gamma'(z) = \int_0^{+\infty} t^{z-1}e^{-t}\ln t\,\mathrm{d}t.$$

En particulier, Γ est analytique sur $\mathbb{H}$.

2. L'égalité $\Gamma(z+1) = z\Gamma(z)$ est valide pour tout $z \in \mathbb{R}_+^*$ (exercice 16 de la page 614). Par principe du prolongement analytique, elle est valable pour tout $z \in \mathbb{H}$.
3. On peut alors prolonger Γ à $\mathbb{C} \setminus \{0, -1, -2, \ldots\}$ en posant :

$$\Gamma(z) = \frac{\Gamma(z+n)}{z(z+1)\cdots(z+n-1)}$$

avec $n > -\mathrm{Re}(z)$. Comme $\Gamma(z+1) = z\Gamma(z)$, la valeur ainsi définie ne dépend pas du choix de $n > -\mathrm{Re}(z)$. Pour tout $z_0 \in \mathbb{C} \setminus \{0, -1, -2, \ldots\}$, on peut choisir un même entier n pour définir le prolongement au voisinage de z_0, ce qui montre que le prolongement est holomorphe au voisinage de z_0 comme composée et quotient de fonctions holomorphes. Cela montre que le prolongement à $\mathbb{C} \setminus \{0, -1, -2, \ldots\}$ est holomorphe. Enfin, lorsque $\mathrm{Re}(z) > 0$, la formule :

$$\Gamma(z) = \lim_{n \to +\infty} \frac{n^z n!}{(z+1)\cdots(z+n)}$$

s'obtient comme dans l'exercice 27 de la page 634. En utilisant la relation $z\Gamma(z) = \Gamma(z+1)$, on voit que cette égalité est valide pour tout $z \in \mathbb{C} \setminus \{0, -1, -2, \ldots\}$.

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 Donner un exemple de fonction de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$ qui ne soit pas analytique sur $\mathbb{R}$.

II.6.2 Soit a un réel, I un intervalle ouvert contenant a et $f : I \to \mathbb{R}$ une fonction de classe $\mathcal{C}^\infty$. On suppose qu'il existe $r, M > 0$ tels que pour tout $x \in I$ et tout $n \geqslant 0$,

$$\left|f^{(n)}(x)\right| \leqslant M n! r^n.$$

1. Montrer que la série de Taylor de f en a a un rayon de convergence non nul.
2. Montrer qu'au voisinage de a, f coïncide avec sa série de Taylor en a.

II.6.3 Montrer qu'il existe une fonction analytique $f : \mathbb{C} \to \mathbb{C}$ telle que $f(z^2) = \cos(z)$ pour tout $z \in \mathbb{C}$. Déterminer l'ensemble des zéros de f.

II.6.4

1. Soit f une application développable en série entière sur un disque $D(0,r)$ avec $r > 0$ et $\alpha \in D(0,r)$ un nombre complexe non nul tel que $f(\alpha) \neq 0$. On considère le développement en série entière : $\dfrac{f(z)}{z-\alpha} = \sum_{n=0}^{+\infty} a_n z^n$. Montrer que $\lim\limits_{n\to\infty} \dfrac{a_n}{a_{n+1}} = \alpha$.
2. Soit P un polynôme n'ayant que des racines simples et $z_0 \in \mathbb{C}$ un nombre complexe tel que $P(z_0) \neq 0$. On suppose qu'il n'y a pas deux zéros de P équidistants de z_0. Montrer que
$$\lim_{n\to\infty} (n+1)\frac{(1/P)^{(n)}(z_0)}{(1/P)^{(n+1)}(z_0)} = \alpha_0 - z_0$$
où α_0 est le zéro de P le plus proche de z_0.

On peut utiliser ce résultat pour chercher les zéros de P.

II.6.5 $*$ Montrer que les solutions de l'équation différentielle $y'' - xy = 0$ sont analytiques sur $\mathbb{C}$.

II.6.6 Soit $f : [0,1] \to \mathbb{C}$ une fonction continue. Montrer que la fonction $F : \mathbb{C} \to \mathbb{C}$ définie par $F(z) = \displaystyle\int_0^1 f(t)\exp(tz)\,\mathrm{d}t$ est analytique.

Exercices sur la section 2

II.6.7 Soit $I \subset \mathbb{R}$ un intervalle non vide et soit $f : I \to \mathbb{C}$ une fonction de I dans $\mathbb{C}$.

1. Montrer que f est analytique sur I si, et seulement si, il existe un domaine $D \subset \mathbb{C}$ contenant I et une fonction analytique $F : D \to \mathbb{C}$ tels que $f = F_{|_I}$.
2. Montrer que si $f : I \to \mathbb{C}$ est analytique et non identiquement nulle, alors les zéros de f sont isolés dans I.

II.6.8 Trouver toutes les fonctions analytiques dans le disque unité qui vérifient $f(2z) = f(z)$.

II.6.9 $*$ Soient τ_1 et τ_2 deux nombres complexes et soit $f : \mathbb{C} \to \mathbb{C}$ une fonction entière telle que $f(z+\tau_1) = f(z+\tau_2) = f(z)$ pour tout $z \in \mathbb{C}$.

1. On suppose que $\tau_1/\tau_2 \in \mathbb{R} \setminus \mathbb{Q}$. Montrer que f est constante.
2. Peut-on aboutir à la même conclusion lorsque $\tau_1/\tau_2 \in \mathbb{Q}$?

II.6.10 Montrer que $\{z \in \mathbb{C} \mid \exists n \in \mathbb{N} \,:\, z^{2^n} = 1\}$ est dense dans $\mathbb{U}$.

II.6.11 Soit U un ouvert de $\mathbb{C}$.
Pour tout $n \in \mathbb{N}^*$, vérifier que $\left\{z \in \mathbb{C} \;\middle|\; |z| \leqslant n \text{ et } D(z,1/n) \subset U\right\}$ est un compact.

Exercices sur la section 3

II.6.12

1. Déterminer le nombre de composantes connexes de l'ensemble $\Omega \subset \mathbb{C}^2$ des couples (z_1, z_2) tels que $z_1 \in \mathbb{C} \setminus]-\infty, 0]$, $z_2 \in \mathbb{C} \setminus]-\infty, 0]$ et $z_1 z_2 \in \mathbb{C} \setminus]-\infty, 0]$.
2. Pour $(z_1, z_2) \in \Omega$, comparer $\mathrm{Log}(z_1 z_2)$ et $\mathrm{Log}(z_1) + \mathrm{Log}(z_2)$.

II.6.13

Soit $U \subset \mathbb{C}$ le domaine défini par $U := \mathbb{C} \setminus \big(]-\infty, -1] \cup [1, +\infty[\big)$ et soit $V := \{z \in \mathbb{C} \mid iz \in U\}$.

1. Pour $z \in U$, on pose $\mathrm{argth}\, z := \dfrac{1}{2}\big(\mathrm{Log}(1+z) - \mathrm{Log}(1-z)\big)$.

   Montrer que l'on définit ainsi une fonction analytique sur U qui prolonge $\mathrm{argth} :]-1, 1[\to \mathbb{R}$.
2. Pour $z \in V$, on pose $\arctan z := -i\,\mathrm{argth}(iz)$. Montrer que l'on définit ainsi une fonction analytique sur V qui prolonge $\arctan : \mathbb{R} \to \mathbb{R}$.

II.6.14

Soit $U \subset \mathbb{C}$ le domaine défini par $U := \mathbb{C} \setminus \big(]-\infty, -1] \cup [1, +\infty[\big)$ et soit $V := \{z \in \mathbb{C} \mid iz \in U\}$.

1. Montrer que si $z \in V$, alors $1 + z^2 \in \mathbb{C} \setminus \mathbb{R}_-$ et $z + \sqrt{1+z^2} \in \mathbb{C} \setminus \mathbb{R}_-$.
2. Montrer que si $z \in U$, alors $1 - z^2 \in \mathbb{C} \setminus \mathbb{R}_-$ et $iz + \sqrt{1-z^2} \in \mathbb{C} \setminus \mathbb{R}_-$.
3. Pour $z \in V$, on pose $\mathrm{argsh}(z) = \mathrm{Log}\big(z + \sqrt{1+z^2}\big)$. Montrer que l'on définit ainsi une fonction analytique sur V qui prolonge $\mathrm{argsh} : \mathbb{R} \to \mathbb{R}$.
4. Pour $z \in U$, on pose $\arcsin(z) = -i\,\mathrm{argsh}(iz)$. Montrer que l'on définit ainsi une fonction analytique sur U qui prolonge $\arcsin :]-1, 1[\to \mathbb{R}$.

II.6.15

Soit $F := \displaystyle\sum_{n \geqslant 0} \frac{(-1)^n}{2n+1} z^{2n+1}$.

Calculer les trois premiers termes significatifs de la série réciproque.

II.6.16 $**$

Soit $F := \displaystyle\sum_{k \geqslant 1} a_k z^k$ une série entière d'ordre $\omega(F) = 1$ et soit $G := \displaystyle\sum_{k \geqslant 1} b_k w^k$ la série réciproque. Nous proposons de démontrer la formule de réversion de Lagrange qui permet d'exprimer les coefficients b_n en fonctions des coefficients des puissances entières de la série $H(z) = z/F(z)$.

1. Pour $n \geqslant 0$, on définit $(a_{k,n})_{k \geqslant n}$ et $(b_{k,n})_{n \geqslant 1}$ par $F^n = \displaystyle\sum_{k \geqslant n} a_{k,n} z^k$ et $G^n = \displaystyle\sum_{k \geqslant n} b_{k,n} w^k$.

   Montrer que les $b_{k,n}$ sont les seules solutions du système :

$$\begin{cases} a_{n,n} b_{n,n} = 1 \\ \displaystyle\sum_{m=n}^{k} a_{m,n} b_{k,m} = 0 \quad \text{si } k > n. \end{cases}$$

2. Pour $n \geqslant 0$, on définit $(c_{k,n})_{k \geqslant 0}$ par $H^n(z) = \displaystyle\sum_{k \geqslant 0} c_{k,n} z^k$ avec $H(z) = z/F(z)$.

   Montrer que, pour $p > n$, on a $\displaystyle\sum_{k=n}^{p} k a_{k,n} c_{p-k,p} = 0$.

3. En déduire que pour $k \geqslant n \geqslant 1$, on a $b_{k,n} = \dfrac{n}{k} c_{k-n,k}$.
4. En déduire que pour $k \geqslant 1$, on a $b_k = \dfrac{1}{k} c_{k-1,k}$.

Exercices sur la section 4

II.6.17 Soit $I \subset \mathbb{R}$ un intervalle ouvert et $f : I \to \mathbb{C}$ une fonction analytique.

1. Montrer que f ne vérifie pas nécessairement le principe du maximum.
2. Étant donné un point $x_0 \in I$, montrer que la fonction f n'est pas nécessairement développable en série entière sur le plus grand intervalle centré en x_0 et contenu dans I.

II.6.18 Soit f une fonction analytique dans $D(0,1)$ telle que $\big|f(z)\big| \to 0$ quand $|z| \to 1$. Montrer que f est identiquement nulle sur $D(0,1)$.

II.6.19 $*$ Soit $f : D(0,1) \to D(0,1)$ une fonction holomorphe telle que $f(0) = 0$.

1. Montrer qu'il existe une fonction analytique $g : D(0,1) \to \mathbb{C}$ telle que $f(z) = zg(z)$ pour tout $z \in D(0,1)$.
2. Montrer que pour tout $z \in D(0,1)$, $\big|g(z)\big| \leqslant 1$.
3. En déduire que $\big|f'(0)\big| \leqslant 1$ et que pour tout $z \in D(0,1)$, $\big|f(z)\big| \leqslant |z|$.

II.6.20

Soit $U \subset \mathbb{C}$ un domaine borné et $f : U \to \mathbb{C}$ une fonction analytique telle que $\mathrm{Im}\big(f(z)\big) \to 0$ quand $z \to \partial U$. Montrer que f est constante sur U.

II.6.21 $*$

Soit $f : D(0,1) \to D(0,1)$ une application analytique telle que $\big|f(z)\big| \to 1$ quand $|z| \to 1$.

1. Montrer que f n'a qu'un nombre fini de zéros dans $D(0,1)$.

Soient $a_1, a_2, \ldots, a_k$ ces zéros. Soient $m_1, m_2, \ldots, m_k$ leurs multiplicités en tant que zéros de f.

2. Montrer que l'application g définie par :

$$g(z) = \prod_{j=1}^{k} \left(\frac{z - a_j}{1 - \bar{a}_j z} \right)^{m_j}$$

est analytique sur $D(0,1)$ et que $\big|g(z)\big| \to 1$ quand $|z| \to 1$.

3. Montrer qu'il existe une fonction analytique $h : D(0,1) \to D(0,1)$ telle que $f(z) = g(z) \cdot h(z)$. Montrer que $\big|h(z)\big| \to 1$ quand $|z| \to 1$ et que h ne s'annule pas sur $D(0,1)$.
4. En déduire qu'il existe un nombre complexe λ de module 1 tel que :

$$f(z) = \lambda \prod_{j=1}^{k} \left(\frac{z - a_j}{1 - \bar{a}_j z} \right)^{m_j}.$$

II.6.22 $*$ Soient τ_1 et τ_2 deux nombres complexes tels que τ_1/τ_2 n'est pas réel. Soit $f : \mathbb{C} \to \mathbb{C}$ une fonction entière telle que $f(z + \tau_1) = f(z + \tau_2) = f(z)$ pour tout $z \in \mathbb{C}$.
Montrer que f est constante.

Équations différentielles

II.7

Dès que le calcul infinitésimal s'est stabilisé, aux XVI^{e} et XVII^{e} siècles, la notion d'équation différentielle s'est imposée comme modèle mathématique *universel* des lois de la nature. Les pères fondateurs de la théorie sont Newton, Leibniz et Jean Bernoulli, ce qui suffit à en illustrer l'importance ! La loi fondamentale $F = m\gamma$ de Newton reliant force et accélération est source inépuisable d'équations différentielles et semble faire de celles-ci un moyen de description général de l'univers matériel. La notion de déterminisme en physique a même pu être *identifiée* avec le théorème (non démontré à l'époque !) d'existence et d'unicité des solutions d'équations différentielles à « conditions initiales » (positions et vitesses) fixées. Les travaux d'Euler et de Cauchy visant à donner des bases solides à cette théorie répondaient donc à une exigence épistémologique. Depuis, les équations différentielles sont devenues un outil merveilleux pour toutes les sciences de la nature ; mais elles ont également fourni un moyen d'attaque de nombreux problèmes de mathématiques pures, en géométrie comme en théorie des nombres.

Les deux premières sections de ce module portent sur les équations différentielles linéaires pour les fonctions de variable réelle. Ils prolongent donc la section du module sur les fonctions vectorielles de [L1] consacrée au même sujet : le lecteur est invité à s'y reporter pour réviser les techniques de base. Les principales nouveautés sont ici l'usage systématique de l'algèbre linéaire et polynomiale et la démonstration générale de l'existence et de l'unicité des solutions, due à Cauchy.

La troisième section présente une sensibilisation aux équations différentielles non linéaires. Leur étude qualitative par Poincaré, en liaison avec la mécanique céleste, est à l'origine d'une bonne partie de la géométrie et de la topologie du XX^{e} siècle. C'est une source de problèmes ouverts difficiles, aux nombreux angles d'attaque. Il ne peut être ici question d'aller plus loin que l'étude de quelques exemples. C'est aussi l'occasion d'introduire la méthode d'Euler pour la résolution numérique approchée.

Conventions générales

Les fonctions étudiées sont à variables et valeurs réelles ou complexes et, chaque fois que possible, les deux cas seront étudiés en même temps : la lettre K désignera alors le corps $\mathbb{R}$

ou le corps $\mathbb{C}$. Pour les fonctions de variable réelle, la variable sera généralement dénommée t, puisque, dans l'immense majorité des applications, elle représente le temps. Lorsque l'on étudie une équation d'origine physique,on notera $x(t)$ la fonction et, comme il est coutumier en mécanique, respectivement $\dot{x}(t)$ et $\ddot{x}(t)$ les dérivées première (vitesse) et seconde (accélération) par rapport au temps t de la fonction $x(t)$ (position au temps t).

1 Équations différentielles linéaires sur $\mathbb{R}$

Nous allons illustrer sur un exemple, simple mais important, diverses approches développées dans la suite de cette section. Le lecteur est encouragé à y revenir tout au long de son étude.

L'exemple de l'oscillateur harmonique

On a vu dans le cours de mécanique de terminale que le mouvement d'un pendule simple peut être décrit, en première approximation, par l'équation différentielle :

$$\ddot{x} + x = 0. \tag{1}$$

Les simplifications qui justifient cette équation sont les suivantes. Tout d'abord, on néglige les frottements et la résistance de l'air, qui introduiraient un terme proportionnel à la vitesse $\dot{x}$. Ensuite, on suppose les oscillations petites, ce qui permet de remplacer le terme réaliste $\sin x$ par son développement limité au premier ordre x. Enfin, les unités ont été choisies telles que, si l désigne la longueur du pendule et g l'accélération de la pesanteur, on ait $\dfrac{l}{g} = 1$. L'équation (1) régit d'ailleurs aussi bien l'évolution de l'intensité d'un circuit RLC sans résistance et sans stimulation externe (oscillateur non amorti et non entretenu).

Du chapitre sur les équations différentielles linéaires du module sur les fonctions vectorielles de [L1], on déduit que les solutions de l'équation (1) se prolongent toutes en des solutions définies sur $\mathbb{R}$ tout entier. Les solutions à valeurs dans $\mathbb{C}$ sont les fonctions de la forme $t \mapsto a_1 e^{\mathrm{i}t} + a_2 e^{-\mathrm{i}t}$ avec $a_1, a_2 \in \mathbb{C}$. Elles forment donc un $\mathbb{C}$-espace vectoriel $\mathrm{Sol}_{\mathbb{C}}((1), \mathbb{R})$ de dimension 2. (Dans cette notation, le $\mathbb{R}$ entre parenthèses désigne l'intervalle de définition et le $\mathbb{C}$ en indice l'espace d'arrivée des solutions considérées.) Les solutions à valeurs dans $\mathbb{R}$ (celles qui ont un sens physique évident) sont les fonctions de la forme $t \mapsto a\cos t + b\sin t$ avec $a, b \in \mathbb{R}$. Elles forment donc un $\mathbb{R}$-espace vectoriel $\mathrm{Sol}_{\mathbb{R}}((1), \mathbb{R})$ de dimension 2. C'est d'ailleurs l'ensemble des points fixes de l'application $\mathbb{R}$-linéaire $f \mapsto \overline{f}$ de $\mathrm{Sol}_{\mathbb{C}}((1), \mathbb{R})$ dans lui-même.

Dans l'un ou l'autre de ces espaces, chaque solution peut être déterminée par ses *conditions initiales* en un temps t_0 : la position $x_0 := x(t_0)$ et la vitesse $\dot{x}_0 := \dot{x}(t_0)$ (on reconnaît le déterminisme classique). On trouve la solution en résolvant, selon que $K := \mathbb{R}$ ou $\mathbb{C}$, le système :

$$\begin{cases} a\cos t_0 + b\sin t_0 = x_0, \\ -a\sin t_0 + b\cos t_0 = \dot{x}_0, \end{cases} \qquad \text{ou} \qquad \begin{cases} a_1 e^{\mathrm{i}t_0} + a_2 e^{-\mathrm{i}t_0} = x_0, \\ a_1 \mathrm{i} e^{\mathrm{i}t_0} - a_2 \mathrm{i} e^{-\mathrm{i}t_0} = \dot{x}_0, \end{cases}$$

dont le déterminant est la valeur en t_0 du *wronskien* $w := x_1\dot{x}_2 - \dot{x}_1 x_2$ des deux solutions fondamentales $x_1 := \cos t$ et $x_2 := \sin t$ ou $x_1 := e^{\mathrm{i}t}$ et $x_2 := e^{-\mathrm{i}t}$. Ici, $w(t_0) = 1$ ou $-2\mathrm{i}$; en conséquence, pour $K := \mathbb{R}$ ou $\mathbb{C}$, l'application $x \mapsto \big(x(t_0), \dot{x}(t_0)\big)$ de $\mathrm{Sol}_K((1), \mathbb{R})$ dans K^2 est un isomorphisme.

Si l'on a deux solutions x_1 et x_2 *quelconques*, leur wronskien $w := x_1\dot{x}_2 - \dot{x}_1 x_2$ a pour dérivée $\dot{w} = x_1\ddot{x}_2 - \ddot{x}_1 x_2 = x_1(-x_2) - (-x_1)x_2 = 0$. Il est donc constant. Sa valeur est déterminée par les conditions initiales des deux solutions : $w_0 := w(t_0) = x_1(t_0)\dot{x}_2(t_0) - \dot{x}_1(t_0)x_2(t_0)$. Ainsi, si l'on a réussi à trouver *une* solution non triviale x_1, toutes les autres solutions s'obtiennent en résolvant l'équation non homogène d'ordre 1 de fonction inconnue x :

$$x_1\dot{x} - \dot{x}_1 x = w_0.$$

Posons $E(x, \dot{x}) := x^2 + \dot{x}^2$. La dérivée par rapport au temps de cette expression est $2\dot{x}(x + \ddot{x})$. Ainsi, pour toute solution $x(t)$ de (1), l'expression $E(x, \dot{x})$ est constante : c'est une *intégrale première*. Elle permet de ramener la résolution de l'équation du second ordre (1) à celle de l'équation du premier ordre :

$$x^2 + \dot{x}^2 = E_0, \tag{2}$$

et cela, sans connaître une solution *a priori* ; en revanche, l'équation (2) n'est pas linéaire ! Lorsque le corps est $K := \mathbb{R}$, la relation (2) a d'ailleurs une signification physique : c'est la loi de conservation de l'énergie. Pour la résoudre, on prendra, pour simplifier, la condition initiale $x(0) = 0$ (origine des temps lorsque le pendule est en position basse). Si $\dot{x}(0) = 0$, on a $E_0 = 0$ et la seule solution est la fonction nulle (pendule immobile). Supposons donc $E_0 > 0$ et soit $]t_1, t_2[$ un intervalle contenant 0 sur lequel $\dot{x}$ est de signe constant ; autrement dit, on observe une portion du trajet du pendule sans rebroussement de chemin. Comme, pour toute solution x sur $]t_1, t_2[$, la fonction $x(t_1 + t_2 - t)$ est aussi solution, on peut supposer $\dot{x} > 0$ et écrire :

$$\forall t \in]t_1, t_2[\ ,\ \dot{x} = \sqrt{E_0 - x^2} \Longrightarrow t = \int_0^t \frac{\dot{x}(s)}{\sqrt{E_0 - x^2(s)}}\,\mathrm{d}s = \int_0^{x(t)/E_0} \frac{\mathrm{d}u}{\sqrt{1-u^2}},$$

par le changement de variable (légitime) $u := x(s)/E_0$. Du module « Fonctions transcendantes » de [L1], on déduit alors que $x(t) = E_0 \sin t$ sur $]t_1, t_2[$. Pour l'extension à une solution globale, voir l'exercice II.7.2 de la page 876.

Exercice 1.

Que devient la loi de conservation de l'énergie dans le cas de l'équation $\ddot{x} + k\dot{x} + x = 0$?

Solution. La dérivée par rapport au temps de $E(x, \dot{x})$ est maintenant $-2k\dot{x}^2$. Si $k > 0$, l'énergie décroît au cours du temps (dissipation sous forme de chaleur due à la résistance de l'air ou au frottement). Si $k < 0$, l'énergie croit au cours du temps : la seule interprétation physique possible est que de l'énergie est apportée au système de l'extérieur. (C'est le cas des circuits RLC avec *amplificateur opérationnel*, voir l'exercice II.7.20 de la page 877.)

Posons $X := \begin{pmatrix} x \\ \dot{x} \end{pmatrix}$. L'équation scalaire (1) d'ordre 2 est équivalente à l'équation vectorielle :

$$\dot{X} = AX, \tag{3}$$

où l'on a posé $A := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. Réciproquement, toute solution $X := \begin{pmatrix} x \\ y \end{pmatrix}$ de (3) vérifie $\dot{x} = y$ et $\dot{y} = -x$, donc est de la forme $\begin{pmatrix} x \\ \dot{x} \end{pmatrix}$, où x est une solution de (1). Posant $\mathcal{X}(t) := e^{(t-t_0)A}$, on peut d'ailleurs remarquer que la fonction matricielle $\mathcal{X}$ est solution de l'équation différentielle matricielle $\dot{\mathcal{X}} = A\mathcal{X}$, avec pour condition initiale $\mathcal{X}(t_0) = I_2$. Comme la fonction matricielle $\mathcal{X}$ est à valeurs dans $GL_2(\mathbb{R})$, toute solution X de (3) peut s'écrire $X(t) := \mathcal{X}(t)X_0$, où X_0 est *a priori* une fonction vectorielle ; mais il est alors facile de vérifier que X_0 est nécessairement constante (dériver la relation $X(t) := \mathcal{X}(t)X_0$ pour en déduire $\dot{X}_0 = 0$), donc égale à $X(t_0)$. L'unique solution de (3) vérifiant la condition initiale $X(t_0) = X_0$ est donc $X(t) := e^{(t-t_0)A}X_0$.

La matrice A est diagonalisable :

$$A = P\ \mathrm{Diag}(\mathrm{i}, -\mathrm{i})\, P^{-1}\ , \quad \text{où } P := \begin{pmatrix} 1 & 1 \\ \mathrm{i} & -\mathrm{i} \end{pmatrix}.$$

On a donc $\mathcal{X}(t) = P\,\mathrm{Diag}\big(e^{\mathrm{i}(t-t_0)}, e^{-\mathrm{i}(t-t_0)}\big)P^{-1} = \begin{pmatrix} \cos(t-t_0) & \sin(t-t_0) \\ -\sin(t-t_0) & \cos(t-t_0) \end{pmatrix}$, ce qui permet de retrouver les solutions obtenues à partir de l'équation scalaire.

Les propriétés du wronskien ont ici leur équivalent. Pour toute solution matricielle de $\dot{\mathcal{X}} = A\mathcal{X}$, posons $W := \det \mathcal{X}$. Si l'on note X_1 et X_2 les colonnes de $\mathcal{X}$, on a $\dot{X}_1 = AX_1$, $\dot{X}_2 = AX_2$ et $W = \det(X_1, X_2)$, d'où, en vertu de l'exercice 7 de la page 108 :

$$\begin{aligned} \dot{W} &= \det\big(\dot{X}_1, X_2\big) + \det\big(X_1, \dot{X}_2\big) \\ &= \det(AX_1, X_2) + \det(X_1, AX_2) \\ &= \mathrm{Tr}(A)\ \det(X_1, X_2) \\ &= \mathrm{Tr}(A)\, W, \end{aligned}$$

donc $\dot{W} = 0$, puisque $\mathrm{Tr}(A) = 0$ (le calcul de la dérivée de W est justifié par les exercices du module sur les fonctions vectorielles de [L1], ou bien par le lemme 1 de la page 807). Ainsi, W est constante. Si $\mathcal{X}(t_0) \in GL_2(\mathbb{R})$, la fonction matricielle $\mathcal{X}$ prend donc toutes ses valeurs dans $GL_2(\mathbb{R})$.

La loi de conservation de l'énergie s'écrit ici : $\|X\|^2 = E_0$. On peut la justifier en dérivant $\|X\|^2 = {}^tXX$, ce qui donne :

$$ {}^t\dot{X}X + {}^tX\dot{X} = {}^t(AX)X + {}^tXAX = {}^tX\big({}^tA + A\big)X = 0,$$

puisque la matrice A est antisymétrique. En fait, on a vu dans le module I.6 (remarques qui suivent la proposition 29 de la page 287) que l'antisymétrie de A entraîne que $\mathcal{X}$ est à valeurs dans $SO_2(\mathbb{R})$, donc que la norme de $X(t) := \mathcal{X}(t)X_0$ est constante.

Deux équations non linéaires

Certains des résultats des sections 1 et 2 pourront sembler tautologiques et certaines distinctions inutiles. L'étude de deux équations *non linéaires* scalaires d'ordre 1 simples montre qu'il n'en est rien. Considérons d'abord l'équation :

$$y' = y^2. \tag{4}$$

Cette équation n'admet *aucune solution globale autre que* 0. Soit en effet f une telle solution. Sur tout intervalle ouvert $I :=]a, b[$ où elle n'est pas nulle, la fonction f vérifie : $\left(\dfrac{1}{f}\right)' = -1$, donc $f(t) = \dfrac{1}{c-t}$ pour une certaine constante c. Puisque f est définie partout, $c \notin I$ et l'on a soit $c \leqslant a$ soit $c \geqslant b$. On supposera que $c \geqslant b$, l'autre cas se traitant exactement de la même manière. L'ensemble $\left\{ b' \leqslant c \mid \forall t \in]a, b'[\ ,\ f(t) = \dfrac{1}{c-t} \right\}$ est fermé car la fonction f est continue ; il est non vide, car il contient b ; et il est majoré par c. Il admet donc un plus grand élément. Quitte à remplacer b par ce plus grand élément, nous pouvons supposer que c'est b lui-même. Puisque $f(t) = \dfrac{1}{c-t}$ sur $]a, b[$, on a (par continuité) $b \neq c$ et $f(b) = \dfrac{1}{c-b} \neq 0$. Au voisinage de b, la fonction f est non nulle ; elle est donc de la forme $\dfrac{1}{c'-t}$ sur un certain intervalle $]b', b''[$ contenant b. L'égalité $\dfrac{1}{c-b} = \dfrac{1}{c'-b}$ entraîne alors $c' = c$, et la relation $f(t) = \dfrac{1}{c-t}$ est valable sur $]a, b''[$, contredisant la maximalité de b.

On peut également s'intéresser à l'équation :

$$y' = 2\sqrt{|y|}. \tag{5}$$

Il y a maintenant pléthore de solutions globales : par exemple, pour tout réel a, la fonction f_a qui vaut 0 sur $]-\infty, a]$ et $(t-a)^2$ sur $[a, +\infty[$ est solution. Or, toutes les solutions f_a telles que $a > 0$ s'annulent en 0. Une solution f de l'équation (5) n'est donc pas uniquement déterminée par la condition initiale $f(0)$ (voir également l'exercice II.7.2 de la page 876).

1.1 Généralités

Rappelons que la lettre K désigne $\mathbb{R}$ ou $\mathbb{C}$.

Définition 1. Une *équation différentielle linéaire scalaire d'ordre* n sur l'intervalle $I \subset \mathbb{R}$ est une équation de la forme :

$$(E_b) \quad y^{(n)} + a_1\, y^{(n-1)} + \cdots + a_n\, y = b, \tag{6}$$

où $a_1, \ldots, a_n, b$ sont des fonctions continues de I dans K.
Pour tout intervalle $J \subset I$, on appelle *solution* ou *intégrale* sur J de l'équation différentielle (6) toute fonction n fois dérivable $f : J \to K$ telle que :

$$\forall t \in J\ ,\ f^{(n)}(t) + a_1(t)\, f^{(n-1)}(t) + \cdots + a_n(t)\, f(t) = b(t). \tag{7}$$

L'ensemble des solutions sur J à valeurs dans K d'une équation différentielle (E_b) sera noté $\mathrm{Sol}_K\big((E_b), J\big)$.

Si f est une solution de (E_b) sur l'intervalle J, sa restriction à tout intervalle $J' \subset J$ est une solution sur J'. Une démarche fondamentale dans l'étude des équations différentielles (linéaires ou non) est d'abord *locale* : on cherche à résoudre l'équation au voisinage de tout point de I ; le voisinage obtenu peut être très petit. Ensuite se pose le problème du *passage du local au global* : on cherche à *prolonger* autant que possible les *solutions locales* ainsi trouvées, voire à *recoller* des solutions locales. Une solution f sur l'intervalle J est dite *globale* si $J = I$. Nous démontrerons (corollaire 30 de la page 830) que, *pour les équations différentielles linéaires*, toute solution peut être prolongée en une solution globale. Mais cela ne va pas de soi, comme l'a montré l'exemple de l'équation (4) page 803.

Point Méthode

Dans beaucoup d'ouvrages, on définit les solutions d'une équation différentielle sur un intervalle *ouvert* : en effet, la plupart des propriétés des fonctions de variable réelle sont plus faciles à formuler et à démontrer dans ce cadre. Dans la pratique, on rencontre souvent des fonctions définies sur un segment ou sur un intervalle semi-ouvert. Il faudra alors être prudent dans l'étude du comportement aux bornes.

Définition 2. On dit que l'équation (E_b) est *homogène* ou *sans second membre* si $b = 0$, qu'elle est *non homogène* ou *avec second membre* dans le cas général. L'*équation différentielle homogène associée à* (E_b) est l'équation :

$$(E_0) \quad y^{(n)} + a_1 y^{(n-1)} + \cdots + a_n y = 0. \tag{8}$$

Les faits suivants généralisent les propriétés énoncées dans [L1] (module IV.6, section 3), et se démontrent de la même manière.

1. En vertu de l'égalité $f^{(n)} = -a_1\, f^{(n-1)} - \cdots - a_n f + b$ et du fait que les a_i, b et $f, \ldots, f^{(n-1)}$ sont continues, toute solution de (E_b) est de classe $\mathcal{C}^n$ sur J.

2. L'ensemble $\mathrm{Sol}_K\big((E_0),J\big)$ des solutions sur J à valeurs dans K de l'équation homogène (E_0) est un K-espace vectoriel.

3. L'ensemble $\mathrm{Sol}_K\big((E_b),J\big)$ des solutions sur J à valeurs dans K de l'équation non homogène (E_b) est soit vide, soit un espace affine de direction $\mathrm{Sol}_K\big((E_0),J\big)$. Plus précisément, si f_0 est une solution particulière de (E_b), les solutions de (E_b) sont les f_0+f, où f est solution de (E_0).

4. Supposons que les fonctions $a_1,\dots,a_n,b$ soient à valeurs dans $\mathbb{R}$. Alors, pour toute fonction $f:J\to\mathbb{C}$, on a les équivalences :

$$f\in\mathrm{Sol}_{\mathbb{C}}\big((E_b),J\big)\iff\overline{f}\in\mathrm{Sol}_{\mathbb{C}}\big((E_b),J\big)\iff\mathrm{Re}\,f,\mathrm{Im}\,f\in\mathrm{Sol}_{\mathbb{R}}\big((E_b),J\big).$$

   L'application $f\mapsto\overline{f}$ est donc une symétrie du $\mathbb{R}$-espace affine $\mathrm{Sol}_{\mathbb{C}}\big((E_b),J\big)$, et son axe est l'espace des solutions réelles :

$$\mathrm{Sol}_{\mathbb{R}}\big((E_b),J\big)=\big\{f\in\mathrm{Sol}_{\mathbb{C}}\big((E_b),J\big)\mid\overline{f}=f\big\}.$$

5. Toute base du $\mathbb{R}$-espace vectoriel $\mathrm{Sol}_{\mathbb{R}}\big((E_0),J\big)$ est une base du $\mathbb{C}$-espace vectoriel $\mathrm{Sol}_{\mathbb{C}}\big((E_0),J\big)$.

   Réciproquement, à partir d'une base du $\mathbb{C}$-espace vectoriel $\mathrm{Sol}_{\mathbb{C}}\big((E_0),J\big)$ de la forme $(f_1,\dots,f_k,\overline{f}_1,\dots,\overline{f}_k,g_1,\dots,g_l)$ dans laquelle les g_i sont à valeurs réelles, on obtient une base $(\mathrm{Re}\,f_1,\dots,\mathrm{Re}\,f_k,\mathrm{Im}\,f_1,\dots,\mathrm{Im}\,f_k,g_1,\dots,g_l)$ du $\mathbb{R}$-espace vectoriel $\mathrm{Sol}_{\mathbb{R}}\big((E_0),J\big)$.

6. Pour résoudre une équation $y^{(n)}+a_1\,y^{(n-1)}+\cdots+a_n\,y=b$ lorsque le second membre est une combinaison linéaire $b=\alpha_1b_1+\cdots+\alpha_kb_k$ ($\alpha_i\in K$), il *suffit* de trouver une solution f_i de chaque équation $y^{(n)}+a_1\,y^{(n-1)}+\cdots+a_n\,y=b_i$ puis de prendre la solution $f:=\alpha_1f_1+\cdots+\alpha_kf_k$ (*principe de superposition*).

Exercice 2.

Quelles conditions poser sur les a_i et b pour garantir que toute solution de (E_b) soit de classe $\mathcal{C}^{n+k}$? De classe $\mathcal{C}^\infty$?

Solution. Si les a_i et b sont de classe $\mathcal{C}^k$, alors toute solution f est de classe $\mathcal{C}^{n+k}$. En effet, on sait d'abord que f est de classe $\mathcal{C}^n$. Si f est de classe $\mathcal{C}^{n+i}$ avec $i<k$, alors le membre droit de l'égalité $f^{(n)}=-a_1\,f^{(n-1)}-\cdots-a_nf+b$ est (au moins) de classe $\min\big(\min(k,i+1),\dots,\min(k,i+n),k\big)=i+1$, donc f est de classe $\mathcal{C}^{n+i+1}$. Par récurrence sur i, on en déduit que f est de classe $\mathcal{C}^{n+k}$.
Si les a_i et b sont de classe $\mathcal{C}^\infty$, alors toute solution f est donc de classe $\mathcal{C}^\infty$.

1.1.1 Le wronskien

Définition 3. Soient $f_1, \dots, f_n$ des fonctions $(n-1)$ fois dérivables sur l'intervalle I. Le *wronskien* des fonctions $f_1, \dots, f_n$ est la fonction :

$$W_n(f_1, \dots, f_n) := \det \begin{pmatrix} f_1 & f_2 & \cdots & f_n \\ f_1' & f_2' & \cdots & f_n' \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)} & f_2^{(n-1)} & \cdots & f_n^{(n-1)} \end{pmatrix}. \tag{9}$$

Si les fonctions $f_1, \dots, f_n$ sont toutes de classe $\mathcal{C}^{n-1}$, alors les fonctions $f_i^{(j)}$ $(1 \leqslant i \leqslant n,$ $0 \leqslant j \leqslant n-1)$ sont toutes continues et il en est de même de leur wronskien.

Soit $\sum_{i=1}^{n} \lambda_i f_i = 0$ une relation linéaire non triviale à coefficients $\lambda_i \in K$ constants entre les fonctions f_i. En dérivant cette relation, on obtient, pour tout entier j, la relation $\sum_{i=1}^{n} \lambda_i f_i^{(j)} = 0$. Les colonnes $C_1, \dots, C_n$ de la matrice wronskienne sont donc liées : $\sum_{i=1}^{n} \lambda_i C_i = 0$, et le wronskien s'annule identiquement : $W_n(f_1, \dots, f_n) = 0$. Nous verrons que, pour des solutions de l'équation homogène (E_0) de la page 804, il y a même équivalence. Mais, pour des fonctions quelconques, il est difficile de formuler une condition nécessaire et suffisante (exercice II.7.4 de la page 876) ; en tous cas, la réciproque évidente est fausse :

Exercice 3.

1. On suppose que f ne s'annule pas sur l'intervalle I.
   À quelle condition le wronskien $W_2(f, g)$ est-il identiquement nul sur I ?
2. Vérifier que les fonctions $f(t) = t^2$ et $g(t) = t|t|$ sont de classe $\mathcal{C}^1$ et que $W_2(f, g) = 0$.

Solution.

1. Si l'on pose $h := \dfrac{g}{f}$, on voit que $h' = \dfrac{W_2(f, g)}{f^2}$: le wronskien est donc nul si, et seulement si, g est proportionnelle à f.
2. La fonction f est polynomiale, donc de classe $\mathcal{C}^\infty$ Il en est de même de la fonction g restreinte à $\mathbb{R}_+^*$ et à $\mathbb{R}_-^*$. On vérifie facilement que, sur $\mathbb{R}$ tout entier, $g' = 2|t|$, qui est continue. Puisque $g = f$ sur $\mathbb{R}_+^*$ et $g = -f$ sur $\mathbb{R}_-^*$, leur wronskien est nul sur $\mathbb{R}^*$, donc sur $\mathbb{R}$ par continuité (pour un exemple de classe $\mathcal{C}^\infty$, voir l'exercice II.7.3 de la page 876).

Lorsque les fonctions $f_1, \dots, f_n$ sont toutes de classe $\mathcal{C}^n$, les fonctions $f_i^{(j)}$ $(1 \leqslant i \leqslant n,$ $0 \leqslant j \leqslant n-1)$ sont toutes de classe $\mathcal{C}^1$ et il en est de même de leur wronskien.

Lemme 1. Soient $F_1, \dots, F_n$ des fonctions dérivables à valeur dans un K-espace vectoriel de dimension n muni d'une base $\mathcal{B}$. Alors la fonction $\det_{\mathcal{B}}(F_1, \dots, F_n)$ est dérivable et sa dérivée est $\sum_{i=1}^{n} \det_{\mathcal{B}}(F_1, \dots, F_i', \dots, F_n)$.

Démonstration. On écrit $F_i(t+h) = F_i(t) + hF_i'(t) + hG_i(t,h)$, où, pour tout t fixé, $\lim_{h\to 0} G_i(t,h) = 0$. La multilinéarité du déterminant permet alors d'écrire :

$$\det_{\mathcal{B}}(F_1, \dots, F_n)(t+h) = \det_{\mathcal{B}}(F_1, \dots, F_n)(t) + h\sum_{i=1}^{n} \det_{\mathcal{B}}(F_1, \dots, F_i', \dots, F_n)(t) + hE(t,h),$$

où, pour tout t fixé, $\lim_{h\to 0} E(t,h) = 0$. ∎

Lemme 2. Si les fonctions $f_1, \dots, f_n$ sont de classe $\mathcal{C}^n$, la dérivée de leur wronskien est donnée par la formule :

$$W_n(f_1, \dots, f_n)' = \det \begin{pmatrix} f_1 & f_2 & \cdots & f_n \\ f_1' & f_2' & \cdots & f_n' \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-2)} & f_2^{(n-2)} & \cdots & f_n^{(n-2)} \\ f_1^{(n)} & f_2^{(n)} & \cdots & f_n^{(n)} \end{pmatrix}. \tag{10}$$

Démonstration. Notons $L_0, \dots, L_{n-1}$ les lignes de la matrice wronskienne considérée ; on a donc $L_i' = L_{i+1}$ pour $i := 0, \dots, n-2$. D'après le lemme 1, la dérivée du déterminant $\det(L_0, \dots, L_{n-1})$ est la somme des déterminants $\det(L_0, \dots, L_{i-1}, L_i', L_{i+1}, \dots, L_{n-1})$. Mais, d'après l'égalité $L_i' = L_{i+1}$, tous ces déterminants sont nuls, sauf le dernier, qui vaut $\det(L_0, \dots, L_{n-2}, L_{n-1}')$. ∎

Théorème 3. Le wronskien $w := W_n(f_1, \dots, f_n)$ de n solutions $f_1, \dots, f_n$ de l'équation différentielle homogène (E_0) de la page 804 vérifie l'équation différentielle :

$$w' = -a_1 w. \tag{11}$$

Démonstration. Si les f_i sont solutions de (E_0), elles sont de classe $\mathcal{C}^n$. Avec les notations du lemme, on a $L_n = -a_1 L_{n-1} - \cdots - a_n L_0$. En développant le déterminant $\det(L_0, \dots, L_{n-2}, L_n)$, on trouve donc $-a_1 \det(L_0, \dots, L_{n-2}, L_{n-1}) = -a_1 w$. ∎

Corollaire 4. Le wronskien de n solutions de l'équation différentielle homogène (E_0) sur l'intervalle J est identiquement nul sur J, ou bien il ne s'y annule pas.

Démonstration. Il ressort de [L1] (module IV.6, section 3.1) que $w(t) = Ce^{A(t)}$, où C est une constante et A une primitive de $-a_1$ sur J. ∎

1.1.2 Systèmes fondamentaux de solutions

Définition 4. On appelle *système fondamental de solutions* de l'équation différentielle homogène (E_0) sur l'intervalle J une famille $(f_1, \dots, f_n)$ de solutions dont le wronskien ne s'annule pas.

Nous n'affirmons pas, pour le moment, l'*existence* d'un système fondamental de solutions : c'est le contenu du théorème 29 de la page 830, qui sera le premier résultat d'analyse de ce module ! Cependant, un fait important mérite d'être noté : la restriction d'un tel système à un intervalle $J' \subset J$ est un système fondamental de solutions sur J' ; c'est trivial, mais l'assertion correspondante pour des familles libres ne le serait pas (voir le cas de t^2 et $t|t|$ sur $\mathbb{R}$ et sur $\mathbb{R}_+^*$).

Théorème 5. Tout système fondamental de solutions de (E_0) sur J est une base du K-espace vectoriel des solutions.

Démonstration. Soit $(f_1, \dots, f_n)$ un tel système et soit f une solution (toutes ces fonctions sont donc de classe $\mathcal{C}^n$). On résout le système :

$$\begin{cases} \lambda_1 f_1 + \cdots + \lambda_n f_n = f, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \lambda_1 f_1^{(n-1)} + \cdots + \lambda_n f_n^{(n-1)} = f^{(n-1)}. \end{cases}$$

C'est possible, puisque le déterminant du système ne s'annule pas. Les λ_i sont *a priori* des fonctions de classe $\mathcal{C}^1$. En dérivant chacune des $(n-1)$ premières relations et en la comparant à la suivante, on obtient les égalités $\lambda_1' f_1^{(j)} + \cdots + \lambda_n' f_n^{(j)} = 0$ pour $j := 0, \dots, n-2$. En dérivant la dernière relation et en utilisant les égalités $f_i^{(n)} = -a_1 f_i^{(n-1)} - \cdots - a_n f_i$ pour $i := 1, \dots, n$, et l'égalité $f^{(n)} = -a_1 f^{(n-1)} - \cdots - a_n f$ on trouve de plus que $\lambda_1' f_1^{(n-1)} + \cdots + \lambda_n' f_n^{(n-1)} = 0$. On a donc un système :

$$\begin{cases} \lambda_1' f_1 + \cdots + \lambda_n' f_n = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \lambda_1' f_1^{(n-1)} + \cdots + \lambda_n' f_n^{(n-1)} = 0. \end{cases}$$

Toujours parce que le déterminant du système ne s'annule pas, on en déduit que les λ_i' sont nuls, donc que les λ_i sont constants. Ainsi, $(f_1, \dots, f_n)$ est une famille génératrice. Mais on savait déjà que c'était une famille libre (*cf.* la remarque qui suit la définition 3). ∎

Corollaire 6. Soit $t_0 \in J$. S'il existe un système fondamental de solutions de (E_0) sur J, l'application « conditions initiales » $f \mapsto \big(f(t_0), \dots, f^{(n-1)}(t_0)\big)$ est un isomorphisme de $\mathrm{Sol}_K\big((E_0), J\big)$ sur K^n.

Démonstration. Il est clair que cette application est linéaire. Une solution f de (8) satisfaisant les *conditions initiales* $f(t_0) = y_0, \dots, f^{(n-1)}(t_0) = y_{n-1}$ est une fonction

$f := \lambda_1 f_1 + \cdots + \lambda_n f_n$ telle que :

$$\begin{cases} \lambda_1 f_1(t_0) + \cdots + \lambda_n f_n(t_0) = y_0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \lambda_1 f_1^{(n-1)}(t_0) + \cdots + \lambda_n f_n^{(n-1)}(t_0) = y_{n-1}. \end{cases}$$

Ce système admet une unique solution dans K^n. ■

Exercice 4.

Soient $f_1, \ldots, f_{p-1}$ des fonctions de classe $\mathcal{C}^p$ sur un intervalle J, dont le wronskien ne s'annule pas. Soit f une fonction de classe $\mathcal{C}^p$ sur J telle que $W_p(f_1, \ldots, f_{p-1}, f)$ est identiquement nul sur I. Montrer que f est combinaison linéaire de $f_1, \ldots, f_{p-1}$.

Solution. En développant le déterminant $W_p(f_1, \ldots, f_{p-1}, f)$ par rapport à sa dernière colonne, on trouve une expression de la forme $W_{p-1}(f_1, \ldots, f_{p-1}) f^{(p)} + b_1 f^{(p-1)} + \cdots + b_p f$, où les b_i sont des fonctions continues sur J. L'annulation de ce wronskien équivaut à une équation différentielle linéaire homogène scalaire d'ordre p sur J, dont $(f_1, \ldots, f_{p-1})$ est un système fondamental de solutions évidentes. La conclusion découle alors du théorème.

Application à la résolution de l'équation non homogène. La méthode que nous allons décrire généralise la méthode de *variation des constantes* de [L1]. Elle a été inventée par Lagrange à la fin du XVIII$^{\text{e}}$ siècle. Supposons trouvé un système fondamental de solutions $(f_1, \ldots, f_n)$ de l'équation différentielle homogène (E_0) de la page 804 sur J. Nous chercherons une solution f de l'équation différentielle non homogène (E_b) sous la forme $f = \lambda_1 f_1 + \cdots + \lambda_n f_n$, où les λ_i sont des *fonctions* à déterminer. Comme dans le théorème 5, nous résolvons *a priori* le système :

$$\begin{cases} \lambda_1 f_1 + \cdots + \lambda_n f_n = f, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \lambda_1 f_1^{(n-1)} + \cdots + \lambda_n f_n^{(n-1)} = f^{(n-1)}. \end{cases}$$

C'est possible pour toute fonction f de classe $\mathcal{C}^n$, puisque le déterminant du système ne s'annule pas, et les λ_i sont des fonctions de classe $\mathcal{C}^1$. On obtient encore les relations $\lambda_1' f_1^{(j)} + \cdots + \lambda_n' f_n^{(j)} = 0$ pour $j := 0, \ldots, n-2$. Cependant, en dérivant la dernière relation, on trouve maintenant : $\lambda_1' f_1^{(n-1)} + \cdots + \lambda_n' f_n^{(n-1)} = f^{(n)} - a_1 f^{(n-1)} - \cdots - a_n f$. Ainsi, pour que f soit solution de (E_b), il faut, et il suffit, que l'on ait $f = \lambda_1 f_1 + \cdots + \lambda_n f_n$, où les λ_i sont solutions du système :

$$\begin{cases} \lambda_1' f_1 + \cdots + \lambda_n' f_n = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \lambda_1' f_1^{(n-2)} + \cdots + \lambda_n' f_n^{(n-2)} = 0, \\ \lambda_1' f_1^{(n-1)} + \cdots + \lambda_n' f_n^{(n-1)} = b. \end{cases}$$

L'algorithme est donc le suivant : on résout ce système (ce qui est possible, par hypothèse) ; on en tire chaque λ_i à une constante près par « quadrature » (*i.e.* par calcul de primitive). Lorsque $n := 1$, on retrouve la méthode de variation des constantes de [L1].

Point Méthode

Selon les égalités : $f^{(k)} = \lambda_1 f_1^{(k)}$ pour $0 \leqslant k \leqslant n-1$, les conditions initiales $f^{(k)}(t_0) = c_k$ s'obtiennent en imposant aux λ_i des valeurs initiales $\lambda_i(t_0) = l_i$ telles que :

$$\begin{pmatrix} f_1(t_0) & f_2(t_0) & \dots & f_n(t_0) \\ f_1'(t_0) & f_2'(t_0) & \dots & f_n'(t_0) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)}(t_0) & f_2^{(n-1)}(t_0) & \dots & f_n^{(n-1)}(t_0) \end{pmatrix} \begin{pmatrix} l_0 \\ l_1 \\ \vdots \\ l_{n-1} \end{pmatrix} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix}.$$

Exemple. Nous allons résoudre par quadratures l'équation $x'' - cx = f$, où $c > 0$. Nous savons que les fonctions $x_1(t) := e^{\sqrt{c}t}$ et $x_2(t) := e^{-\sqrt{c}t}$ sont solutions de l'équation homogène $x'' - cx = 0$. Leur wronskien est $x_1x_2' - x_1'x_2 = -2\sqrt{c}$, elles forment donc un système fondamental de solutions sur $\mathbb{R}$. On cherche une solution sous la forme $x := \lambda x_1 + \mu x_2$, avec les conditions :

$$\begin{cases} \lambda' x_1 + \mu' x_2 = 0, \\ \lambda' x_1' + \mu' x_2' = f, \end{cases} \quad \textit{i.e.} \quad \begin{cases} \lambda'(t) e^{\sqrt{c}t} + \mu'(t) e^{-\sqrt{c}t} = 0, \\ \sqrt{c}\lambda'(t) e^{\sqrt{c}t} - \sqrt{c}\mu'(t) e^{-\sqrt{c}t} = f(t). \end{cases}$$

On en tire : $\lambda'(t) = \dfrac{1}{2\sqrt{c}} f(t) e^{-\sqrt{c}t}$ et $\mu'(t) = -\dfrac{1}{2\sqrt{c}} f(t) e^{\sqrt{c}t}$, d'où :

$$\lambda(t) = \lambda_0 + \int_{t_0}^{t} \frac{1}{2\sqrt{c}} f(s) e^{-\sqrt{c}s}\, \mathrm{d}s \quad \text{et} \quad \mu(t) = \mu_0 - \int_{t_0}^{t} \frac{1}{2\sqrt{c}} f(s) e^{\sqrt{c}s}\, \mathrm{d}s.$$

Les constantes λ_0 et μ_0 sont fixées par les conditions initiales. Si l'on impose $x(t_0) = x_0$ et $x'(t_0) = x_0'$, on doit résoudre :

$$\begin{pmatrix} e^{\sqrt{c}t_0} & e^{-\sqrt{c}t_0} \\ \sqrt{c}e^{\sqrt{c}t_0} & -\sqrt{c}e^{-\sqrt{c}t_0} \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \mu_0 \end{pmatrix} = \begin{pmatrix} x_0 \\ x_0' \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} \lambda_0 \\ \mu_0 \end{pmatrix} = -\frac{1}{2\sqrt{c}} \begin{pmatrix} -\sqrt{c}e^{-\sqrt{c}t_0} & -e^{-\sqrt{c}t_0} \\ -\sqrt{c}e^{\sqrt{c}t_0} & e^{\sqrt{c}t_0} \end{pmatrix} \begin{pmatrix} x_0 \\ x_0' \end{pmatrix},$$

soit $\lambda_0 = \dfrac{e^{-\sqrt{c}t_0}}{2\sqrt{c}} \left(\sqrt{c}x_0 + x_0'\right)$ et $\mu_0 = \dfrac{e^{\sqrt{c}t_0}}{2\sqrt{c}} \left(\sqrt{c}x_0 - x_0'\right)$.

Exercice 5.

Résoudre par quadratures l'équation $\ddot{x} + x = b$. Donner la forme précise de la solution satisfaisant aux conditions initiales $x(t_0) = x_0$, $\dot{x}(t_0) = \dot{x}_0$.

Solution. Le wronskien de $(\cos, \sin)$ est 1 : c'est donc un système fondamental de solutions sur $\mathbb{R}$. On cherche une solution sous la forme $x := \lambda \cos + \mu \sin$, avec les conditions :

$$\begin{cases} \dot{\lambda} \cos + \dot{\mu} \sin = 0, \\ -\dot{\lambda} \sin + \dot{\mu} \cos = b. \end{cases}$$

On a donc $\dot\lambda = -b\sin$ et $\dot\mu = b\cos$. Le lecteur vérifiera sans peine que ces relations entraînent bien que $x := \lambda\cos + \mu\sin$ satisfait $\ddot x + x = b$. Pour satisfaire les conditions initiales, on doit résoudre le système :

$$\begin{cases} \lambda(t_0)\cos t_0 + \mu(t_0)\sin t_0 = x_0, \\ -\lambda(t_0)\sin t_0 + \mu(t_0)\cos t_0 = \dot x_0, \end{cases}$$

d'où $\lambda(t_0) = x_0\cos t_0 - \dot x_0 \sin t_0$, $\mu(t_0) = x_0\sin t_0 + \dot x_0\cos t_0$. Finalement :

$$\lambda(t) = x_0\cos t_0 - \dot x_0\sin t_0 - \int_{t_0}^{t} b(s)\sin s\,\mathrm{d}s$$

$$\mu(t) = x_0\sin t_0 + \dot x_0\cos t_0 + \int_{t_0}^{t} b(s)\cos s\,\mathrm{d}s.$$

1.2 Équations à coefficients constants

Nous supposons maintenant que les fonctions $a_1, \dots, a_n$ qui apparaissent dans l'équation (E_b) de la page 804 sont *constantes* sur l'intervalle $I := \mathbb{R}$. Les solutions de l'équation homogène associée (E_0) sont alors de classe $\mathcal{C}^\infty$ (exercice 2 de la page 805).

Pour simplifier autant que se peut les lourds calculs qui vont suivre, nous noterons (improprement) $e^{\lambda t}$ la fonction $t \mapsto e^{\lambda t}$ de $\mathcal{C}^\infty(I, K)$. De même, nous noterons t^j la fonction polynomiale $t \mapsto t^j$ et $K_d[t]$ le K-espace vectoriel des fonctions polynomiales de degré inférieur ou égal à d de $\mathbb{R}$ dans K. Nous introduisons de plus l'endomorphisme $D : f \mapsto f'$ du K-espace vectoriel $\mathcal{C}^\infty(I, K)$.

L'application $f \mapsto f^{(n)} + a_1 f^{(n-1)} + \cdots + a_n f$ n'est autre que l'endomorphisme $P(D)$ de $\mathcal{C}^\infty(I, K)$ défini par le polynôme : $P := X^n + a_1 X^{n-1} + \cdots + a_n \in K[X]$. Le K-sous–espace vectoriel de $\mathcal{C}^\infty(I, K)$ formé des solutions de (E_0) sur $\mathbb{R}$ est le noyau de $P(D)$:

$$E := \mathrm{Sol}_K\big((E_0), \mathbb{R}\big) = \mathrm{Ker}\, P(D).$$

Définition 5. L'équation (algébrique) $P(\lambda) = 0$ est appelée *équation caractéristique* associée à l'équation différentielle (linéaire, non homogène, à coefficients constants) :

$$P(D)y = b. \qquad (E_b)$$

Nous commencerons par supposer que $K := \mathbb{C}$. Dans ce cas, le polynôme P est scindé :

$$P = (X - \lambda_1)^{r_1} \cdots (X - \lambda_k)^{r_k}.$$

Puisque $P(D)$ est nul sur E, on déduit du lemme des noyaux (module I.5) la décomposition :

$$E = E_1 \oplus \cdots \oplus E_k\,, \qquad E_i := \mathrm{Ker}(D - \lambda_i\,\mathrm{Id})^{r_i} \qquad (1 \leqslant i \leqslant k). \qquad (12)$$

Rappelons que, dans la résolution de l'équation $f' - \lambda f = b$ par « variation des constantes », on pose $f = e^{\lambda t} g$ et l'on se ramène à résoudre $g' = e^{-\lambda t} b$. Le lemme suivant formalise cette idée, en vue d'un usage plus systématique.

Lemme 7. Notons φ_λ ($\lambda \in \mathbb{C}$) l'automorphisme $f \mapsto e^{\lambda t} f$ de $\mathcal{C}^\infty(I, \mathbb{C})$.
Pour $Q \in \mathbb{C}[X]$, notons $Q_\lambda(X) := Q(X - \lambda)$. Alors $Q_\lambda(D) = \varphi_\lambda \circ Q(D) \circ \varphi_\lambda^{-1}$.

Démonstration. C'est trivialement vrai lorsque $Q = a$ est constant, car alors $Q(D) = a\,\mathrm{Id}$ commute avec φ_λ, et $Q_\lambda(D) = a\,\mathrm{Id}$. C'est vrai lorsque $Q(X) := X$; cela revient en effet à l'égalité : $f' - \lambda f = e^{\lambda t}(e^{-\lambda t} f)'$. C'est donc vrai lorsque $Q(X) := X - a$, en vertu du calcul : $\varphi_\lambda \circ (D - a\,\mathrm{Id}) \circ \varphi_\lambda^{-1} = \varphi_\lambda \circ D \circ \varphi_\lambda^{-1} - a\,\mathrm{Id}$.
Enfin, si $Q = RS$, on a $Q_\lambda = R_\lambda S_\lambda$, d'où $Q_\lambda(D) = R_\lambda(D) \circ S_\lambda(D)$. Par ailleurs, $\varphi_\lambda \circ Q(D) \circ \varphi_\lambda^{-1} = \big(\varphi_\lambda \circ R(D) \circ \varphi_\lambda^{-1}\big) \circ \big(\varphi_\lambda \circ S(D) \circ \varphi_\lambda^{-1}\big)$. Si l'égalité est vraie pour R et pour S, elle est donc vraie pour Q. On en conclut qu'elle est vraie pour tout polynôme. ∎

Comme dans toute situation où deux endomorphismes sont conjugués, ce lemme permet de *transporter* les propriétés de l'un vers l'autre. Précisément, supposons que V est un sous–espace vectoriel de $\mathcal{C}^\infty(I, \mathbb{C})$ stable par $Q(D)$. Alors le sous–espace vectoriel $V' := \varphi_\lambda(V) = e^{\lambda t} V$ est stable par $\varphi_\lambda \circ Q(D) \circ \varphi_\lambda^{-1} = Q_\lambda(D)$. On peut donc considérer $Q(D)$ et $Q_\lambda(D)$ comme des endomorphismes respectifs de V et de V', et l'on a un *diagramme commutatif* (module « Fondements » de [L1], paragraphe 2.2.1) :

$$\begin{array}{ccc} V & \xrightarrow{Q(D)} & V \\ \big\downarrow{\scriptstyle \varphi_\lambda} & & \big\downarrow{\scriptstyle \varphi_\lambda} \\ V' & \xrightarrow{Q_\lambda(D)} & V' \end{array} \tag{13}$$

On en déduit aisément $\mathrm{Ker}\, Q_\lambda(D) = \varphi_\lambda\big(\mathrm{Ker}\, Q(D)\big)$, $\mathrm{Im}\, Q_\lambda(D) = \varphi_\lambda\big(\mathrm{Im}\, Q(D)\big)$ et que $Q(D)$ est un automorphisme de V si, et seulement si, $Q_\lambda(D)$ est un automorphisme de V'. Voici une illustration de ce raisonnement.

Proposition 8. Soient $\lambda \in \mathbb{C}$ et $r \in \mathbb{N}$. Alors :

$$\mathrm{Ker}(D - \lambda\,\mathrm{Id})^r = e^{\lambda t}\mathbb{C}_{r-1}[t].$$

Démonstration. D'après le lemme 7, $(D - \lambda\,\mathrm{Id})^r = \varphi_\lambda \circ D^r \circ \varphi_{\lambda^{-1}}$, d'où $\mathrm{Ker}(D - \lambda\,\mathrm{Id})^r = \varphi_\lambda\big(\mathrm{Ker}\, D^r\big)$. On est donc ramené à prouver l'égalité $\mathrm{Ker}\, D^r = \mathbb{C}_{r-1}[t]$, qui est évidente. ∎

Théorème 9.
Le $\mathbb{C}$-espace vectoriel $\mathrm{Sol}_{\mathbb{C}}\big((E_0), \mathbb{R}\big)$ admet pour base la famille $\big(t^j e^{\lambda_i t}\big)_{\substack{1 \leqslant i \leqslant k \\ 0 \leqslant j \leqslant r_i - 1}}$.

Démonstration. D'après la proposition, pour $1 \leqslant i \leqslant k$, la famille $(t^j e^{\lambda_i t})_{0 \leqslant j \leqslant r_i - 1}$ est une base de $E_i := \mathrm{Ker}(D - \lambda_i\,\mathrm{Id})^{r_i}$. Il suffit donc d'invoquer la décomposition (12). ∎

Remarque. Nous n'avons pas démontré que cette base était un système fondamental de solutions de (E_0), c'est-à-dire que son wronskien ne s'annule pas : cela découlera

du corollaire 30 de la page 830 (et le lecteur vérifiera sans peine qu'il n'y a pas de cercle vicieux).

Supposons maintenant que les coefficients a_i du polynôme P soient réels : $P \in \mathbb{R}[X]$. La décomposition en facteurs de degré 1 de P sur le corps algébriquement clos $\mathbb{C}$ prend alors la forme :

$$P = (X-\lambda_1)^{r_1}(X-\overline{\lambda}_1)^{r_1}\cdots(X-\lambda_k)^{r_k}(X-\overline{\lambda}_k)^{r_k}(X-\mu_1)^{s_1}\cdots(X-\mu_\ell)^{s_\ell},$$

où les μ_j sont réels mais pas les λ_j. Écrivons ces derniers $\lambda_j = \alpha_j + \mathrm{i}\beta_j$.

Théorème 10. Le $\mathbb{R}$-espace vectoriel $\mathrm{Sol}_{\mathbb{R}}\big((E_0),\mathbb{R}\big)$ admet pour base la famille formée des $(t^j e^{\alpha_i t}\cos\beta_i t)_{\substack{1\leqslant i\leqslant k\\ 0\leqslant j\leqslant r_i-1}}$, des $(t^j e^{\alpha_i t}\sin\beta_i t)_{\substack{1\leqslant i\leqslant k\\ 0\leqslant j\leqslant r_i-1}}$ et des $(t^j e^{\mu_i t})_{\substack{1\leqslant i\leqslant \ell\\ 0\leqslant j\leqslant s_i-1}}$.

Démonstration.

Cela découle des faits rappelés à la section 1.1 et du théorème 9 de la page précédente. ∎

1.2.1 Solutions de l'équation non homogène (E_b) : cas où b est une exponentielle-polynôme

Les méthodes précédentes permettent de résoudre (E_b) dans le cas où $b = t^p e^{\lambda t}$. Nous ne faisons pas ici d'hypothèse sur le corps K ($\mathbb{R}$ ou $\mathbb{C}$).

Proposition 11. (i) Pour tout $m \in \mathbb{N}$, le K-espace vectoriel $e^{\lambda t}K_m[t]$ est stable par l'endomorphisme $P(D)$.
(ii) Si λ n'est pas racine de P, l'endomorphisme $P(D)$ induit un automorphisme de $e^{\lambda t}K_m[t]$.
(iii) Si λ est racine de P de multiplicité r, l'image de $e^{\lambda t}K_m[t]$ par $P(D)$ est $\{0\}$ si $m < r$, et $e^{\lambda t}K_{m-r}[t]$ sinon.

Démonstration. (i) De la formule $(e^{\lambda t}Q)' = e^{\lambda t}(\lambda Q + Q')$, on déduit que $e^{\lambda t}K_m[t]$ est stable par D, donc par $P(D)$.
(ii) et (iii) On invoque le diagramme commutatif (13) et les propriétés qui s'en déduisent. Pour cela, on prend $V := K_m[t]$ et $Q(X) := P(X+\lambda)$, d'où $Q_\lambda = P$. On est ainsi ramené à étudier l'effet de $Q(D)$ sur $K_m[t]$. Écrivons $Q(X) = X^r R(X)$, où $r \in \mathbb{N}$ et $R(0) \neq 0$. Alors $Q(D) = D^r \circ R(D)$. Comme l'endomorphisme D de $K_m[t]$ est nilpotent (il vérifie $D^{m+1} = 0$), $R(D)$ est inversible, car de la forme $a\,\mathrm{Id} + DS(D)$, où $a := R(0) \neq 0$ et $DS(D)$ est nilpotent. Et, bien sûr, l'image de $K_m[t]$ par D^r est 0 si $r > m$, et $K_{m-r}[t]$ sinon. ∎

Pratiquement, pour résoudre (E_b) lorsque $b := t^p e^{\lambda t}$, on doit distinguer deux cas :

1. Si λ n'est pas une racine de P, on cherche une solution sous la forme $f = e^{\lambda t} g(t)$, avec $g \in K_p[t]$. Pour cela, on écrit $g(t) := g_0 + \cdots + g_p t^p$, on calcule le polynôme $e^{-\lambda t} P(D)(g)$ en fonction des g_i et on l'identifie à $f(t)$ (*méthode des coefficients indéterminés*). La solution $(g_0, \ldots, g_p)$ est unique.

2. Si λ est racine de P de multiplicité r, on cherche une solution sous la forme $f = e^{\lambda t} g$, avec $\deg g = p + r$. Le calcul est le même que précédemment, mais les inconnues sont $(g_0, \ldots, g_{p+r})$ et les solutions forment un espace affine de dimension r.

Le premier cas est d'ailleurs le cas particulier $r := 0$ du second cas.

Point Méthode

(i) Dans le second cas, on peut imposer à g d'être de la forme $g(t) = t^r h(t)$. En effet, les termes de la forme $e^{\lambda t} t^i$ avec $i < r$ sont solutions de l'équation homogène associée et l'on peut donc les soustraire sans modifier la condition (E_b).
(ii) Si b est une combinaison linéaire de termes de la forme indiquée, on résoudra séparément chacune des équations correspondantes et l'on appliquera le « principe de superposition » (page 805).

La nécessité de recourir à un degré plus élevé lorsque λ est l'une des racines de P porte le nom de *résonance*. Sa signification physique est illustrée par l'exemple suivant.

Exercice 6.

Appliquer ce qui précède à l'équation : $\ddot{x} + x = \cos \omega t$. On vérifiera le résultat à l'aide de l'exercice 5 de la page 810.

Solution. Ici, $P(X) = X^2 + 1$, dont les racines sont $\pm \mathrm{i}$ (elles sont simples). On retrouve le fait que $\mathrm{Sol}_{\mathbb{C}}\big((E_0), \mathbb{R}\big) = \mathrm{Vect}_{\mathbb{C}}(e^{\mathrm{i}t}, e^{-\mathrm{i}t})$ et $\mathrm{Sol}_{\mathbb{R}}\big((E_0), \mathbb{R}\big) = \mathrm{Vect}_{\mathbb{R}}(\cos t, \sin t)$.
On résoudra séparément les équations $\ddot{x} + x = e^{\mathrm{i}\omega t}$ (solution f_1) et $\ddot{x} + x = e^{-\mathrm{i}\omega t}$ (solution f_2) : la solution recherchée sera $f := \dfrac{1}{2}(f_1 + f_2)$.
Si $\mathrm{i}\omega, -\mathrm{i}\omega \neq \pm \mathrm{i}$, c'est-à-dire si $\omega \neq \pm 1$, on doit pouvoir trouver des solutions *uniques* de la forme $f_1 = c_1 e^{\mathrm{i}\omega t}$ et $f_2 = c_2 e^{-\mathrm{i}\omega t}$. De fait, on vérifie que $(D^2 + \mathrm{Id}) e^{\pm \mathrm{i}\omega t} = (1 - \omega^2) e^{\pm \mathrm{i}\omega t}$, d'où $c_1 = c_2 = \dfrac{1}{1 - \omega^2}$ et la solution $f := \dfrac{1}{1 - \omega^2} \cos \omega t$ de E_b.
Si $\omega = \pm 1$, on peut aussi bien supposer que $\omega = 1$. Dans ce cas, on devra chercher des solutions de la forme $f_1 = (c_1 t + d_1) e^{\mathrm{i}t}$ et $x_2 = (c_2 t + d_2) e^{-\mathrm{i}t}$. Comme $(D^2 + \mathrm{Id})\big(e^{\pm \mathrm{i}t}\big) = 0$, on peut aussi bien prendre $d_1 = d_2 := 0$ (c'est ce que dit le (i) du « point méthode » ci-dessus). D'autre part, $(D^2 + \mathrm{Id})\big(te^{\mathrm{i}t}\big) = 2\mathrm{i}e^{\mathrm{i}t}$ et $(D^2 + \mathrm{Id})\big(te^{-\mathrm{i}t}\big) = -2\mathrm{i}e^{-\mathrm{i}t}$, d'où $c_1 = -c_2 = \dfrac{1}{2\mathrm{i}}$ et l'on trouve la solution $f := \dfrac{1}{2} t \sin t$.

Appliquons les résultats de l'exercice 5 de la page 810. On cherche x sous la forme $\lambda\cos+\mu\sin$, avec, par exemple, $x(0)=\dot{x}(0)=0$. On prend donc :

$$\lambda(t)=-\int_0^t \cos\omega s\,\sin s\mathrm{d}s=-\int_0^t \frac{\sin(1+\omega)s+\sin(1-\omega)s}{2}\,\mathrm{d}s$$
$$\mu(t)=\int_0^t \cos\omega s\,\cos s\,\mathrm{d}s=\int_0^t \frac{\cos(1+\omega)s+\cos(1-\omega)s}{2}\,\mathrm{d}s.$$

Si $\omega\neq\pm1$, on trouve :

$$\lambda(t)=\frac{\cos(1+\omega)t-1}{2(1+\omega)}+\frac{\cos(1-\omega)t-1}{2(1-\omega)}\qquad\text{et}\qquad\mu(t)=\frac{\sin(1+\omega)t}{2(1+\omega)}+\frac{\sin(1-\omega)t}{2(1-\omega)}\cdot,$$

d'où :

$$\begin{aligned}x(t)&=\frac{\cos(1+\omega)t\cos t-\cos t+\sin(1+\omega)t\sin t}{2(1+\omega)}\\&\qquad+\frac{\cos(1-\omega)t\cos t-\cos t+\sin(1-\omega)t\sin t}{2(1-\omega)}\\&=\frac{\cos\omega t-\cos t}{2(1+\omega)}+\frac{\cos\omega t-\cos t}{2(1-\omega)}=\frac{\cos\omega t-\cos t}{1-\omega^2}\cdot\end{aligned}$$

À une solution de l'équation homogène près, c'est la fonction f trouvée plus haut. Si $\omega=\pm1$, on trouve $\lambda(t)=\dfrac{\cos 2t-1}{4}$ et $\mu(t)=\dfrac{\sin 2t+2t}{4}$, d'où :

$$x(t)=\frac{\cos 2t\,\cos t-\cos t}{4}+\frac{\sin 2t\,\sin t+2t\,\sin t}{4}=\frac{t\,\sin t}{2},$$

soit la solution f trouvée plus haut. Puisque cette solution particulière est non bornée et que les solutions de (E_0) sont toutes bornées, il en découle qu'*aucune solution de* (E_b) *n'est bornée*.

L'équation $\ddot{x}+x=0$ modélise un système physique (pendule, circuit RLC) ayant une *période propre* : toutes ses solutions sont périodiques de période 2π. L'équation $\ddot{x}+x=\cos\omega t$ modélise le même système soumis à une stimulation externe de période $\dfrac{2\pi}{\omega}\cdot$ En général, le système obtenu présente un comportement qui superpose les deux périodicités et *toutes ses solutions sont bornées*. Dans le cas spécial où la stimulation a la même périodicité ($\omega=\pm1$), *toutes les solutions sont non bornées*. Le mouvement *s'amplifie*.

1.2.2 Solutions de l'équation non homogène (E_b) : cas général

On suppose encore les a_i constants, mais b est une fonction continue quelconque. Nous allons montrer que l'équation (E_b) a toujours une solution, et donner une *formule intégrale* explicite pour une telle solution. Remarquons d'abord que, pour tout entier k, la restriction de D est une application linéaire surjective de $\mathcal{C}^{k+1}(\mathbb{R},K)$ sur $\mathcal{C}^k(\mathbb{R},K)$.

Fixons une fois pour toutes $t_0\in\mathbb{R}$. L'application linéaire I qui, à $f\in\mathcal{C}(\mathbb{R},K)$, associe la fonction $t\mapsto\displaystyle\int_{t_0}^t f(s)\,\mathrm{d}s$, envoie $\mathcal{C}^k(\mathbb{R},K)$ dans $\mathcal{C}^{k+1}(\mathbb{R},K)$ et l'on a $D\circ I=\mathrm{Id}$.

Reprenons la notation $\varphi_\lambda : f \mapsto e^{\lambda t} f$. En conjuguant D par l'automorphisme φ_λ de $\mathcal{C}^\infty(\mathbb{R}, K)$. Comme celui-ci laisse stable chaque $\mathcal{C}^k(\mathbb{R}, K)$), on obtient une application linéaire $D_\lambda := \varphi_\lambda \circ D \circ \varphi_\lambda^{-1}$ de $\mathcal{C}^{k+1}(\mathbb{R}, K)$ sur $\mathcal{C}^k(\mathbb{R}, K)$: en fait, $D_\lambda = D - \lambda \operatorname{Id} : f \mapsto f' - \lambda f$.

De même, en conjuguant I par l'automorphisme φ_λ, on obtient une application linéaire $I_\lambda := \varphi_\lambda \circ I \circ \varphi_\lambda^{-1}$ de $\mathcal{C}^k(\mathbb{R}, K)$ dans $\mathcal{C}^{k+1}(\mathbb{R}, K)$. Elle se calcule comme suit :

$$(I_\lambda f)(t) = e^{\lambda t} \int_{t_0}^t e^{-\lambda s} f(s)\, \mathrm{d}s = \int_{t_0}^t e^{\lambda(t-s)} f(s)\, \mathrm{d}s.$$

De plus, $I_\lambda f$ est la seule fonction g telle que $g' - \lambda g = f$ et $g(t_0) = 0$.
En particulier, $D_\lambda \circ I_\lambda$ induit l'application identité de $\mathcal{C}^k(\mathbb{R}, K)$.

Proposition 12. Soit $P := X^n + a_1 X^{n-1} + \cdots + a_n$. L'application $P(D)$ de $\mathcal{C}^n(\mathbb{R}, K)$ sur $\mathcal{C}(\mathbb{R}, K)$ est surjective. l'équation (E_b) (à coefficients a_i constants) a donc toujours des solutions globales.

Démonstration. D'après les faits généraux énumérés à la section 1.1, il suffit d'examiner le cas où $K := \mathbb{C}$. Le polynôme P est alors scindé, ce que l'on écrira ici : $P = (X - \lambda_1) \cdots (X - \lambda_n)$. L'application linéaire $P(D)$ est composée d'applications de la forme $D_{\lambda_i} : \mathcal{C}^{n-i+1}(\mathbb{R}, \mathbb{C}) \to \mathcal{C}^{n-i}(\mathbb{R}, \mathbb{C})$ qui sont toutes surjectives. ∎

On prendra garde que D n'est pas ici considéré comme un endomorphisme. Cependant, la plupart des règles usuelles restent valides.
Par exemple, $(QR)(D) = Q(D) \circ R(D)$, etc.

En fait, l'application $I_{\lambda_1} \circ \cdots \circ I_{\lambda_n}$ de $\mathcal{C}(\mathbb{R}, K)$ dans $\mathcal{C}^n(\mathbb{R}, K)$ associe à b une solution de (E_b). La résolution de l'équation scalaire du $n^{ème}$ ordre $P(D)(y) = b$ est ainsi ramenée à la résolution successive d'équations scalaires du premier ordre $y' - \lambda y = z$; chacune de ces équations se ramène à son tour à une quadrature (c'est-à-dire à une intégration).

Exercice 7.

Appliquer ce qui précède à l'équation : $\ddot{x} + x = \cos \omega t$. On supposera $\omega \neq \pm 1$. (On pourra commencer par résoudre l'équation : $\ddot{x} + x = e^{\mathrm{i}\omega t}$.)

Solution. Pour résoudre (comme suggéré) $\ddot{x} + x = e^{\mathrm{i}\omega t}$, on résoudra successivement $\dot{y} + \mathrm{i}y = e^{\mathrm{i}\omega t}$, puis $\dot{x} - \mathrm{i}x = y$. On prendra $t_0 := 0$. La première équation a pour solution :

$$g(t) = e^{-\mathrm{i}t} \int_0^t e^{\mathrm{i}s} e^{\mathrm{i}\omega s}\, \mathrm{d}s = e^{-\mathrm{i}t} \frac{e^{\mathrm{i}(1+\omega)t} - 1}{\mathrm{i}(1+\omega)} = \frac{e^{\mathrm{i}\omega t} - e^{-\mathrm{i}t}}{\mathrm{i}(1+\omega)}.$$

La deuxième équation a pour solution :

$$f(t) = e^{\mathrm{i}t}\int_0^t e^{-\mathrm{i}s}\frac{e^{\mathrm{i}\omega s}-e^{-\mathrm{i}s}}{\mathrm{i}(1+\omega)}\,\mathrm{d}s = \frac{e^{\mathrm{i}t}}{\mathrm{i}(1+\omega)}\Big(\frac{e^{\mathrm{i}(\omega-1)t}-1}{\mathrm{i}(\omega-1)} - \frac{e^{-2\mathrm{i}t}-1}{-2\mathrm{i}}\Big).$$

Les combinaisons linéaires de $e^{\mathrm{i}t}$, $e^{-\mathrm{i}t}$ étant solutions de l'équation homogène associée, nous retiendrons seulement le premier terme : $\dfrac{e^{\mathrm{i}\omega t}}{1-\omega^2}$.
Pour résoudre l'équation *à coefficients réels* $\ddot{x}+x=\cos\omega t$, il suffit maintenant de prendre la partie réelle de la solution trouvée, soit $\dfrac{\cos\omega t}{1-\omega^2}$.
(Pour le cas où $\omega=\pm1$, voir l'exercice II.7.6 de la page 876.)

1.2.3 Calcul symbolique

Il est possible de donner des formules intégrales tout à fait générales pour des solutions de (E_b). La méthode qui suit se rattache au *calcul symbolique*, étudié en [MPA] avec la transformation de Laplace.

Lemme 13. Soit f une fonction continue. Alors, pour $k\geqslant 1$, la fonction :

$$F(t) := \int_{t_0}^t \frac{(t-s)^{k-1}}{(k-1)!}f(s)\,\mathrm{d}s$$

est l'unique fonction de classe $\mathcal{C}^k$ telle que $F(t_0)=\cdots=F^{(k-1)}(t_0)=0$ et $F^{(k)}=f$.

Démonstration. Puisque If est la primitive de f qui s'annule en t_0, la fonction $F:=I^kf$ obtenue en itérant I vérifie les conditions imposées. On lui applique la formule de Taylor avec reste intégral (module « Approximation » de [L1]) :

$$F(t)=\sum_{i=0}^{k-1}\frac{(t-t_0)^i}{i!}F^{(i)}(t_0)+\int_{t_0}^t\frac{(t-t_0)^{k-1}}{(k-1)!}F^{(k)}(s)\,\mathrm{d}s,$$

autrement dit, $\big(I^kf\big)(t)=F(t)=\displaystyle\int_{t_0}^t\frac{(t-s)^{k-1}}{(k-1)!}f(s)\,\mathrm{d}s$. ∎

En conjuguant par φ_λ, on en déduit la formule :

$$\big(I_\lambda^kf\big)(t)=\int_{t_0}^t\frac{(t-s)^{k-1}}{(k-1)!}e^{\lambda(t-s)}f(s)\,\mathrm{d}s.$$

On prouve sans peine que c'est la seule fonction F telle que :

$$D_\lambda^kF=f\qquad\text{et}\qquad F(t_0)=\cdots=F^{(k-1)}(t_0)=0.$$

L'opérateur I_λ^k nous fournit donc une solution de l'équation $P(D)f=b$ quand le polynôme P admet une unique racine λ.

Lemme 14. On a $I_\lambda\circ I_\mu=\dfrac{1}{\lambda-\mu}(I_\lambda-I_\mu)$.

Démonstration. La fonction $g := \dfrac{1}{\lambda-\mu}(I_\lambda f - I_\mu f)$ est nulle en t_0 et :

$$D_\lambda g = \frac{1}{\lambda-\mu}(f - D_\lambda I_\mu f) = \frac{1}{\lambda-\mu}\big(f - f + (\lambda-\mu)f\big) = f,$$

en vertu des égalités $D_\lambda I_\lambda f = D_\mu I_\mu f = f$ et $D_\lambda(I_\mu f) = D_\mu(I_\mu f) + (\lambda-\mu)(I_\mu f)$.
On a donc bien $g = I_\lambda(I_\mu f)$. ∎

On serait tenté de calculer :

$$\begin{aligned}(D_\lambda \circ D_\mu)g &= \frac{1}{\lambda-\mu}(D_\mu \circ D_\lambda)I_\lambda f - \frac{1}{\lambda-\mu}(D_\lambda \circ D_\mu)I_\mu f\\ &= \frac{1}{\lambda-\mu}D_\lambda f - \frac{1}{\lambda-\mu}D_\mu f = f,\end{aligned}$$

puisque $D_\lambda \circ D_\mu = D_\mu \circ D_\lambda$. Malheureusement, cela n'a pas de sens en général puisque f est seulement supposée continue : la fonction g est de classe $\mathcal{C}^2$, mais les deux termes qui la composent sont *a priori* seulement de classe $\mathcal{C}^1$.

Soit $P := \prod_{i=1}^{k}(X-\lambda_i)^{r_i}$ un polynôme de degré $n := \sum_{i=1}^{k} r_i \geqslant 1$. Alors $F := I_{\lambda_1}^{r_1}\cdots I_{\lambda_k}^{r_k} f$ est l'unique fonction de classe $\mathcal{C}^n$ telle que $P(D)F = f$ et $F(t_0) = \cdots = F^{(n-1)}(t_0) = 0$. Nous allons voir comment la calculer effectivement.

Proposition 15. Il existe une famille $(a_{i,j})_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}}$ de n constantes telle que, quelle que soit f continue, on ait :

$$I_{\lambda_1}^{r_1}\cdots I_{\lambda_k}^{r_k} f = \sum_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}} a_{i,j} I_{\lambda_i}^{j} f.$$

Démonstration. Il s'agit de deux récurrences imbriquées. Pour $k := 1$, l'assertion est évidente. Pour passer de k à $k+1$, on doit calculer, pour un $\mu \notin \{\lambda_1,\dots,\lambda_k\}$ et un $r \geqslant 1$, l'opérateur composé $I_\mu^s \circ I_{\lambda_1}^{r_1}\cdots I_{\lambda_k}^{r_k}$. on le fait par récurrence sur s. Si $s := 1$, cela découle de l'hypothèse de récurrence et du lemme 14. Sinon, c'est immédiat par linéarité. ∎

Déterminons les coefficients $a_{i,j}$. Pour cela, on prend f de classe $\mathcal{C}^\infty$: les $I_{\lambda_i}^j f$ le sont donc aussi. Si l'on pose $Q_{i,j} := \dfrac{P}{(X-\lambda_i)^j}$, qui est un polynôme pour $j \leqslant r_i$, on a :

$$P(D)(I_{\lambda_i}^j f) = Q_{i,j}(D)(D_{\lambda_i}^j I_{\lambda_i}^j f) = Q_{i,j}(D)f.$$

On applique alors $P(D)$ à l'égalité de la proposition, et l'on obtient :

$$f = P(D)I_{\lambda_1}^{r_1}\cdots I_{\lambda_k}^{r_k} f \sum_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}} a_{i,j}P(D)(I_{\lambda_i}^j f) = \sum_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}} a_{i,j}Q_{i,j}(D)f.$$

Autrement dit, si l'on note $R := 1 - \sum_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}} a_{i,j}Q_{i,j}$, l'opérateur différentiel $R(D)$ annule toutes

les fonctions de classe $\mathcal{C}^\infty$. Cela n'est possible que si $R = 0$. On a donc :

$$1 = \sum_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}} a_{i,j}Q_{i,j} = \sum_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}} a_{i,j}\frac{P}{(X-\lambda_i)^j}, \text{ d'où : } \frac{1}{P} = \sum_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}} a_{i,j}\frac{1}{(X-\lambda_i)^j}.$$

Nous reconnaissons la *décomposition en éléments simples* de la fraction rationnelle $\dfrac{1}{P}$ (module sur les polynômes de [L1]). Avec ces notations, nous avons démontré :

Théorème 16. L'unique solution de (E_b) telle que $f(t_0) = \cdots = f^{(n-1)}(t_0) = 0$ est la fonction :

$$f := \sum_{\substack{1\leqslant i\leqslant k\\ 1\leqslant j\leqslant r_i}} a_{i,j}\int_{t_0}^{t}\frac{(t-s)^{j-1}}{(j-1)!}e^{\lambda_i(t-s)}b(s)\,\mathrm{d}s.$$

1.3 Vectorialisation et existence des solutions globales

1.3.1 Vectorialisation

À l'équation (E_b), nous allons associer une matrice et un vecteur colonne de fonctions :

$$A(t) := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0\\ 0 & 0 & 1 & \dots & 0 & 0\\ 0 & 0 & 0 & \dots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \dots & 0 & 1\\ -a_n & -a_{n-1} & -a_{n-2} & \dots & -a_2 & -a_1 \end{pmatrix} \quad \text{et} \quad B(t) := \begin{pmatrix} 0\\ 0\\ 0\\ \vdots\\ 0\\ b(t) \end{pmatrix}. \tag{14}$$

Ainsi, A et B sont des fonctions vectorielles continues sur I, à valeurs respectivement dans $M_n(K)$ et dans K^n (ici identifié à $M_{n,1}(K)$.

Soit maintenant f une fonction de classe $\mathcal{C}^n$ sur l'intervalle $J \subset I$ et notons $F(t)$ le vecteur colonne $\begin{pmatrix} f(t)\\ \vdots\\ f^{(n-1)}(t) \end{pmatrix}$: c'est donc une fonction vectorielle de classe $\mathcal{C}^1$ sur J.

Proposition 17. Pour que la fonction f soit une solution de l'équation (E_b) sur $J \subset I$, il faut, et il suffit, que l'on ait la relation :

$$\forall t \in J\ ,\ F'(t) = A(t)F(t) + B(t). \tag{15}$$

Pour que la fonction f soit une solution de l'équation (E_0) sur $J \subset I$, il faut, et il suffit, que l'on ait la relation :

$$\forall t \in J\ ,\ F'(t) = A(t)F(t). \tag{16}$$

Démonstration. On a ${}^t\big(A(t)F(t)\big) = (f'(t), \dots, f^{(n-1)}(t), -a_nf(t) - \cdots - a_1f^{(n-1)}(t))$, d'où ${}^t\big(F'(t) - A(t)F(t)\big) = (0, \dots, 0, f^{(n)}(t) + a_1(t)f^{(n-1)}(t) + \cdots + a_n(t)f(t))$, ce qui justifie les deux équivalences logiques. ∎

Définition 6. Une *équation différentielle linéaire vectorielle d'ordre* 1, également appelée *système différentiel linéaire de rang* n sur l'intervalle [1] $I \subset \mathbb{R}$ est une équation de la forme :

$$(S_B) \quad Y' = AY + B, \tag{17}$$

où A et B sont des fonctions continues sur I respectivement à valeurs dans $M_n(K)$ et dans $M_{n,1}(K) = K^n$. Le *système différentiel homogène associé* est le système :

$$(S_0) \quad Y' = AY. \tag{18}$$

Pour tout intervalle ouvert $J \subset I$, on appelle *solution* sur J du système différentiel (S_B), resp. du système différentiel homogène (S_0), toute fonction $F : J \to K^n$ de classe $\mathcal{C}^1$ qui satisfait à la relation (15), resp. à la relation (16).
L'ensemble des solutions sur J du système différentiel (S_B), resp. (S_0), sera noté $\mathrm{Sol}_K\big((S_B), J\big)$, resp. $\mathrm{Sol}_K\big((S_0), J\big)$.

Si l'on note $A := (a_{i,j})$ et $B := {}^t(b_1, \dots, b_n)$, et $Y := {}^t(y_1, \dots, y_n)$, on a l'équivalence suivante, qui explique le terme de « système différentiel » :

$$(S_B) \iff \begin{cases} y_1' = a_{1,1}y_1 + \cdots + a_{1,n}y_n + b_1, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n' = a_{n,1}y_1 + \cdots + a_{n,n}y_n + b_n. \end{cases} \tag{19}$$

Remarque. Soit V un K-espace vectoriel de dimension finie. Par extension, on peut définir une équation différentielle linéaire d'ordre 1 dans V. Pour cela, on considère une application continue $t \mapsto A(t)$ de I dans $\mathcal{L}_K(V)$ et une application continue $t \mapsto A(t)$ de I dans V (section 4 de la page 440). L'équation (S_B) admet pour solutions sur J des fonctions $F : J \to V$ de classe $\mathcal{C}^1$ qui satisfont à la relation (15) de la page 819.

Exercice 8.

Soit P une fonction de classe $\mathcal{C}^1$ sur I, à valeurs dans $GL_n(K)$. Montrer que la « transformation de jauge » $X = PY$ transforme le système $X' = AX$ en un système $Y' = BY$ que l'on explicitera. Que devient cette transformation si P est constante, $P \in GL_n(K)$?

Solution. La relation $X' = AX$ se traduit en $PY' + P'Y = APY$, qui équivaut à $Y' = BY$, où $B := P^{-1}AP - P^{-1}P'$. Dans le cas particulier où P est constante, on trouve simplement une relation de similitude : $B := P^{-1}AP$.

Comme dans le cas des équations scalaires, la restriction à $J' \subset J$ d'une solution sur J est une solution sur J'. Réciproquement, nous verrons (corollaire 28 de la page 829) que toute solution sur un intervalle J peut être prolongée en une solution *globale*, c'est-à-dire définie sur I. Insistons encore sur le fait que cette propriété est caractéristique des équations et des systèmes *linéaires*.

[1] Relire à ce sujet le « point méthode » de la page 804.

Exercice 9.

Montrer que F est solution de (S_B) si, et seulement si, $G := \begin{pmatrix} F \\ 1 \end{pmatrix}$ est solution d'un certain système linéaire homogène de rang $n+1$, que l'on précisera.

Solution.

Il suffit de prendre le système de matrice $\begin{pmatrix} A & B \\ 0_n & 0 \end{pmatrix}$, où 0_n désigne le zéro de $M_{1,n}(K)$.

Comme dans le cas des équations scalaires, les faits suivants sont évidents :

1. Il suffit qu'une solution F soit dérivable pour qu'elle soit de classe $\mathcal{C}^1$. Plus généralement, si A et B sont de classe $\mathcal{C}^k$ (resp. de classe $\mathcal{C}^\infty$), toute solution est de classe $\mathcal{C}^{k+1}$ (resp. de classe $\mathcal{C}^\infty$).
2. L'ensemble $\mathrm{Sol}_K\big((S_B), J\big)$ est soit vide soit un espace affine de direction $\mathrm{Sol}_K\big((S_0), J\big)$. Plus précisément, si F_0 est une solution particulière de (S_B), les solutions de (S_B) sont les $F_0 + F$, où F est solution de (S_0).
3. Si A est à valeurs réelles, F est une solution complexe de (S_B) si, et seulement si, $\mathrm{Re}\, F$ est une solution réelle de $(S_{\mathrm{Re}\, B})$ et $\mathrm{Im}\, F$ une solution réelle de $(S_{\mathrm{Im}\, B})$.
4. Pour résoudre (S_B) lorsque $B = \sum \lambda_i B_i$, il suffit de résoudre chaque (S_{B_i}) puis de former la même combinaison linéaire des solutions.

Proposition 18. L'application $f \mapsto {}^t(f, \dots, f^{(n-1)})$ est un isomorphisme affine de $\mathrm{Sol}_K\big((E_b), J\big)$ sur $\mathrm{Sol}_K\big((S_B), J\big)$ et un isomorphisme linéaire de $\mathrm{Sol}_K\big((E_0), J\big)$ sur $\mathrm{Sol}_K\big((S_0), J\big)$.

Démonstration. Si $F := {}^t(f_1, \dots, f_n)$ est une fonction vectorielle qui vérifie (15) ou (16), pour $1 \leqslant i \leqslant n-1$, on a $f_{i+1} = f_1'$ et F est de la forme ${}^t(f, \dots, f^{(n-1)})$. Il suffit alors d'appliquer la proposition 17 de la page 819. ∎

Les équations peuvent donc être transformées en systèmes qui leur sont équivalents. Dans une certaine mesure, la réciproque est vraie, ce que nous nous contenterons d'illustrer dans un cas particulier.

Exercice 10.

On prend $A := \begin{pmatrix} a & b \\ d & e \end{pmatrix}$, $B := \begin{pmatrix} c \\ f \end{pmatrix}$ et $Y := \begin{pmatrix} x \\ y \end{pmatrix}$. On suppose a, b, c, d, e, f de classe $\mathcal{C}^\infty$. Ramener la résolution du système $Y' = AY + B$ à celle d'une équation scalaire d'ordre 2. (On considérera des intervalles ouverts où b ne s'annule pas, des intervalles ouverts où d ne s'annule pas, enfin, des intervalles ouverts où b et d sont identiquement nuls.)

Solution. Le système $Y' = AY + B$ équivaut à : $\begin{cases} x' = ax + by + c, \\ y' = dx + ey + f. \end{cases}$ Sur un intervalle

où b ne s'annule pas, on peut écrire $y = \dfrac{x' - ax - c}{b}$ et substituer dans la deuxième équation, ce qui nous ramène à l'équation linéaire scalaire du second ordre en x :

$$\left(\frac{x' - ax - c}{b}\right)' = dx + e\frac{x' - ax - c}{b} + f$$
$$\iff \quad x'' - \left(a + \frac{b'}{b} + e\right)x' + \left(ea + \frac{b'}{b}a - a' - bd\right)x = c' + bf - ce - c\frac{b'}{b}\cdot$$

Une fois cette équation résolue, on en tire $y = \dfrac{x' - ax - c}{b}\cdot$ De même, sur un intervalle où d ne s'annule pas, on peut écrire $x = \dfrac{y' - ey - f}{d}$, reporter dans la première équation et se ramener à une équation du second ordre en y. Enfin, sur un intervalle où b et d sont nuls, on est ramené à deux équations scalaires du premier ordre : $x' = ax + c$ et $y' = ey + f$. Cette solution ne prétend pas être complète : voir l'exercice II.7.9 de la page 876.

1.3.2 Solutions matricielles fondamentales

Soit A une fonction continue sur I à valeurs dans $M_n(K)$. Pour tout $t \in I$, l'application $M \mapsto A(t)M$ est un endomorphisme de l'espace vectoriel $M_n(K)$ et l'application qui, à t, associe cet endomorphisme est continue de I dans $\mathcal{L}\big(M_n(K)\big)$. En accord avec la remarque de la page 820, on peut donc définir une équation différentielle linéaire d'ordre 1 *matricielle* $\mathcal{Y}' = A\mathcal{Y}$. Il est clair que la fonction $\mathcal{F} := [F_1; \dots; F_n]$ de J dans $M_n(K)$ est solution de cette équation si, et seulement si, ses colonnes $F_1, \dots, F_n$ sont solutions de l'équation (S_0). Nous considérerons une telle fonction $\mathcal{F}$ comme une *solution matricielle de l'équation vectorielle homogène* (S_0).

Par extension, on appelle *wronskien* de la solution matricielle $\mathcal{F}$ la fonction $W := \det \mathcal{F}$. Nous allons adapter à ce contexte le théorème 3 de la page 807 et ses conséquences (systèmes fondamentaux de solutions).

Théorème 19. Soit $\mathcal{F}$ une solution matricielle de (S_0) sur J. Alors son wronskien W vérifie :

$$W' = \mathrm{Tr}(A)W. \tag{20}$$

Démonstration. Écrivons $\mathcal{F} := [F_1; \dots; F_n]$, d'où $W = \det(F_1, \dots, F_n)$. Du lemme 1 de la page 807 et de l'exercice 7 de la page 108, on déduit :

$$W' = \sum_{i=1}^{n} \det(F_1, \dots, F_i', \dots, F_n) = \sum_{i=1}^{n} \det(F_1, \dots, AF_i, \dots, F_n) = \mathrm{Tr}(A)W.$$

Noter que le théorème 3 de la page 807 est un cas particulier de celui-ci, puisque la trace de la matrice $A(t)$ décrite dans (14) est $-a_1(t)$.

Corollaire 20. Le wronskien d'une solution matricielle sur un intervalle ouvert J est identiquement nul sur J ou bien ne s'y annule pas.

Définition 7. Une *solution matricielle fondamentale* du système homogène (S_0) sur J est une solution matricielle dont le wronskien ne s'annule pas.

Théorème 21. Soit $\mathcal{F}$ une solution matricielle fondamentale de (S_0) sur J. Alors :

$$\mathrm{Sol}_K\big((S_0), J\big) = \{\mathcal{F}C \mid C \in K^n\}. \tag{21}$$

En particulier, les colonnes de $\mathcal{F}$ forment une base de $\mathrm{Sol}_K\big((S_0), J\big)$.

Démonstration. Pour toute fonction F de classe $\mathcal{C}^1$ de J dans K^n, la fonction $C := \mathcal{F}^{-1}F$ est de classe $\mathcal{C}^1$. En dérivant la relation $F = \mathcal{F}C$ et en remplaçant $\mathcal{F}'$ par $A\mathcal{F}$, on voit que $F' - AF = \mathcal{F}C'$. Par conséquent, F est solution de (S_0) si, et seulement si, $C' = 0$.
On sait d'autre part (module de calcul matriciel de [L1]) que les $\mathcal{F}C$ sont les combinaisons linéaires des colonnes de $\mathcal{F}$. L'existence et l'unicité du $C \in K^n$ ci-dessus donnent la seconde assertion. ∎

Corollaire 22. Soit $t_0 \in J$. S'il existe une solution matricielle fondamentale, l'application « conditions initiales » $F \mapsto F(t_0)$ est un isomorphisme de $\mathrm{Sol}_K\big((S_0), J\big)$ sur K^n.

Démonstration. La linéarité est évidente. L'unique solution F correspondant à une condition initiale $Y_0 \in K^n$ est la solution $F := \mathcal{F}C$, où $C := \mathcal{F}^{-1}(t_0)Y_0$. ∎

Application à la résolution du système non homogène. Pour déterminer une solution particulière du système non homogène (S_B) sur J à partir d'une solution matricielle fondamentale de (S_0), on utilise encore la méthode de variation des constantes. On cherche la solution F sous la forme $\mathcal{F}C$, de sorte que (comme on l'a vu plus haut), $F' - AF = \mathcal{F}C'$. Pour que F soit solution de (S_B), il est donc nécessaire et suffisant que $C' = \mathcal{F}^{-1}B$. On a donc ramené la résolution de (S_B) à une quadrature.

1.3.3 Solutions matricielles fondamentales des systèmes à coefficients constants

Nous allons étudier le système (S_0) lorsque $A \in M_n(K)$. Soit $t_0 \in \mathbb{R}$. L'équation matricielle $\mathcal{Y}' = A\mathcal{Y}$ admet en particulier la solution matricielle $\mathcal{F}(t) := e^{(t-t_0)A}$ sur $\mathbb{R}$, qui satisfait à la condition initiale $\mathcal{F}(t_0) = I_n$. L'unique solution F (vectorielle) de (S_0) qui satisfait à la condition initiale $F(t_0) = Y_0$ est donc $F := e^{(t-t_0)A}Y_0$.

Pour résoudre (S_B), nous chercherons une solution sous la forme $F := e^{(t-t_0)A}G$ (variation des constantes). On est ramené à l'équation $e^{(t-t_0)A}G' = B$, donc à $G' = e^{-(t-t_0)A}B$,

donc à une quadrature :

$$G(t) := Y_0 + \int_{t_0}^{t} e^{-(s-t_0)A} B(s)\,\mathrm{d}s,$$

d'où l'unique solution de condition initiale $F(t_0) = Y_0$:

$$F(t) := e^{(t-t_0)A} Y_0 + e^{(t-t_0)A} \int_{t_0}^{t} e^{-(s-t_0)A} B(s)\,\mathrm{d}s = e^{(t-t_0)A} Y_0 + \int_{t_0}^{t} e^{(t-s)A} B(s)\,\mathrm{d}s.$$

Exercice 11.
Résoudre l'équation $\ddot{x} - 2\dot{x} + x = t$, avec la condition initiale $f(0) = f'(0) = 0$.

Solution. Par vectorialisation, on est ramené au système (S_B), avec $A := \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}$ et $B := \begin{pmatrix} 0 \\ t \end{pmatrix}$. Pour calculer $\exp(A)$, on utilise en général sa décomposition de Dunford sous la forme (1), page 279. La matrice A est ici unipotente et l'on a $A = I_2 + N$, où la matrice $N := \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}$ est de carré nul, d'où :

$$e^{tA} = e^t(I_2 + tN) = e^t \begin{pmatrix} 1-t & t \\ -t & 1+t \end{pmatrix}.$$

Le lecteur vérifiera que les colonnes de cette matrice sont bien de la forme $\begin{pmatrix} f \\ \dot{f} \end{pmatrix}$, où f est une solution de l'équation homogène $\ddot{x} - 2\dot{x} + x = 0$.
On obtient enfin la solution cherchée de (S_B) sous la forme :

$$\begin{aligned}
\begin{pmatrix} f \\ \dot{f} \end{pmatrix} &:= e^t \begin{pmatrix} 1-t & t \\ -t & 1+t \end{pmatrix} \int_0^t e^{-s} \begin{pmatrix} 1+s & -s \\ s & 1-s \end{pmatrix} \begin{pmatrix} 0 \\ s \end{pmatrix} \mathrm{d}s \\
&= e^t \begin{pmatrix} 1-t & t \\ -t & 1+t \end{pmatrix} \int_0^t \begin{pmatrix} -s^2 e^{-s} \\ (s-s^2)e^{-s} \end{pmatrix} \mathrm{d}s \\
&= e^t \begin{pmatrix} 1-t & t \\ -t & 1+t \end{pmatrix} \begin{pmatrix} (t^2+2t+2)e^{-t} - 2 \\ (t^2+t+1)e^{-t} - 1 \end{pmatrix} \\
&= \begin{pmatrix} (t-2)e^t + (t+2) \\ (t-1)e^t + 1 \end{pmatrix}.
\end{aligned}$$

On a donc $f(t) := (t-2)e^t + (t+2)$, et l'on vérifie sans peine que $f'(t) = (t-1)e^t + 1$, $f(0) = f'(0) = 0$ et $f - 2f' + f'' = t$.

Remarque. Soit $P := X^n + a_1 X^{n-1} + \cdots + a_n$. On suppose les a_i constants. La matrice A du système associé à l'équation $P(D)f = b$ (équation (14) page 819) a pour polynôme caractéristique P. Lorsque l'on calcule $e^{(t-s)A}$, il apparaît des coefficients de la forme $(t-s)^i e^{\lambda_j (t-s)}$, pour chaque valeur propre λ_j de multiplicité r_j de

A et chaque entier $i < r_j$. Par conséquent, l'expression $\displaystyle\int_{t_0}^{t} e^{(t-s)A}B(s)\,\mathrm{d}s$ fait apparaître des termes $\displaystyle\int_{t_0}^{t}(t-s)^i e^{\lambda_j(t-s)}b(s)\,\mathrm{d}s$: comme les λ_j sont justement les racines de P, avec multiplicité r_j, on reconnaît les formules de résolution des sections 1.2.2 et 1.2.3.

Point Méthode

La *théorie de la réduction* (module I.5) est ici un outil indispensable :

1. Soient $A, B \in M_n(K)$ et $P \in GL_n(K)$. Si $B := P^{-1}AP$, les solutions du système $X' = AX$ s'obtiennent à partir de celles du système $Y' = BY$ par la transformation $X = PY$ (exercice 8 de la page 820). On a donc intérêt à réduire la matrice A à une forme B plus simple.
2. Dans le cas d'une matrice diagonale $A := \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$, une matrice fondamentale de solutions est $\mathrm{Diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_n t})$.
3. En combinant les deux faits précédents, on obtient la version plus géométrique suivante : si $V_1, \ldots, V_n$ forment une base de vecteurs propres de A (associés aux valeurs propres $\lambda_1, \ldots, \lambda_n$), alors $e^{\lambda_1 t}V_1, \ldots, e^{\lambda_n t}V_n$ forment un système fondamental de solutions de $X' = AX$.
4. Plus généralement, si l'on dispose d'un vecteur propre V de A, associé à la valeur propre λ, alors $e^{\lambda t}V$ est une solution de $X' = AX$.

Utilisation de la géométrie euclidienne. L'étude du champ électromagnétique fait apparaître des équations dans l'espace vectoriel euclidien orienté $\mathbb{R}^3$, de la forme :

$$Y' = U \wedge Y + V,$$

où $U, V \in \mathbb{R}^3$ sont des vecteurs constants et où $U \wedge Y$ désigne le produit vectoriel. On supposera $U, V \neq 0$. Comme l'application $Y \mapsto U \wedge Y$ est linéaire, il s'agit d'un système à coefficients constants. Pour résoudre ce système, on commencera pas décomposer V en $V_1 + V_2$ et Y en $Y_1 + Y_2$, où $Y_1, V_1 \in \mathrm{Vect}(U)$ et $Y_2, V_2 \perp U$. Comme les projections sont des applications linéaires, Y_1 et Y_2 sont de classe $\mathcal{C}^1$ et notre système équivaut à :

$$Y_1' = V_1 \quad \text{et} \quad Y_2' = U \wedge Y_2 + V_2.$$

Les solutions de la première équation sont de la forme $tV_1 + Y_1(0)$. Puisque $U \perp V_2$, il existe $P \in SO_3(\mathbb{R})$ et $\lambda, \mu \in \mathbb{R}_+$ tels que $U = \lambda P e_3$ et $V = \mu P e_1$, où l'on a noté (e_1, e_2, e_3) la base canonique de $\mathbb{R}^3$. Notant $v := Pe_1$, $w := Pe_2$ et $u := Pe_3$, on voit que $u = \dfrac{U}{\|U\|}$, $v = \dfrac{V}{\|V\|}$, et (v, w, u) est une base orthonormée directe (donc $w = u \wedge v$). On écrira de plus $V_1 = \nu u$.

On opère le changement de fonction inconnue $Y_2 = PZ$, ce qui nous ramène à l'équation :

$$Z' = \lambda e_3 \wedge Z + \mu e_1.$$

Puisque $Z' \perp e_3$, le produit scalaire $Z.e_3$ est constant et l'on a $Z = Z_0 + X$, où Z_0 est constant et où X est à valeurs dans $\mathrm{Vect}(e_1, e_2)$. Identifions ce dernier plan orienté à $\mathbb{C}$. L'équation satisfaite par X prend la forme :

$$X' = \lambda \mathrm{i} X + \mu,$$

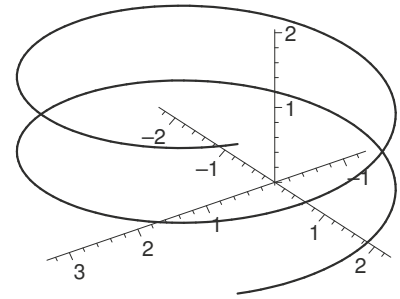
dont les solutions sont de la forme $\alpha e^{\lambda \mathrm{i} t} + \mu$, soit, en termes vectoriels :

$$(a\cos\lambda t - b\sin\lambda t + \mu)e_1 + (a\sin\lambda t + b\cos\lambda t)e_2.$$

En fin de compte, les solutions de notre équation ont la forme :

$$\begin{aligned} F(t) =& (a\cos\lambda t - b\sin\lambda t + \mu)v \\ &+ (a\sin\lambda t + b\cos\lambda t)w + \nu t u + F_0, \end{aligned}$$

soit un mouvement hélicoïdal.
Pour la figure ci-contre, on a pris $a := 1$, $b := 2$, $\lambda := 2\pi$, $\mu := 1$, $\nu := 1$ et $F_0 := 0$.



2 Existence et comportement des solutions

2.1 Le théorème de Cauchy-Lipschitz linéaire

2.1.1 Le théorème de Cauchy-Lipschitz linéaire vectoriel

Soit $A : I \to M_n(K)$ une fonction matricielle continue.

Lemme 23. Soit J un intervalle et soit $t_0 \in J$. L'équation vectorielle d'ordre 1 avec condition initiale :

$$\begin{cases} Y' = AY, \\ Y(t_0) = Y_0 \end{cases} \tag{22}$$

est équivalente à l'*équation intégrale* :

$$Y(t) = Y_0 + \int_{t_0}^{t} A(s)Y(s)\,\mathrm{d}s. \tag{23}$$

Plus précisément, les fonctions dérivables vérifiant (22) sont les fonctions continues vérifiant (23).

Démonstration. Soit F une solution sur J de l'équation (22). Alors :

$$\forall t \in J\ ,\ \int_{t_0}^{t} A(s)F(s)\,\mathrm{d}s = \int_{t_0}^{t} F'(s)\,\mathrm{d}s = F(t) - F(t_0) = F(t) - Y_0,$$

et F est solution sur J de l'équation (23).
Réciproquement, si F est solution sur J de (23), on a d'abord $F(t_0) = Y_0$, puis $F'(t) = A(t)F(t)$, et F est solution sur J de (22). ∎

Ce lemme permet de ramener la résolution d'une équation différentielle à une *recherche de point fixe*. Il vaut aussi bien pour des solutions matricielles que vectorielles. Il sera commode de l'appliquer d'abord à des solutions matricielles. Pour toute fonction matricielle continue $\mathcal{F} : I \to M_n(K)$, nous définirons la fonction matricielle $\mathcal{G} = \Phi(\mathcal{F})$ par la formule :

$$\forall t \in I\ ,\ \mathcal{G}(t) := \int_{t_0}^{t} A(s)\mathcal{F}(s)\,\mathrm{d}s.$$

La fonction $\mathcal{G} : I \to M_n(K)$ est évidemment continue, et Φ est un endomorphisme du K-espace vectoriel $\mathcal{C}\big(I, M_n(K)\big)$.

Remarque. Nous cherchons $\mathcal{F}$ telle que $\mathcal{F} = Y_0 + \Phi(\mathcal{F})$, ce que l'on peut écrire $(\mathrm{Id} - \Phi)(\mathcal{F}) = Y_0$. Il est alors tentant d'inverser $\mathrm{Id} - \Phi$ en écrivant : $(\mathrm{Id} - \Phi)^{-1} = \mathrm{Id} + \Phi + \Phi^2 + \cdots$

La démonstration du lemme qui suit s'inspire de cette idée, sans lui donner toutefois une signification formelle. Des idées similaires (*méthode de Picard*) apparaissent dans les problèmes de point fixe (module II.1) en analyse numérique (module II.8).

Lemme 24. Soit $\|-\|$ une norme subordonnée sur $M_n(K)$ (module II.1).
Soient $J \subset I$ un intervalle, $\mathcal{F}$ une fonction matricielle continue sur J et $C, D > 0$ tels que :

$$\forall t \in J\ ,\ \|\mathcal{F}(t)\| \leqslant C \quad \text{et} \quad \|A(t)\| \leqslant D.$$

Alors :

$$\forall k \in \mathbb{N}\ ,\ \forall t \in J\ ,\ \left\|\Phi^k(\mathcal{F})(t)\right\| \leqslant CD^k \frac{|t - t_0|^k}{k!}.$$

Démonstration. Pour $k := 0$, l'inégalité est contenue dans les hypothèses. Supposons-la vérifiée pour un entier k. Soit $t \in J$; nous supposerons d'abord que $t \geqslant t_0$. Alors :

$$\begin{aligned}\left\|\Phi^{k+1}(\mathcal{F})(t)\right\| &\leqslant \int_{t_0}^{t} \|A(s)\| \left\|\Phi^k(\mathcal{F})(s)\right\| \mathrm{d}s \\ &\leqslant \int_{t_0}^{t} DCD^k \frac{(s - t_0)^k}{k!}\,\mathrm{d}s = CD^{k+1}\frac{(t - t_0)^{k+1}}{(k+1)!}.\end{aligned}$$

Lorsque $t \leqslant t_0$, il faut faire attention aux signes :

$$\begin{aligned}\left\|\Phi^{k+1}(\mathcal{F})(t)\right\| &\leqslant \int_{t}^{t_0} \|A(s)\| \left\|\Phi^k(\mathcal{F})(s)\right\| \mathrm{d}s \\ &\leqslant \int_{t}^{t_0} DCD^k \frac{(t_0 - s)^k}{k!}\,\mathrm{d}s = CD^{k+1}\frac{(t_0 - t)^{k+1}}{(k+1)!}.\end{aligned}$$

Voici, enfin, le premier résultat d'analyse de ce module !

Théorème et définition 25. L'équation matricielle avec condition initiale :

$$\begin{cases} \mathcal{Y}' = A\mathcal{Y}, \\ \mathcal{Y}(t_0) = I_n \end{cases} \tag{24}$$

admet une unique solution globale sur I. Cette solution est une solution matricielle fondamentale. Nous noterons $R(t, t_0)$ cette solution (la notation ne fait donc pas référence au système concerné) et nous l'appellerons *résolvante* de l'équation vectorielle (24).

Démonstration. Il s'agit, d'après le lemme 23, dont nous reprenons les notations, de trouver une fonction matricielle $\mathcal{F}$ telle que $\mathcal{F} = I_n + \Phi(\mathcal{F})$.
Supposons d'abord que l'on a trouvé une telle fonction $\mathcal{F}$. En itérant cette relation, on obtient l'égalité :

$$\forall k \in \mathbb{N}^* \ , \ \mathcal{F} = I_n + \Phi(I_n) + \cdots + \Phi^{k-1}(I_n) + \Phi^k(\mathcal{F}).$$

En effet, c'est évident pour $k := 1$ (c'est l'hypothèse) et, si c'est vrai pour k, en remplaçant $\mathcal{F}$ par $I_n + \Phi(\mathcal{F})$, on voit que c'est encore vrai pour $k+1$; la conclusion vient par récurrence. Nous allons majorer le « reste » $\Phi^k(\mathcal{F})$. On fixe un intervalle ouvert $J :=]a, b[$ tel que $[a, b] \subset I$. Il existe une constante $C > 0$ telle que $\|\mathcal{F}\| \leqslant C$ sur $[a, b]$ (fonction continue sur un compact), donc sur J. D'après le lemme 24, pour tout $k \in \mathbb{N}$, on a :

$$\forall t \in J \ , \ \left\|\Phi^k(\mathcal{F})(t)\right\| \leqslant CD^k \frac{|t - t_0|^k}{k!} \leqslant CD^k \frac{|b - a|^k}{k!},$$

puisque $|t - t_0| \leqslant |b - a|$ sur J. On en déduit d'abord que $\Phi^k(\mathcal{F})$ tend uniformément vers 0 sur J (pour tout $r > 0$, la suite des $\dfrac{r^k}{k!}$ tend vers 0), donc l'égalité :

$$\mathcal{F} = \sum_{k \geqslant 0} \Phi^k(I_n),$$

où la convergence de la série est uniforme sur tout compact de I. S'il y a une solution, elle est donc unique et c'est la somme de cette série de fonctions matricielles.
Réciproquement, la même majoration déduite du lemme, appliquée à $\Phi^k(I_n)$ (donc avec $C := 1$) implique que la série $\sum_{k \geqslant 0} \Phi^k(I_n)$ converge normalement sur tout compact de I (car la série des $\dfrac{r^k}{k!}$ converge absolument). Sa somme est une fonction continue $\mathcal{F}$, et un calcul simple montre que $\mathcal{F} = I_n + \Phi(\mathcal{F})$.
Il reste à voir que $\mathcal{F}$ est une solution matricielle fondamentale. Mais son déterminant vaut 1 en t_0 et cela découle donc de la section 1.3.2. ∎

Corollaire 26. Soit D un majorant de $\|A(t)\|$ sur un intervalle $J \subset I$ qui contient t_0. Alors :

$$\forall t \in J \ , \ \|R(t, t_0) - I_n\| \leqslant e^{D|t - t_0|} - 1. \tag{25}$$

Démonstration. La fonction matricielle $R(t, t_0) - I_n$ est la somme des $\Phi^k(I_n)$ pour $k \geqslant 1$, et chaque norme $\left\|\Phi^k(I_n)\right\|$ est majorée par $D^k \dfrac{|t - t_0|^k}{k!}$. ∎

Exercice 12.
Calculer $R(t,t_0)$ lorsque A est constante.
Solution. Puisque $e^{(t-t_0)A}$ est une solution satisfaisant à la même condition initiale, on a $R(t,t_0)=e^{(t-t_0)A}$ dans ce cas.

Théorème 27 (Théorème de Cauchy-Lipschitz linéaire vectoriel).
(i) Soient $t_0\in I$ et $Y_0\in K^n$. L'équation (S_0) admet une unique solution globale F satisfaisant la condition initiale $F(t_0)=Y_0$.
(ii) L'application « conditions initiales » $F\mapsto F(t_0)$ est un isomorphisme de $\mathrm{Sol}_K\big((S_0),I\big)$ sur K^n.
(iii) Le K-espace vectoriel $\mathrm{Sol}_K\big((S_0),I\big)$ est dimension n.

Démonstration. La première assertion découle des théorèmes 25 de la page ci-contre et 21 de la page 823. La seconde en est une paraphrase et la troisième une conséquence immédiate. ∎

Pratiquement, la solution F est donnée par la formule : $F(t):=R(t,t_0)Y_0$, où $R(t,t_0)$ est la résolvante.

Corollaire 28. Pour tout intervalle non trivial $J\subset I$, l'application de restriction $F\mapsto F_{|J}$ est un isomorphisme de $\mathrm{Sol}_K\big((S_0),I\big)$ sur $\mathrm{Sol}_K\big((S_0),J\big)$. Toute solution sur un intervalle non trivial J se prolonge en une unique solution globale.

La résolvante est la série des $\Phi^k(I_n)$, qui peuvent être calculés itérativement comme suit : $\mathcal{F}_0(t):=I_n$ (fonction constante) et, pour $k\geqslant 0$, $\mathcal{F}_{k+1}(t)=\displaystyle\int_{t_0}^{t}A(s)\mathcal{F}_k(s)\,\mathrm{d}s$.

Exercice 13.
Exprimer $\mathcal{F}_k(t)$ comme une intégrale multiple (on supposera $t\geqslant t_0$). En déduire une majoration de sa norme. Appliquer au cas particulier où A est constante.
Solution. On trouve sans peine l'*intégrale itérée* :

$$\mathcal{F}_k(t)=\int_{t_0\leqslant s_1\leqslant\cdots\leqslant s_k\leqslant t}A(s_k)\cdots A(s_1)\,\mathrm{d}s_1\cdots\mathrm{d}s_k.$$

Ainsi, si $\|A\|\leqslant D$, la norme de $\mathcal{F}_k(t)$ est majorée par D^k multiplié par le volume du domaine de $\mathbb{R}^k$ défini par les inégalités $t_0\leqslant s_1\leqslant\cdots\leqslant s_k\leqslant t$. C'est un exercice amusant, donc laissé au lecteur, que de démontrer que ce volume est $\dfrac{(t-t_0)^k}{k!}$, soit par la géométrie euclidienne soit à l'aide du chapitre sur les intégrales multiples de [L2-L3]).
Si A est constante, l'intégrale multiple est A^k multiplié par le volume du domaine indiqué, d'où la formule $\mathcal{F}_k(t)=\dfrac{(t-t_0)^k}{k!}A^k$.
On reconnaît le terme général de la série exponentielle ; on obtient donc dans ce cas :

$$R(t,t_0)=e^{(t-t_0)A}.$$

2.1.2 Le théorème de Cauchy-Lipschitz scalaire

Nous noterons encore (E_b) et (E_0) les équations différentielles (6) et (8) de la page 804, dont nous conservons les notations. Nous notons (S_B) et (S_0) les systèmes associés à ces équations à la section 1.3.1.

Théorème 29 (Théorème de Cauchy-Lipschitz linéaire scalaire).
(i) Soient $t_0 \in I$ et $(y_0, \dots, y_{n-1}) \in K^n$. L'équation (E_b) admet une unique solution globale f soumise aux conditions initiales $f(t_0) = y_0, \dots, f^{(n-1)}(t_{n-1}) = y_{n-1}$. Le K-espace vectoriel $\mathrm{Sol}\big((E_0), I\big)$ est de dimension n. L'application « conditions initiales » $f \mapsto \big(f(t_0), \dots, f^{(n-1)}(t_{n-1})\big)$ est un isomorphisme de $\mathrm{Sol}\big((E_0), I\big)$ sur K^n.
(ii) L'équation (E_0) admet un système fondamental de solutions sur I.

Démonstration. (i) Il découle du théorème 27 de la page précédente que le système (S_B) admet une unique solution globale Y telle que $Y(t_0) = (y_0, \dots, y_{n-1})$. Il suffit alors d'appliquer la proposition 18 de la page 821.
(ii) Soit $(f_1, \dots, f_n)$ la famille d'éléments de $\mathrm{Sol}\big((E_0), I\big)$ dont l'image par l'isomorphisme ci-dessus est la base canonique de K^n. Sa matrice wronskienne en t_0 est I_n et son wronskien est donc non identiquement nul. ∎

Corollaire 30. (i) Pour toute famille $(f_1, \dots, f_n)$ de solutions de (E_0) sur tout intervalle non trivial $J \subset I$ et pour tout $t_0 \in J$, les conditions suivantes sont équivalentes : le wronskien de $(f_1, \dots, f_n)$ est non nul en t_0 ; la famille $(f_1, \dots, f_n)$ est un système fondamental de solutions sur J ; la famille $(f_1, \dots, f_n)$ est une base de $\mathrm{Sol}\big((E_0), J\big)$.
(ii) Toute solution sur un intervalle non trivial $J \subset I$ admet un unique prolongement en une solution globale.

Par exemple, la base déterminée dans le théorème 9 de la page 812 est un système fondamental de solutions.

Exercice 14.
Démontrer que la fonction définie sur $\mathbb{R}^*$ par $f(t) = e^{-1/t^2}$ et prolongée par $f(0) = 0$ est de classe $\mathcal{C}^\infty$ mais n'est solution d'aucune équation linéaire différentielle scalaire homogène.

Solution. Il est clair que f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^*$. Ses dérivées sont de la forme $f^{(n)}(t) = e^{-1/t^2} P_n(1/t)$, où les P_n sont définis par :

$$P_0(x) = 1 \qquad \text{et} \qquad P_{n+1}(x) = 2x^3 P_n(x) - x^2 P_n'(x).$$

Ce sont donc des polynômes et l'on a, pour tout n, l'égalité :

$$\lim_{t \to 0} f^{(n)}(t) = 0.$$

On en déduit que le prolongement par $f(0) = 0$ fait de f une fonction de classe $\mathcal{C}^\infty$ telle que :

$$\forall n \in \mathbb{N}\ ,\ f^{(n)}(0) = 0.$$

Si f était solution d'une équation linéaire différentielle scalaire homogène (E_0) d'ordre n, les égalités $f(0) = \cdots = f^{(n-1)}(0) = 0$, jointes au théorème 29, entraîneraient que f est nulle.

Exercice 15.
Soient p et q deux fonctions continues sur I et f une solution non triviale de l'équation $x'' + pf' + qf = 0$. Montrer que, dans tout segment $[a,b] \subset I$, la fonction f n'a qu'un nombre fini de zéros. (Si $(t_n)_{n\in\mathbb{N}}$ est une suite strictement croissante de zéros de f dans $[a,b]$, de limite α, on a $f(\alpha) = 0$; en appliquant le théorème de Rolle sur chaque $]t_k, t_{k+1}[$, prouver que $f(\alpha) = f'(\alpha) = 0$ et déduire du théorème 29 de la page ci-contre que $f = 0$.) Que peut-on en déduire sur l'ensemble ordonné des zéros de f ?
Solution. Si f a une infinité de zéros dans $[a,b]$, ceux-ci ont un point d'accumulation $\alpha \in [a,b]$. Il y a donc une suite de zéros de f dans $[a,b]$ de limite α et l'on peut en extraire une suite $(t_n)_{n\in\mathbb{N}}$ strictement croissante ou bien strictement décroissante (voir par exemple l'exercice type corrigé du module « Fondements » de [L1]). Pour le raisonnement qui suit, on peut aussi bien supposer que cette suite est strictement croissante.
Par continuité, on a $f(\alpha) = 0$. Comme $f(t_k) = f(t_{k+1}) = 0$, on déduit du théorème de Rolle l'existence de $u_k \in]t_k, t_{k+1}[$ tel que $f'(u_k) = 0$. La suite $(u_n)_{n\in\mathbb{N}}$ converge vers α et, f étant de classe $\mathcal{C}^1$, f' est continue et $f'(\alpha) = 0$. En appliquant l'assertion d'unicité du théorème 29 de la page précédente au temps initial $t := \alpha$, on trouve que $f = 0$, contradiction.
On peut en déduire que, dans l'ensemble des zéros de f, tout élément qui n'est pas le maximum admet un successeur ; et tout élément qui n'est pas le minimum admet un prédécesseur.

2.1.3 La résolvante

Proposition 31 (Relations de Chasles pour la résolvante).
Notons $R(t,t_0)$ la résolvante du système (S_0). Alors, quels que soient $t_0, t_1, t_2 \in I$, on a :

$$R(t_0,t_0) = I_n,$$
$$R(t_2,t_1)R(t_1,t_0) = R(t_2,t_0),$$
$$R(t_1,t_0) = R(t_0,t_1)^{-1}.$$

Si V est une solution matricielle fondamentale arbitraire sur I, on a :

$$R(t,s) = V(t)V(s)^{-1}.$$

Démonstration. La première relation découle tautologiquement de la définition de la résolvante (théorème 25 de la page 828).
Les fonctions matricielles $R(t,t_0)$ et $R(t,t_1)R(t_1,t_0)$ sont toutes deux solutions de (S_0) et prennent la même valeur en $t := t_1$; elles sont donc égales, d'après l'assertion d'unicité dans le théorème de Cauchy-Lipschitz.
La troisième relation est conséquence immédiate des deux premières.
Enfin, la fonction matricielle $V(t)V(t_0)^{-1}$ est solution de (S_0) et vaut I_n en $t := t_0$, c'est donc $R(t,t_0)$.

Corollaire 32. La fonction de deux variables $R(t,s)$ est de classe $\mathcal{C}^1$ (au sens du module II.5).

La proposition suivante est à rapprocher de la proposition 29 de la page 287 et des commentaires qui la suivent.

Proposition 33. (i) On a la formule :

$$\det R(t,t_0) = \exp\left(\int_{t_0}^{t} \operatorname{Tr} A(s)\,\mathrm{d}s\right).$$

Si $\operatorname{Tr} A$ est identiquement nul, alors la résolvante est à valeurs dans $SL_n(K)$.
(ii) On suppose que A est à valeurs dans l'espace des matrices antisymétriques réelles. Alors la résolvante est à valeurs dans $SO_n(\mathbb{R})$.

Démonstration. (i) D'après le théorème 19 de la page 822, le membre gauche de l'égalité vérifie l'équation $W' = \operatorname{Tr}(A)W$, avec la condition initiale $W(t_0) = 1$; la formule en découle, et l'assertion qui suit est alors immédiate.
(ii) Notons $\mathcal{F}(t) := {}^tR(t,t_0)R(t,t_0)$. Alors :

$$\mathcal{F}'(t) := {}^t\big(A(t)R(t,t_0)\big)R(t,t_0) + {}^tR(t,t_0)A(t)R(t,t_0) = 0,$$

et la fonction $\mathcal{F}$ est constante. Comme $\mathcal{F}(t_0) = I_n$, la résolvante est à valeurs dans $O_n(\mathbb{R})$, donc dans $SO_n(\mathbb{R})$ d'après (i). ∎

La première formule, ainsi que le calcul de la résolvante lorsque A est constante, pourraient laisser espérer une formule générale : $R(t,t_0) = \exp\left(\int_{t_0}^{t} A(s)\,\mathrm{d}s\right)$. L'exercice II.7.12 de la page 876 montre qu'il n'en est rien en général. L'exercice II.7.13 montre que c'est lié à la non commutativité de la multiplication matricielle. L'exercice II.7.15 discute des cas où la formule peut être appliquée.

On a déjà vu comment s'expriment les solutions de (S_0) en fonction de la résolvante et des conditions initiales. La méthode de variation des constantes fournit, pour (S_B), la solution : $R(t,t_0)\displaystyle\int_{t_0}^{t} R(s,t_0)^{-1}B(s)\,\mathrm{d}s$. Tenant compte des relations de Chasles et de la forme générale des solutions de (S_0), on trouve que *l'unique solution de* (S_B) *satisfaisant à la condition initiale* $F(t_0) = Y_0$ *est donnée par la formule :*

$$F(t) := R(t,t_0)Y_0 + \int_{t_0}^{t} R(t,s)B(s)\,\mathrm{d}s.$$

Si l'on part d'une solution matricielle fondamentale $V(s)$ arbitraire, on tire de la proposi-

tion 31 la formule :

$$F(t) := V(t)V(t_0)^{-1}Y_0 + \int_{t_0}^{t} V(t)V(s)^{-1}B(s)\,\mathrm{d}s.$$

Proposition 34. Soit D un majorant de $\|A(t)\|$ sur un intervalle ouvert $J \subset I$ qui contient t_0. Alors la solution de (S_B) satisfaisant à la condition initiale $F(t_0) = Y_0$ vérifie l'inégalité :

$$\forall t \in J\ ,\ \|F(t) - Y_0\| \leqslant \left(e^{D|t-t_0|} - 1\right)\|Y_0\|\,. \tag{26}$$

Démonstration. C'est une conséquence immédiate du corollaire 26 de la page 828. ∎

2.1.4 Le théorème fondamental sur les courbes paramétrées

Il a été énoncé dans le module « Courbes paramétrées » de [L1] (sections 3.4 et 4.3). Il dit qu'une courbe plane (resp. de l'espace) est déterminée, à un déplacement près, par sa courbure (resp. sa courbure et sa torsion). Une interprétation en a été donnée dans la section 2.3 du module I.6. Nous allons l'illustrer en termes de systèmes différentiels linéaires.

Commençons par le cas d'une courbe plane. La courbure κ étant donnée en fonction de l'abscisse curviligne s, et notant $\mathcal{T}$ la matrice dont les lignes sont les coordonnées des vecteurs $\overrightarrow{T}$ et $\overrightarrow{N}$, on a :

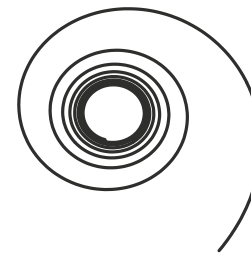
$$\mathcal{T}' = A\mathcal{T}\ ,\ \text{où}\ A := \begin{pmatrix} 0 & \kappa \\ -\kappa & 0 \end{pmatrix}.$$

On peut choisir un repère orthonormé de telle sorte que $\mathcal{T}(0) = I_2$, et l'on voit que l'on aura $\mathcal{T}(s) \in SO_2(\mathbb{R})$ pour tout s, puisque A est antisymétrique (proposition 33 de la page précédente).

Pour résoudre numériquement le système, à l'aide de Maple, le plus simple est d'introduire l'angle $\alpha(s)$ du vecteur tangent au dessus de l'horizontale, de sorte que l'on a : $\overrightarrow{T} = (\cos\alpha, \sin\alpha)$ et $\kappa(s) = \alpha'(s)$. On est alors conduit à résoudre le système non linéaire :

$$\begin{cases} x'(s) = \cos\alpha(s), \\ y'(s) = \sin\alpha(s), \\ \alpha'(s) = \kappa(s). \end{cases}$$

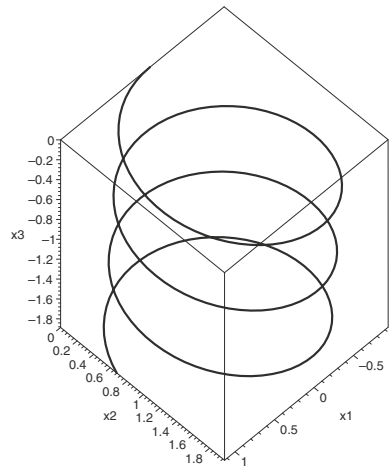
Les conditions initiales sur x, y déterminent la courbe à une translation près, celle sur α à une rotation près. Un exemple célèbre est la « courbe de l'autoroute » ou *spirale de Cornu* ou *clothoïde*. Elle résout le problème suivant : comment construire une bretelle d'autoroute de manière à ce que le conducteur qui roule à vitesse constante tourne son volant à vitesse constante ? La courbure doit donc être proportionnelle au trajet parcouru. Nous prendrons $\kappa(s) = s$. Avec les conditions initiales $x(0) = y(0) = \alpha(0) = 0$, on trouve la courbe suivante :



Dans le cas d'une courbe de l'espace, nous prendrons pour $\mathcal{T}$ la matrice de lignes $\overrightarrow{T}$, $\overrightarrow{N}$ et $\overrightarrow{B}$ et pour matrice A la matrice antisymétrique $\begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & -\tau \\ 0 & \tau & 0 \end{pmatrix}$. Nous résoudrons le système associé dans le cas d'une courbure κ et d'une torsion τ constantes. Les instructions Maple ci-dessous montrent comment poser le problème :

```
restart: with(DEtools,DEplot3d): kappa(s) := 1 : tau(s) := 0.1:
Eq := {diff(x1(s),s) = t1(s),
diff(x2(s),s) = t2(s), diff(x3(s),s) = t3(s),
diff(t1(s),s) = k(s)*n1(s),
diff(t2(s),s) = k(s)*n2(s),
diff(t3(s),s) = k(s)*n3(s),
diff(n1(s),s) = -k(s)*t1(s) - tau(s)*b1(s) ,
diff(n2(s),s) = -k(s)*t2(s) - tau(s)*b2(s) ,
diff(n3(s),s) = -k(s)*t3(s) - tau(s)*b3(s) ,
diff(b1(s),s) = tau(s)*n1(s),
diff(b2(s),s) = tau(s)*n2(s),
diff(b3(s),s) = tau(s)*n3(s)}
DEplot3d(Eq, [x1,x2,x3,t1,t2,t3,n1,n2,n3,b1,b2,b3], s=0..20,
{[0,0,0,0,1,0,0,0,1,0,0,0,1]}, stepsize=0.1, scene=[x1,x2,x3],
linecolor=black);
```

La courbe obtenue est une hélice (module « Courbes paramétrées » de [L1]) :



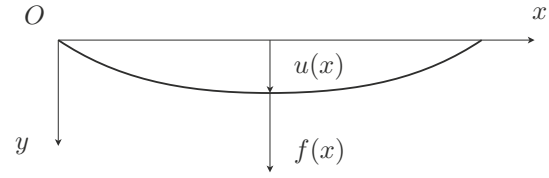
2.1.5 Problèmes de conditions aux limites

Puisqu'un système de rang n admet un espace vectoriel de solutions de dimension n, il est naturel de penser qu'en imposant n conditions à une solution, on détermine uniquement celle-ci (existence comprise). C'est bien ce que dit le théorème de Cauchy-Lipschitz dans le cas de *conditions initiales*. Il existe cependant d'autres types de conditions, qui apparaissent naturellement, pour lesquelles cela ne marche pas si bien.

Exemple. Une poutre métallique soumise à son propre poids fléchit. Considérons la poutre au repos (sans action de la pesanteur) comme le segment d'extrémités $(0,0)$ et $(1,0)$ du plan vertical rapporté à l'axe horizontal Ox et à l'axe vertical Oy. Ce dernier sera orienté *vers le bas*. Supposons que l'élément de longueur dx de position $x \in [0,1]$ soit soumis à son poids $f(x)dx$ (où $f > 0$). La poutre déformée sera décrite par sa *flèche* $u(x)$, où u est solution de l'équation différentielle avec *conditions aux limites* :

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), \\ u(0) = u(1) = 0. \end{cases}$$

Les conditions aux limites expriment le fait que les extrémités de la poutre sont fixées. La fonction $c(x)$ est liée au matériau. Elle est strictement positive.
Nous allons étudier l'équation ci-dessus sous diverses hypothèses.



Supposons d'abord c constante. Si $c = 0$, il s'agit d'un problème d'intégrations successives ; comme u sera déterminé à une fonction $\alpha x + \beta$ près, il y aura bien une solution unique. Si $c > 0$ (resp. $c < 0$), nous reconnaissons une équation du type rencontré (avec des notations légèrement différentes) dans l'exemple de la page 810 (resp. dans l'exercice 5 de la page 810).

Exercice 16.

On suppose $c := -\pi^2$. Le problème admet-il une solution ?

Solution. Une base de solutions du système homogène associé est $\cos \pi x$, $\sin \pi x$. Une solution particulière donnée par les formules de la section 1.2.3 est :

$$u_0(x) := \frac{1}{2\mathrm{i}\pi} \int_0^x e^{\mathrm{i}\pi(x-s)} f(s)\,\mathrm{d}s - \frac{1}{2\mathrm{i}\pi} \int_0^x e^{-\mathrm{i}\pi(x-s)} f(s)\,\mathrm{d}s = \frac{1}{\pi} \int_0^x \sin \pi(x-s)\, f(s)\,\mathrm{d}s.$$

La solution générale est donc $u(x) = \lambda \cos \pi x + \mu \sin \pi x + \dfrac{1}{\pi} \displaystyle\int_0^x \sin \pi(x-s) f(s)\,\mathrm{d}s$.

La condition $u(0) = 0$ donne $\lambda := 0$. La condition $u(1) = 0$ donne alors :

$$\int_0^1 \sin \pi s\, f(s)\,\mathrm{d}s = 0.$$

Le système avec conditions aux limites a une solution si, et seulement si, cette égalité est vraie. De plus, si elle est vraie, il y a une infinité de solutions puisque l'on n'a aucune condition sur μ.
Le problème est que les fonctions $\cos \pi x$ et $\sin \pi x$ prennent les mêmes valeurs en 0 et 1 et que les conditions aux limites ne se traduisent pas par un système de Cramer en λ et μ.

Le système ci-dessus sera résolu numériquement dans l'exercice II.8.0 de la page 949 du module II.8, avec c constante positive. Nous allons ici simplement supposer que c est une fonction telle que $c \geqslant -\pi^2$ et qui n'est pas la constante $-\pi^2$. Soit (u_1, u_2) une base de solutions du système homogène associé. Considérons la fonction $v(x) := u_1(0)u_2(x) - u_2(0)u_1(x)$. Elle vérifie l'équation homogène $v'' = cv$, et elle n'est pas identiquement nulle : en effet, $v'(0)$ est la valeur du wronskien de u_1, u_2 en 0, donc n'est pas nulle. Comme $v(0) = 0$, en appliquant l'exercice II.7.24 de la page 878, on voit qu'elle ne s'annule pas sur $]0,1]$. En particulier, $v(1) \neq 0$.

Exercice 17.
Montrer que le problème admet une solution (s'inspirer de l'exercice 16).
Solution. Soit u_0 une solution particulière et cherchons u sous la forme $u = u_0 + \lambda u_1 + \mu u_2$. La condition $u(0) = 0$ donne $\lambda u_1(0) + \mu u_2(0) = -u(0)$. La condition $u(1) = 0$ donne $\lambda u_1(1) + \mu u_2(1) = -u(1)$. Le déterminant de ce système est $v(1) \neq 0$, c'est un système de Cramer, il admet donc une solution.

2.2 Aspects qualitatifs

La première idée pour comprendre le comportement d'un système physique régi par une équation différentielle est de tracer le graphe de ses solutions. Cependant, ce graphe ne rend justement pas très bien compte de la relation différentielle ! En fait, les physiciens qui étudient le mouvement d'un point ont coutume de représenter dans un même espace la position et la vitesse de ce point à tout moment. Autrement dit, pour une équation différentielle scalaire d'ordre 2 (ce qui est le cas de la mécanique newtonienne), ils représentent les solutions avec leurs dérivées.

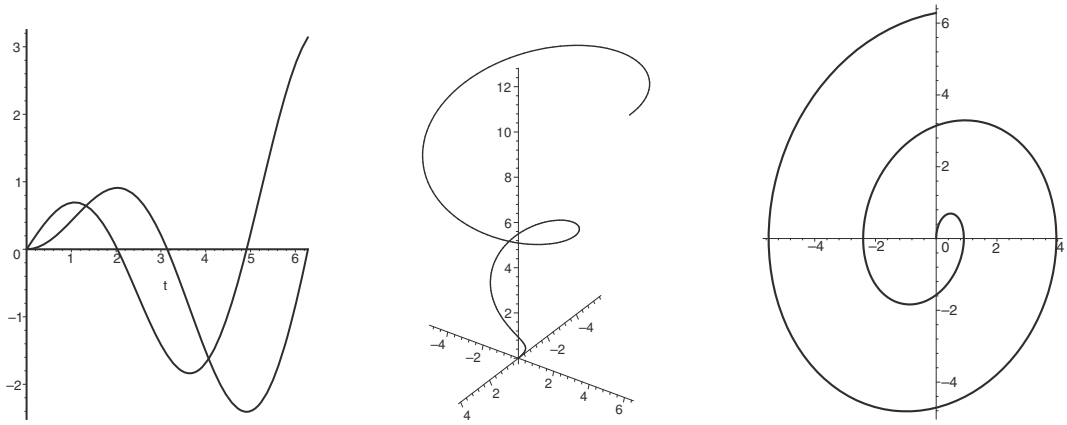
Définition 8. Soit f une solution globale de l'équation différentielle (E_b) de la page 804. La *trajectoire* associée à la solution f est le graphe de la fonction $t \mapsto \big(f(t), \dots, f^{(n-1)}(t)\big)$. De même, la trajectoire d'une solution F du système (S_B) est le graphe de la fonction $t \mapsto F(t)$. C'est donc, dans les deux cas, l'image d'une courbe paramétrée de classe $\mathcal{C}^1$ dans l'*espace des phases élargi* $I \times K^n$.

D'après les théorèmes 29 de la page 830 et 27 de la page 829, par tout point de l'espace des phases élargi, il passe une unique trajectoire. Celles-ci forment donc une *partition* de $I \times K^n$.

Exercice 18.
Quelles sont les trajectoires des solutions de l'équation $\ddot{x} + x = 0$?
Solution. Toute solution globale est de la forme $f(t) := a\cos t + b\sin t$, d'où $f'(t) = b\cos t - a\sin t$. La trajectoire est donc l'image de la courbe paramétrée $t \mapsto (t, a\cos t + b\sin t, b\cos t - a\sin t)$. Si $a := b := 0$, c'est la droite $I \times \{(0,0)\}$. Sinon, c'est une hélice dont l'axe est cette droite.

Exercice 19.
Quelle est la trajectoire de l'équation $\ddot{x} + x = \cos t$ qui passe par $(0,0,0)$?
Solution. Selon l'exercice 6 de la page 814, une solution particulière de l'équation est $\frac{1}{2}t\sin t$. La solution générale est donc de la forme $f(t) := \frac{1}{2}t\sin t + a\cos t + b\sin t$. On a alors $f'(t) = \frac{1}{2}t\cos t + b\cos t + \left(\frac{1}{2} - a\right)\sin t$. La condition initiale $f(0) = 0$ impose $a := 0$. La condition initiale $f'(0) = 0$ impose alors $b := 0$. La trajectoire est donc l'image de la courbe paramétrée $t \mapsto \left(t, \frac{1}{2}t\sin t, \frac{1}{2}t\cos t + \frac{1}{2}\sin t\right)$. On a représenté ci-

dessous les graphes de f et f', la trajectoire et la *courbe intégrale*, c'est-à-dire la projection de la trajectoire dans le plan de phases.



2.2.1 Portrait de phases d'une équation linéaire autonome

Nous verrons à la section 3.3 de la page 865 que l'on appelle *autonome* une équation d'ordre n dans laquelle $y^{(n)}$ ne dépend que de $y, \dots, y^{(n-1)}$ et pas explicitement du temps. Les équations linéaires autonomes sont donc les équations scalaires ou vectorielles à coefficients constants (y compris le second membre b ou B). Nous ne parlerons, dans leur cas, que de solutions globales sur $\mathbb{R}$.

Théorème 35. Soient $Y_0 := (y_0, \dots, y_{n-1}) \in K^n$ des conditions initiales et $t_0, u_0 := t_0 + \tau \in \mathbb{R}$ des temps initiaux. Soient F la solution de l'équation autonome (E_b) satisfaisant à la condition initiale $F(t_0) = Y_0$ et G la solution satisfaisant à la condition initiale $G(u_0) = Y_0$. Les solutions F et G sont alors décalées dans le temps :

$$\forall t \in \mathbb{R}\ ,\ F(t) = G(t + \tau).$$

Soient de même f la solution de l'équation autonome (E_b) satisfaisant aux conditions initiales $f(t_0) = y_0, \dots, f^{(n-1)}(t_0) = y_{n-1}$ et g la solution satisfaisant aux conditions initiales $g(u_0) = y_0, \dots, g^{(n-1)}(u_0) = y_{n-1}$. Les solutions f et g sont alors décalées dans le temps :

$$\forall t \in \mathbb{R}\ ,\ f(t) = g(t + \tau).$$

Démonstration. Notons provisoirement $H(t) := G(t + \tau)$.

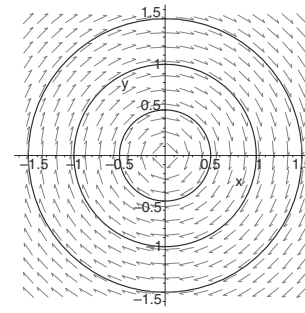
Alors $H' = AH$ et $H(t_0) = G(u_0) = Y_0$. D'après l'assertion d'unicité du théorème de Cauchy-Lipschitz, $H = F$.
Le cas de (E_b) s'en déduit immédiatement par vectorialisation. ∎

Définition 9. Soit f une solution globale de l'équation différentielle linéaire autonome (E_b). On appelle *courbe intégrale* associée à f la projection sur K^n de la trajectoire correspondante, c'est-à-dire l'image de la fonction $t \mapsto \big(f(t), \dots, f^{(n-1)}(t)\big)$. C'est donc l'image d'une courbe paramétrée de classe $\mathcal{C}^1$ dans l'*espace des phases* K^n. De même, la courbe intégrale associée à la solution globale F du système (S_B) est la projection sur K^n de la trajectoire correspondante, c'est-à-dire l'image de la fonction $t \mapsto F(t)$.

Dans [L1], où seules des équations à coefficients constants ont été étudiées, ces courbes ont été appelées *trajectoires*.

Corollaire 36. Par chaque point $Y_0 := (y_0, \dots, y_{n-1})$ de l'espace des phases K^n passe une et une seule courbe intégrale de (E_b) ou de (S_B).

L'ensemble des courbes intégrales d'une équation autonome forme donc une partition de l'espace des phases : cette partition est appelée *portrait de phases* de l'équation. Toutes les figures de la section 3.3 du module IV.6 de [L1] représentent des portraits de phases.



Exemple. Le portrait de phases de l'équation $\ddot{x} + x = 0$ est la partition du plan en cercles centrés à l'origine (y compris le cercle de rayon 0, correspondant à la solution constante).

Équations autonomes et champs de vecteurs

Soit $\big(x(t), y(t)\big)$ une solution du système autonome $\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$. En tout point de paramètre t de la courbe intégrale correspondante, la tangente est dirigée par le vecteur $V(x,y) := \begin{pmatrix} ax + by + e \\ cx + dy + f \end{pmatrix}$. (Pour le cas où $V(x,y) = 0$, voir l'exercice ci-dessous.) On appelle *champ de vecteurs plan* l'application $(x,y) \mapsto V(x,y)$. Il revient donc au même de se donner une équation autonome ou un champ de vecteurs, mais les champs de vecteurs permettent une vision géométrique. Leur utilisation sera plus accentuée dans l'étude des équations différentielles comme *systèmes dynamiques* de [MPA]. Naturellement, on peut définir un champ de vecteurs dans $\mathbb{R}^n$, donc pour n'importe quelle équation autonome, mais nous ne le ferons que pour les équations d'ordre 2. Les dessins de courbes intégrales de l'étude des circuits RLC, page 839, sont réalisés sur fond de champ vecteurs.

Exercice 20.

Supposons que $V(x_0, y_0) = 0$. On dit alors que (x_0, y_0) est un *point critique*, ou encore *singulier* du champ de vecteurs V. Que peut-on dire de la courbe intégrale passant par ce point ?

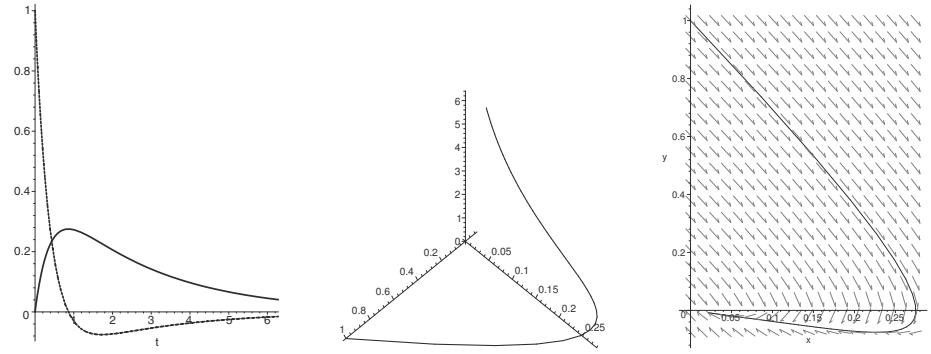
Solution. La fonction constante $\big(x(t), y(t)\big) := (x_0, y_0)$ est alors solution, et c'est la seule d'après le théorème de Cauchy-Lipschitz. La courbe intégrale est alors réduite à un point.

Exemple. L'intensité du courant dans un circuit RLC (capacité $C > 0$, inductance $L > 0$, résistance $R \geqslant 0$) sans stimulation extérieure vérifie l'équation différentielle :

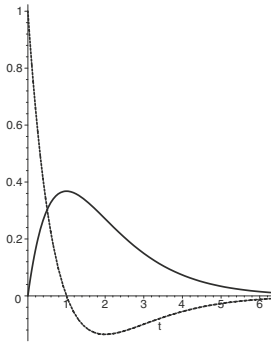
$$L\ddot{I} + R\dot{I} + CI = 0. \tag{27}$$

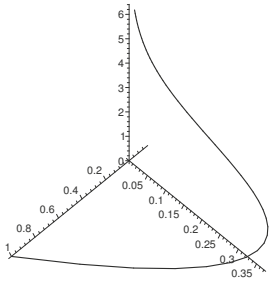
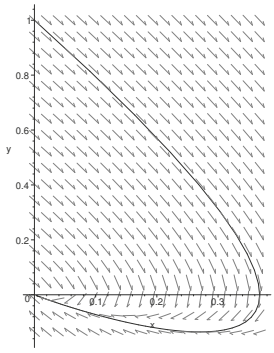
(cas de l'oscillateur amorti). L'équation caractéristique associée (définition 5 de la page 811) est : $LX^2 + RX + C = 0$. On distingue donc plusieurs cas selon le signe du discriminant $R^2 - 4LC$. Dans chaque cas, nous représenterons le graphe de la solution $I(t)$ et de sa dérivée $\dot{I}(t)$, soumises à la condition initiale $I(0) := 1$, $\dot{I}(0) = 0$; la trajectoire correspondante, c'est-à-dire l'image de la courbe paramétrée $t \mapsto \Big(t, I(t), \dot{I}(t)\Big)$; et la courbe intégrale, c'est-à-dire l'image de la courbe paramétrée $t \mapsto \Big(I(t), \dot{I}(t)\Big)$. Les illustrations ci-dessous ont été réalisées à l'aide de Maple, en prenant $L := C := 1$ et $R := 3, 2, 1, 0, -1$ et les conditions initiales $\big(I(0), \dot{I}(0)\big) := (0, 1)$.

- Si $R > 2\sqrt{LC}$, il y a deux racines réelles λ_1 et λ_2, toutes deux strictement négatives puisque $C > 0$ et (ici) $R > 0$. Les solutions sont donc de la forme $I(t) := a_1 e^{\lambda_1 t} + a_2 e^{\lambda_2 t}$. Les trajectoires sont les images des courbes paramétrées $t \mapsto (t, a_1 e^{\lambda_1 t} + a_2 e^{\lambda_2 t}, \lambda_1 a_1 e^{\lambda_1 t} + \lambda_2 a_2 e^{\lambda_2 t})$. Les courbes intégrales sont les images des courbes paramétrées $t \mapsto (a_1 e^{\lambda_1 t} + a_2 e^{\lambda_2 t}, \lambda_1 a_1 e^{\lambda_1 t} + \lambda_2 a_2 e^{\lambda_2 t})$.



- Si $R = 2\sqrt{LC}$, on a une racine double $\lambda < 0$ et les solutions ont la forme $I(t) := (a_1 + a_2 t)e^{\lambda t}$. Les trajectoires sont les images des courbes paramétrées $t \mapsto \big(t, (a_1 + a_2 t)e^{\lambda t}, (a_2 + \lambda a_1 + \lambda a_2 t)e^{\lambda t}\big)$. Les courbes intégrales sont les images des courbes paramétrées $t \mapsto \big((a_1 + a_2 t)e^{\lambda t}, (a_2 + \lambda a_1 + \lambda a_2 t)e^{\lambda t}\big)$.

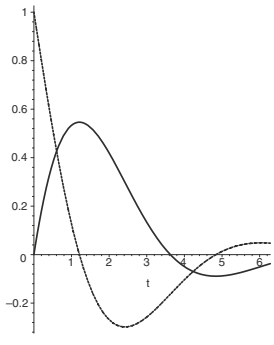


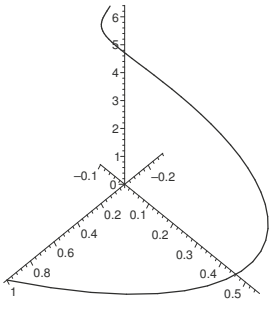
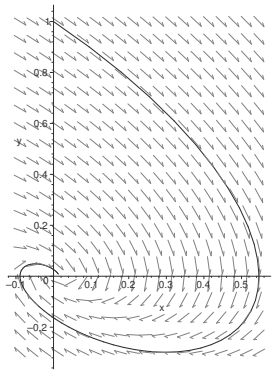



Dans les deux cas que nous venons d'examiner, c'est-à-dire si $R \geqslant 2\sqrt{LC}$, toutes les solutions tendent rapidement vers 0 lorsque t tend vers $+\infty$ et elles n'ont aucun aspect oscillatoire : c'est l'amortissement dû à la résistance qui prime. Au contraire, dans les deux cas qui suivent, c'est-à-dire si $0 \leqslant R < 2\sqrt{LC}$, la présence de racines complexes se manifeste par un comportement oscillatoire.

- Lorsque $0 < R < 2\sqrt{LC}$, il y a deux racines complexes conjuguées $\lambda \pm \mathrm{i}\mu$ de partie réelle $\lambda := -\dfrac{R}{2L}$ strictement négative. Les solutions ont donc la forme $I(t) := e^{\lambda t}(a\cos\omega t + b\sin\omega t)$, avec $\omega := \sqrt{\dfrac{C}{L} - \dfrac{R}{2L}^2}$. Il y a un « régime pério-dique amorti » de période $\dfrac{2\pi}{\omega}$. Les trajectoires sont les images des courbes paramétrées $t \mapsto \big(t, a\cos e^{\lambda t}\omega t + be^{\lambda t}\sin\omega t, (b+\lambda a)e^{\lambda t}\cos\omega t + (-a+\lambda b)e^{\lambda t}\sin\omega t\big)$. Les courbes intégrales sont les images des courbes paramétrées :

$$t \mapsto \big(a\cos e^{\lambda t}\omega t + be^{\lambda t}\sin\omega t, (b + \lambda a)e^{\lambda t}\cos\omega t + (-a + \lambda b)e^{\lambda t}\sin\omega t\big).$$



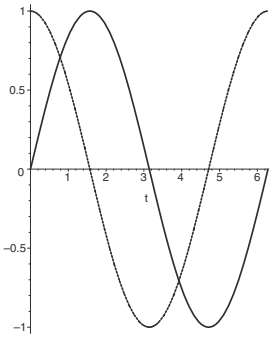



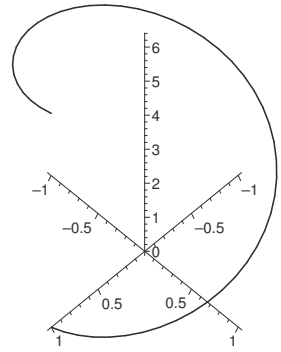
- Lorsque $R := 0$, on retombe sur le cas de l'oscillateur harmonique, les solutions sont toutes périodiques de pulsation $\omega := \sqrt{\dfrac{C}{L}}$ et de période $\dfrac{2\pi}{\omega}$.
  Les trajectoires sont les images des courbes paramétrées

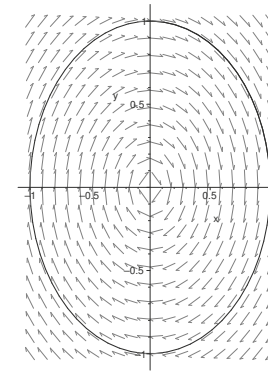
$$t \mapsto \big(t, a\cos\omega t + b\sin\omega t, (b+\lambda a)\cos\omega t + (-a+\lambda b)\sin\omega t\big).$$

  Les courbes intégrales sont les images des courbes paramétrées :

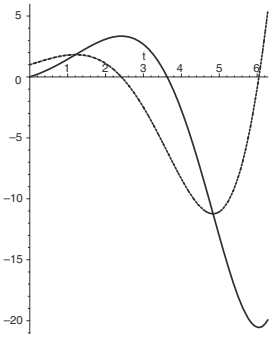
$$t \mapsto \big(a\cos\omega t + b\sin\omega t, (b+\lambda a)\cos\omega t + (-a+\lambda b)\sin\omega t\big).$$

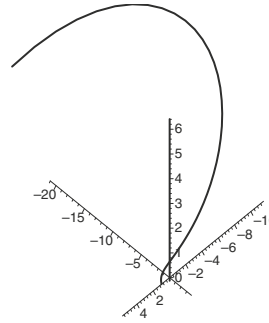


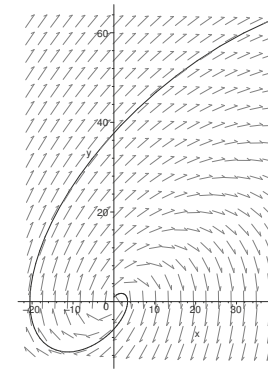




- Pour le cas $R < 0$, voir l'exercice II.7.20 de la page 877. Nous l'illustrons ici :







Pour le cas d'un circuit avec stimulation extérieure, voir l'exercice II.7.19 de la page 877.

Remarque. Lorsque $R > 0$, toutes les courbes intégrales ont 0 dans leur adhérence. Ce fait a une influence sur l'existence d'intégrales premières (page 843).

2.2.2 Intégrales premières d'une équation linéaire autonome

Dans l'étude de l'oscillateur harmonique au début de ce module, nous avons pu voir tout le parti que l'on peut tirer de la loi de conservation de l'énergie. Cette idée admet une généralisation.

Définition 10. On appelle *intégrale première* de l'équation différentielle linéaire autonome (S_B) ou (E_b) une fonction sur l'espace des phases qui est constante sur chaque courbe intégrale.

Exemple. Un circuit RLC sans résistance admet pour intégrale première la fonction $E(y_0, y_1) := Cy_0^2 + Ly_1^2$. En effet, de l'équation $L\ddot{I} + CI = 0$, on tire sans peine que $CI^2 + L\dot{I}^2$ est constant.

Théorème 37. Soit $E(y_0, \dots, y_{n-1})$ une fonction de classe $\mathcal{C}^1$ sur l'espace des phases. Pour que E soit une intégrale première de l'équation autonome (E_b), il faut, et il suffit, que l'on ait :

$$\sum_{i=0}^{n-2} y_{i+1}\frac{\partial E}{\partial y_i} = (a_1 y_{n-1} + \cdots + a_n y_0 - b)\frac{\partial E}{\partial y_{n-1}}\cdot \tag{28}$$

Démonstration. Soit f une solution globale de (E_b). dire que E est constante sur la courbe intégrale correspondante revient à dire que la fonction $E(f, \dots, f^{(n-1)})$ est constante, c'est-à-dire que sa dérivée est nulle. Mais cette dérivée vaut :

$$\sum_{i=0}^{n-1} \frac{\partial E}{\partial y_i}(f, \dots, f^{(n-1)})f^{(i+1)} = f^{(n)}\frac{\partial E}{\partial y_{n-1}}(f, \dots, f^{(n-1)}) + F(f, \dots, f^{(n-1)}),$$

où l'on a noté :

$$F(y_0, \dots, y_{n-1}) := \sum_{i=0}^{n-2} \frac{\partial E}{\partial y_i} y_{i+1}.$$

Supposons vérifiée la relation (28). Alors, en remplaçant $f^{(n)}$ par $-a_1 f^{(n-1)} - \cdots - a_n f + b$, on obtient bien l'égalité voulue, et E est une intégrale première. Supposons réciproquement que E est une intégrale première, donc que, pour toute solution f, l'égalité ci-dessus est vérifiée. En remplaçant $f^{(n)}$ par $-a_1 f^{(n-1)} - \cdots - a_n f + b$, puis en prenant une valeur particulière de t, on voit que la relation (28) est vérifiée en tout point de la courbe intégrale correspondant à f. Comme il passe une courbe intégrale en chaque point, la relation est identiquement vérifiée sur l'espace des phases.

Exercice 21.

Dans le cas d'une équation autonome d'ordre 2, comment interpréter la condition (28) en termes de champ de vecteurs ?

Solution. Notons $\ddot{x} + a\dot{x} + bx = c$ l'équation ; le système autonome et le champ de vecteurs associés sont donc :

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} = \begin{pmatrix}0 & 1\\ -b & -a\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix} + \begin{pmatrix}0\\ c\end{pmatrix} \quad \text{et} \quad V(x, y) = (y, -bx - ay + c).$$

Notons $E(x,y)$ l'intégrale première. La condition s'écrit ici $y\dfrac{\partial E}{\partial x} = (bx+ay-c)\dfrac{\partial E}{\partial y}$. (Nous n'avons fait que traduire (28) dans nos notations.) Elle équivaut donc à dire que le gradient $\left(\dfrac{\partial E}{\partial x}, \dfrac{\partial E}{\partial y}\right)$ est orthogonal à $V(x,y)$. C'est en harmonie avec le fait que le gradient d'une fonction $E(x,y)$ est orthogonal aux tangentes des lignes de niveau $E(x,y)=c$.

Exercice 22.

Trouver toutes les intégrales premières de classe $\mathcal{C}^1$ de l'oscillateur harmonique.

Solution. Ce sont des fonctions telles que :

$$y_1\frac{\partial E}{\partial y_0} = y_0\frac{\partial E}{\partial y_1}.$$

Le plus simple, pour résoudre cette *équation aux dérivées partielles*, est en fait d'écrire que E est constante sur les courbes intégrales : $E(r\cos t, r\sin t)$ ne dépend pas de t. On écrit donc $E(r\cos t, r\sin t) = \varphi(r)$, où r est *a priori* arbitraire. Prenant $t := 0$, on voit que $\varphi(r) = E(r,0)$ est une fonction de classe $\mathcal{C}^1$. Finalement, les intégrales premières de classe $\mathcal{C}^1$ sont les fonctions $E(x,y) := \varphi\left(\sqrt{x^2+y^2}\right)$, où φ est de classe $\mathcal{C}^1$. L'énergie en fait bien partie.

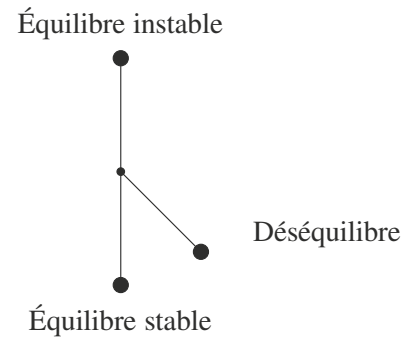
Notons que la constance de E sur les courbes intégrale est bien conséquence de l'équation aux dérivées partielles ci-dessus : la dérivée par rapport à t de $E(r\cos t, r\sin t)$ est $y_0\dfrac{\partial E}{\partial y_1} - y_1\dfrac{\partial E}{\partial y_0}$ évalué en $(r\cos t, r\sin t)$.

Portraits de phase et existence d'intégrales premières continues non triviales. On a vu que, dans le cas d'un circuit RLC amorti ($R>0$), toutes les courbes intégrales avaient l'origine dans leur adhérence. Soit E une intégrale première continue sur le plan des phases. Alors sa valeur sur toute courbe intégrale est égale à $E(0)$. Ainsi, dans ce cas, *toutes les intégrales premières continues sont constantes.*

2.2.3 Stabilité

Une position d'équilibre d'un système physique est dite stable si, lorsqu'on la perturbe un peu, le système ne s'en éloigne pas exagérément : la position basse d'un pendule est stable, mais pas sa position haute.

On peut adapter cette notion en considérant une solution d'une équation différentielle comme une position d'équilibre du système correspondant. Nous supposerons maintenant que l'intervalle ouvert I est de la forme $]a,+\infty[$, le cas de $-\infty$ se traitant de la même manière.



Définition 11.
On dit qu'une solution F de l'équation différentielle $(S_B) : Y' = AY + B$ sur l'intervalle ouvert I satisfaisant à la condition initiale $F(t_0) = Y_0$ est *stable (au sens de Liapounov)* si, quel que soit $\varepsilon > 0$, il existe $\delta > 0$ tel que, pour toute solution G satisfaisant à la condition initiale $G(t_0) = Z_0$, on ait :

$$\|Y_0 - Z_0\| \leqslant \varepsilon \Longrightarrow \|F - G\| \leqslant \delta \text{ sur } I. \tag{29}$$

On dit qu'elle est est *asymptotiquement stable* si elle est stable et si, de plus, il existe une constante $C > 0$ telle que :

$$\|Y_0 - Z_0\| < C \Longrightarrow \lim_{t\to+\infty} \|F(t) - G(t)\| = 0. \tag{30}$$

Nous dirons que la solution f de l'équation (E_b) est *stable*, resp. *asymptotiquement stable*, si la solution correspondante F du système (S_B) l'est.
Naturellement, ces conditions ne dépendent pas du choix de la norme $\|-\|$ sur K^n.

Cette définition a été donnée sous une forme générale s'appliquant aussi bien aux équations différentielles non linéaires. Dans le cas des équations linéaires (le nôtre), nous allons voir qu'elle se simplifie notablement.
Si la solution F de (S_B) satisfaisant à la condition initiale $F(t_0) = Y_0$ est fixée, il revient au même de se donner une solution G de (S_B) satisfaisant à la condition initiale $G(t_0) = Z_0$, ou une solution $H := F - G$ de (S_0) satisfaisant à la condition initiale $H(t_0) = Y_0 - Z_0$. On en déduit que, *pour qu'il existe une solution stable (resp. asymptotiquement stable) de* (S_B)*, il faut, et il suffit, que* 0 *soit une solution stable (resp. asymptotiquement stable) de* (S_0). Nous dirons alors plus brièvement que le système (S_0) est stable (resp. asymptotiquement stable). De plus, comme les solutions de (S_0) dépendent linéairement de leurs conditions initiales, on peut toujours, à un facteur près, leur appliquer les conditions de la définition. Il en résulte les critères suivants :

Proposition 38. (i) Pour que le système linéaire homogène (S_0) soit stable (resp. asymptotiquement stable), il faut, et il suffit, que toutes ses solutions soient bornées (resp. qu'elles tendent vers 0) en $+\infty$.
(ii) Soit $\mathcal{F}$ une solution matricielle fondamentale arbitraire. Pour que le système (S_0) soit stable (resp. asymptotiquement stable), il faut, et il suffit, que $\mathcal{F}$ soit bornée (resp. qu'elle tende vers 0) en $+\infty$.
(iii) Soit $t_0 \in I$. Pour que le système (S_0) soit stable (resp. asymptotiquement stable), il faut, et il suffit, que la résolvante $R(t, t_0)$ soit bornée (resp. qu'elle tende vers 0) en $+\infty$.

Exemples. L'équation $\ddot{x} + x = 0$ de l'oscillateur harmonique est stable mais pas asymptotiquement stable. L'équation $L\ddot{I} + R\dot{I} + CI = 0$ d'un circuit RLC est stable si, et seulement si, $R \geqslant 0$ et asymptotiquement stable si, et seulement si, $R > 0$.

Plus généralement, le cas d'un système (S_0) à coefficients constants : $A \in M_n(K)$, il s'agit d'étudier le comportement de e^{tA} lorsque $t \to +\infty$. Comme indiqué dans le module I.6, il

convient de mettre A sous la forme (1), page 279 :

$$A = P\ \mathrm{Diag}(\lambda_1 I_{r_1} + T_1, \dots, \lambda_k I_{r_k} + T_k)\, P^{-1}$$
$$\Longrightarrow P^{-1} e^{tA} P = \mathrm{Diag}\big(e^{\lambda_1 t}\exp(tT_1), \dots, e^{\lambda_k t}\exp(tT_k)\big).$$

Tout revient donc à considérer le comportement des blocs $e^{\lambda_i t}\exp(tT_i)$. Puisque T_i est nilpotente, les coefficients de $\exp(tT_i)$ sont des polynômes de degré maximum $d_i - 1$, où d_i est l'ordre de nilpotence de T_i. On en déduit que, lorsque $t \to +\infty$, le bloc $e^{\lambda_i t}\exp(tT_i)$ est borné si, et seulement si, soit $\mathrm{Re}\,\lambda_i < 0$, soit $\mathrm{Re}\,\lambda_i = 0$ et $T_i = 0$; et ce même bloc tend vers 0 si, et seulement si, $\mathrm{Re}\,\lambda_i < 0$. En résumé :

1. Pour que le système (S_0) soit stable, il faut, et il suffit, que toutes les valeurs propres de A dans $\mathbb{C}$ aient une partie réelle négative ou nulle et que celles de partie réelle nulle soient associées à des espaces caractéristiques qui sont des espaces propres. Cela revient à dire que les blocs de Jordan correspondants sont de taille 1.
2. Pour que le système (S_0) soit asymptotiquement stable, il faut, et il suffit, que toutes les valeurs propres de A dans $\mathbb{C}$ aient une partie réelle strictement négative.

Le cas d'une équation à coefficients constants donne lieu à un critère un peu plus simple : voir l'exercice II.7.21 de la page 877.

2.2.4 Le flot d'une équation linéaire autonome

Soit (S_B) une équation linéaire vectorielle autonome. Pour toute condition initiale $Y_0 \in K^n$ et pour tout temps initial $t_0 \in \mathbb{R}$, notons provisoirement F_{t_0,Y_0} l'unique solution de (S_B) qui prend la valeur Y_0 en t_0. D'après le théorème 35 de la page 837, on a, pour tout autre temps initial $u_0 := t_0 + \tau$:

$$\forall t \in \mathbb{R}\ ,\ F_{u_0,Y_0}(t+\tau) = F_{t_0,Y_0}(t).$$

En remplaçant t par $t_0 + t$, on obtient :

$$\forall t \in \mathbb{R}\ ,\ F_{u_0,Y_0}(u_0+t) = F_{t_0,Y_0}(t_0+t).$$

La position de F_{t_0,Y_0} au temps $t_0 + t$ ne dépend donc que de sa position Y_0 au temps t_0 et de l'intervalle de temps t, et non du temps initial t_0. Nous noterons $\Phi(t, Y_0)$ cette position.

Proposition 39. L'application $(t, Y_0) \mapsto \Phi(t, Y_0)$ est une action du groupe $\mathbb{R}$ sur l'espace des phases K^n.

Démonstration. Il faut d'abord vérifier l'égalité $\Phi(0, Y_0) = Y_0$, qui est tautologique. Soient ensuite $u, v \in \mathbb{R}$ et $Y_0 \in K^n$. Notons $Y_1 := \Phi(u, Y_0)$. La deuxième égalité à vérifier est $\Phi(v, Y_1) = \Phi(u+v, Y_0)$. Avec les notations précédentes, la fonction F_{t_0,Y_0} est la solution de (S_B) dont la condition initiale en $t_0 + u$ est Y_1, d'où l'égalité : $F_{t_0,Y_0} = F_{t_0+u,Y_1}$. Appliquant ces deux fonctions à $t_0 + u + v$, on trouve $\Phi(u+v, Y_0) = \Phi(v, Y_1)$, c'est-à-dire $\Phi(u+v, Y_0) = \Phi\big(v, \Phi(u, Y_0)\big)$, ce qui est la définition d'une action de groupe. ∎

Cette action est appelé *flot* de l'équation (S_B). Ses orbites sont les courbes intégrales. Le flot de (E_b) se définit alors comme le flot du système (S_B) associé.

Exercice 23.
Décrire le flot de l'équation $\ddot{x} + x = 0$.
Solution. La solution de condition initiale (x, y) au temps 0 est la fonction $t \mapsto x\cos t + y\sin t$. Ainsi, $F_{0,(x,y)}(t) = (x\cos t + y\sin t, y\cos t - x\sin t)$ et :

$$\Phi\big(t,(x,y)\big) = (x\cos t + y\sin t, y\cos t - x\sin t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Ce flot agit sur $\mathbb{R}^2$ par rotations d'angle $-t$.

2.3 Équations à coefficients périodiques

On étudie maintenant des systèmes différentiels linéaires homogènes à coefficients périodiques. Nous supposerons donc que la matrice $A(t)$ est définie continue sur $I := \mathbb{R}$ et qu'elle admet une période $T > 0$:

$$\forall t \in \mathbb{R}\ ,\ A(t+T) = A(t).$$

Dans ce cas, les solutions de (S_0) sont définies sur $\mathbb{R}$, et il en est de même de la matrice résolvante $R(t, t_0)$. Celle-ci vérifie la relation de périodicité suivante :

$$\forall t, t_0 \in \mathbb{R}\ ,\ R(t+T, t_0+T) = R(t, t_0).$$

En effet, les deux membres de l'égalité sont solutions du système de matrice $A(t+T) = A(t)$; et satisfont à la condition initiale $\mathcal{F}(t_0) = I_n$.

On appelle *application de Floquet* la matrice :

$$C(t_0) := R(t_0 + T, t_0).$$

On peut la considérer comme un automorphisme de l'espace des phases : une solution qui passe par $Y_0 \in K^n$ à l'instant t_0 passe par $C(t_0)Y_0$ à l'instant $t_0 + T$, c'est à dire après l'écoulement d'une période.

L'application de Floquet dépend de t_0, mais sa classe de conjugaison n'en dépend pas :

$$\forall t_1, t_0 \in \mathbb{R}\ ,\ C(t_1) = R(t_1, t_0)C(t_0)R(t_1, t_0)^{-1}.$$

Cela découle en effet des relations de Chasles pour la résolvante et de la relation de périodicité énoncée plus haut. Il s'ensuit que les valeurs propres de $C(t_0)$ ne dépendent pas de t_0. On les appelle *nombres caractéristiques* du système (S_0).

Pour les mêmes raisons, l'application de Floquet est périodique :

$$\forall t_0 \in \mathbb{R}\ ,\ C(t_0 + T) = C(t_0).$$

En combinant les diverses égalités qui précèdent, on obtient l'utile relation :

$$\forall t, t_0 \in \mathbb{R}\ ,\ \forall k \in \mathbb{N}\ ,\ R(t + kT, t_0) = R(t, t_0)C(t_0)^k. \tag{31}$$

En effet, le membre de gauche se décompose (relation de Chasles) en :

$$R(t + kT, t_0 + kT) = R(t, t_0) \qquad \text{et} \qquad R(t_0 + kT, t_0) = C(t_0)^k.$$

Exercice 24.
Que peut-on dire de l'application de Floquet lorsque $\operatorname{Tr} A$ est identiquement nulle ? Lorsque A est antisymétrique réelle ?
Solution. D'après la proposition 33 de la page 832, $C(t_0) \in SL_n(K)$ dans le premier cas, $C(t_0) \in SO_n(\mathbb{R})$ dans le deuxième cas.

2.3.1 Solutions périodiques

Proposition 40. (i) La solution F de (S_0) satisfaisant à la condition initiale $F(t_0) = Y_0$ est périodique de période T (resp. de période kT, $k \in \mathbb{N}^*$) si, et seulement si, $C(t_0)Y_0 = Y_0$ (resp. si, et seulement si, $C(t_0)^k Y_0 = Y_0$).
(ii) Le système S_0 admet des solutions T-périodiques non triviales (resp. des solutions kT-périodiques non triviales, $k \in \mathbb{N}^*$) si, et seulement si, l'un des nombres caractéristiques est 1 (resp. est une racine $k^{ème}$ de l'unité).

Démonstration. L'assertion (ii) est conséquence immédiate de l'assertion (i).
Puisque $F(t) = R(t, t_0)Y_0$, en vertu de (31), l'égalité $F(t + kT) = F(t)$ équivaut à $C(t_0)^k Y_0 = Y_0$. L'autre partie de (i) est le cas particulier $k := 1$.

Exercice 25.
Soit $\lambda \in K$. On pose $Y = e^{\lambda t} Z$. Quelle équation différentielle est vérifiée par Z ? Quels sont ses nombres caractéristiques ? À quelle condition le système (S_0) admet-il des solutions non triviales de la forme $e^{\lambda t} G$, où G est T-périodique ?
Solution. L'équation satisfaite par Z est $Z' = (A - \lambda I_n)Z$. C'est un système à coefficients constants de résolvante $e^{-\lambda t} R(t, t_0)$, donc d'application de Floquet $e^{-\lambda T} C(t_0)$. Ses nombres caractéristiques sont ceux de (S_0) multipliés par $e^{-\lambda T}$. Il admet des solutions non triviales T-périodiques si, et seulement si, l'un de ces nombres caractéristiques est 1. La condition demandée est donc : $\lambda = \dfrac{1}{T} \log c$, où $\log c$ désigne un logarithme d'un nombre caractéristique de (S_0) (autrement dit, un complexe γ tel que $e^{\gamma} = c$, voir le module II.6).

Proposition 41. Soit $B \in M_n(\mathbb{C})$ une matrice telle que $e^{TB} = C(t_0)$ (théorème 20 de la page 279). Alors $R(t, t_0) = Z(t)e^{tB}$, où la fonction matricielle Z est T-périodique.

Démonstration. Posons $Z(t) := R(t, t_0)e^{-tB}$. Alors :

$$Z(t+T) = R(t+T, t_0)e^{-TB}e^{-tB} = R(t, t_0)C(t_0)e^{-TB}e^{-tB} = R(t, t_0)e^{-tB} = Z(t).$$

Toute solution de (S_0) est donc de la forme $Z(t)e^{tB}Y_0$ où Z est T-périodique, ce qui permet de retrouver le fait que les valeurs propres de B sont les $\dfrac{1}{T} \log c$.

2.3.2 Stabilité des équations à coefficients périodiques

Proposition 42. Le système à coefficients périodiques (S_0) est stable, resp. asymptotiquement stable, si, et seulement si, $C(t_0)^k$ est bornée, resp. tend vers 0 lorsque l'entier k tend vers $+\infty$.

Démonstration. On va invoquer le critère sur la résolvante énoncé au (iii) de la proposition 38 de la page 844. Si le système est stable, resp. asymptotiquement stable, ce critère implique en particulier que $R(t_0 + kT, t_0) = C(t_0)^k$ est bornée, resp. tend vers 0 lorsque l'entier k tend vers $+\infty$.
Pour la réciproque, on note a le maximum de $\|R(t, t_0)\|$ lorsque $t \in [t_0, t_0 + T]$. D'après la relation (31), on a :

$$\forall t \in \mathbb{R}\ ,\ \|R(t, t_0)\| \leqslant a \|C(t_0)\|^k\ ,$$

où k est la partie entière de $\dfrac{t - t_0}{T}$ et $\|-\|$ une norme subordonnée. La conclusion s'ensuit. ∎

Le critère semble faire jouer un rôle privilégié à un point particulier t_0, mais ce n'est pas le cas : pour un autre point, on sait que $C(t_1)$ est conjuguée à $C(t_0)$; or, les propriétés ci-dessus (comme d'ailleurs celles du corollaire ci-dessous) sont conservées par conjugaison.

Corollaire 43.
(i) Soit $\|-\|$ une norme subordonnée. Si $\|C(t_0)\| \leqslant 1$, resp. si $\|C(t_0)\| < 1$, le système (S_0) est stable, resp. asymptotiquement stable.
(ii) Si toutes les valeurs propres de $C(t_0)$ sont de module $|\lambda| < 1$, le système est asymptotiquement stable.
(iii) Si l'une au moins des valeurs propres de $C(t_0)$ est de module $|\lambda| > 1$, le système est instable.

Ces propriétés relèvent de l'algèbre linéaire et les deux dernières se prouvent par trigonalisation. C'est de la même manière qu'on voit que la réciproque de (iii) (ou l'analogue de (ii) pour une inégalité large) est fausse : si par exemple $C(t_0)$ n'admet que la valeur propre 1 mais n'est pas l'identité, ses puissances ne tendent pas vers 0.

Exemple. Soit f une fonction réelle continue de période $T > 0$. L'équation $\ddot{x} + f(t)x = 0$ se ramène au système de matrice $A := \begin{pmatrix} 0 & 1 \\ -f(t) & 0 \end{pmatrix}$.
Comme $\operatorname{Tr} A = 0$, l'application de Floquet est dans $SL_2(\mathbb{R})$. L'étude des racines du trinôme du second degré $X^2 - \operatorname{Tr}(C(t_0))\, X + 1$ montre que :
- Si $|\operatorname{Tr} C(t_0)| < 2$, le système est asymptotiquement stable.
- Si $|\operatorname{Tr} C(t_0)| > 2$, le système est instable.

Si $\operatorname{Tr} C(t_0) = 2\varepsilon$, $\varepsilon = \pm 1$, les deux valeurs propres sont égales à ε mais on ne peut pas conclure en toute généralité. Si $C(t_0) = \varepsilon I_2$, le système est stable (et même à solutions périodiques). Sinon, il est instable.

Par exemple, si $f = 0$, la matrice de Floquet est $C(t_0) = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix}$.

Résonance paramétrique

Dans l'exemple ci-dessus, prenons $f := \omega^2 + \varepsilon g$, où $\omega \neq 0$, $\varepsilon > 0$ est petit et g est de période T. La matrice de (S_0) est alors $A = A_0 + \varepsilon B$, où $A_0 := \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}$ et $B := \begin{pmatrix} 0 & 0 \\ -g(t) & 0 \end{pmatrix}$. Notons respectivement $R(t, t_0)$ et $R_\varepsilon(t, t_0)$ les résolvantes associées à A_0 et à A_ε. D'après l'exercice II.7.30 de la page 878, on a :

$$\|R_\varepsilon(t_0 + T, t_0) - R(t_0 + T, t_0)\| = O(\varepsilon) \text{ lorsque } \varepsilon \to 0.$$

Il en découle que l'application de Floquet $C_\varepsilon(t_0)$ du système de matrice A est très proche de e^{TA_0}. Les valeurs propres de cette dernière matrice sont $e^{\pm\omega \mathrm{i} T}$, sa trace est donc $2\cos\omega T$, d'où :

$$\operatorname{Tr} C(t_0) = 2\cos\omega T + \mathrm{O}(\varepsilon) \text{ lorsque } \varepsilon \to 0.$$

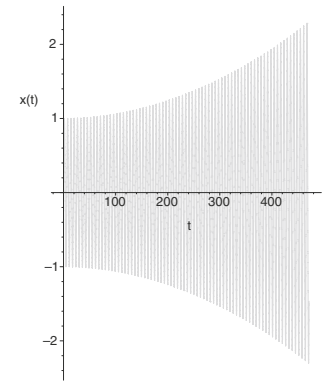
Notons $T_0 := \dfrac{2\pi}{\omega}$ la période propre du système non perturbé (de matrice A_0).
On en conclut :

Proposition 44. Si T n'est pas un multiple de la demi-période $\dfrac{T_0}{2}$, le système perturbé est asymptotiquement stable pour ε suffisamment petit.

Démonstration.

En effet, dans ce cas, $|2\cos\omega T| < 2$ et $|\operatorname{Tr} C(t_0)| < 2$ pour ε assez petit. ∎

Lorsque T est multiple de la demi-période, il peut y avoir instabilité : c'est le phénomène de *résonance paramétrique*. Il est assez difficile de l'établir rigoureusement, mais nous en proposerons une illustration numérique. Dans l'exemple illustré ci-contre, nous avons pris $\omega := 1$ et $\varepsilon := 0,01$. et les conditions initiales $x(0) := 0$, $\dot{x}(0) := 1$. Le logiciel Maple calcule la matrice de Floquet au bout d'une période π du système perturbé (la période du système de départ est 2π) :



$$C \approx \begin{pmatrix} -1{,}00003124649242126 & 0.00785897128871685356 \\ 0{,}00784911131075739267 & -1{,}00003125069915599 \end{pmatrix}.$$

Ses valeurs propres sont $\lambda \approx -0{,}9921772087$ et $\mu \approx -1{,}007885288$.
Comme $|\mu| > 1$, il y a instabilité.

3 Équations différentielles non linéaires

Nous allons maintenant *effleurer* l'étude d'équations plus générales, que nous appellerons « non linéaires » pour signifier qu'elles ne sont pas nécessairement linéaires. Les résultats de cette section s'appliquent donc en particulier aux équations linéaires. Nous ne considérerons que des fonctions *réelles*. On suppose choisie une norme $\| \ \|$ sur chaque espace vectoriel $\mathbb{R}^n$: il sera clair que les résultats qui suivent ne dépendent pas du choix de cette norme.

3.1 Généralités

3.1.1 Vocabulaire

La forme la plus générale d'une équation différentielle scalaire d'ordre n est $A(t, y, \dots, y^{(n)}) = 0$, où A est une fonction de $n+1$ variables réelles. On peut considérer cette relation comme un problème de fonction implicite. Pour une fonction A convenable (par exemple, si l'on peut appliquer le théorème des fonctions implicites), on pourra exprimer $y^{(n)}$ en fonction des autres termes :

$$y^{(n)} = \varphi(t, y, \dots, y^{(n-1)}). \tag{32}$$

Une telle équation différentielle (scalaire d'ordre n) est dite *sous forme résolue*.

Exemples. L'équation différentielle $\dot{x}^2 + x^2 = E$ n'est pas sous forme résolue. Sur tout intervalle où $\dot{x}$ garde un signe constant, elle se ramène à l'une des équations sous forme résolue $\dot{x} = \sqrt{E - x^2}$ ou $\dot{x} = -\sqrt{E - x^2}$.
L'équation différentielle linéaire $ty' - y = 0$ n'est pas sous forme résolue. Elle se ramène à l'équation sous forme résolue $y' = \dfrac{1}{t}y$, mais à condition de restreindre de $\mathbb{R}$ à $\mathbb{R}^*$ le domaine dans lequel varie la variable t.

L'équation scalaire (32) d'ordre n se ramène à une équation vectorielle d'ordre 1 par le même procédé de vectorialisation que précédemment. On pose $Y := (y, \dots, y^{(n-1)})$ et l'on définit une fonction $\Phi(t, Y)$ de deux variables (l'une réelle, l'autre dans $\mathbb{R}^n$) en posant :

$$\Phi(t, y_0, \dots, y_{n-1}) := \big(y_1, \dots, y_{n-1}, \varphi(t, y_0, \dots, y_{n-1})\big).$$

L'équation (32) est alors équivalente à la suivante :

$$Y' = \Phi(t, Y). \tag{33}$$

Définition 12. On appelle *équation différentielle vectorielle d'ordre* 1 *(résolue)* sur l'ouvert non vide $U \subset \mathbb{R} \times \mathbb{R}^n$ une équation de la forme (33), où Φ est une fonction continue de U dans $\mathbb{R}^n$. Une *solution* de (33) sur un intervalle $J \subset \mathbb{R}$ est une fonction dérivable F de J dans $\mathbb{R}^n$ telle que :

$$\forall t \in J \ , \ \big(t, F(t)\big) \in U \quad \text{et} \quad F'(t) = \Phi\big(t, F(t)\big). \tag{34}$$

La *trajectoire* de la solution F est l'image par J de l'application $t \mapsto \big(t, F(t)\big)$ dans *l'espace des phases élargi* $\mathbb{R} \times \mathbb{R}^n$. La *courbe intégrale* de la solution F est l'image par J de l'application F dans *l'espace des phases* $\mathbb{R}^n$.

Il découle de la relation (34) que toute solution F sur J est de classe $\mathcal{C}^1$; plus généralement, si la fonction Φ est de classe $\mathcal{C}^k$ (resp. de classe $\mathcal{C}^\infty$), alors la solution F est de classe $\mathcal{C}^{k+1}$ (resp. de classe $\mathcal{C}^\infty$). La restriction de la solution F à un intervalle $J' \subset J$ est une solution sur J'. On dit que la solution F est *maximale* si elle n'est pas restriction d'une solution sur un intervalle contenant strictement J. Si l'ouvert U est de la forme $U := I \times V$, où I est un intervalle (nécessairement ouvert) de $\mathbb{R}$, nous appellerons *globale* toute solution sur I.

Exemples.

1. Soit $A(t)$ une fonction matricielle continue sur l'intervalle I.
   La fonction $\Phi(t, Y) := A(t)Y$ définit une équation vectorielle d'ordre 1 sur $I \times \mathbb{R}^n$ (sections 1 et 2).
2. Les équations $y' = y^2$ et $y' = 2\sqrt{|y|}$ de l'introduction sont des équations scalaires d'ordre 1 définies sur $\mathbb{R} \times \mathbb{R}$. La fonction Φ (ou φ) ne dépend pas explicitement de t : de telles équations, dites *autonomes*, sont étudiées à la section 3.3.
3. L'équation résolue $y' = \dfrac{1}{t}y$, définie sur $\mathbb{R}^* \times \mathbb{R}$ admet pour solutions maximales les fonctions de la forme $f(t) = at$ sur $\mathbb{R}_+^*$ et les fonctions de la forme $f(t) = at$ sur $\mathbb{R}_-^*$.
4. L'équation différentielle $y' = 2ty^2$ est scalaire d'ordre 1 est définie sur $\mathbb{R} \times \mathbb{R}$. Elle n'est pas linéaire.

Exercice 26.
Déterminer toutes les solutions maximales de $y' = 2ty^2$.

Solution. Il y a, bien entendu, la solution globale identiquement nulle. Si f est une solution qui ne s'annule pas sur un intervalle ouvert J, on a $\left(\dfrac{1}{f}\right)' = -2t$ sur J, donc $f_a(t) = \dfrac{1}{a - t^2}$ sur J, où $a \in \mathbb{R}$. Les solutions globales sont les f_a telles que $a < 0$.
Les autres solutions maximales sont f_0 sur $\mathbb{R}_+^*$, f_0 sur $\mathbb{R}_-^*$ et, pour tout $a > 0$, f_a sur $]-\infty, -\sqrt{a}[$, f_a sur $]-\sqrt{a}, \sqrt{a}[$ et f_a sur $]\sqrt{a}, +\infty[$ (voir l'exercice II.7.31 de la page 878).

Définition 13. On appelle *problème de Cauchy* une équation différentielle (33) assortie d'une *condition initiale* $(t_0, Y_0) \in U$:

$$\begin{cases} Y' = \Phi(t, Y), \\ Y(t_0) = Y_0. \end{cases} \tag{35}$$

Une *solution* de ce problème de Cauchy est une solution de $Y' = \Phi(t, Y)$ sur un intervalle J tel que t_0 est intérieur à J et dont la trajectoire passe par (t_0, Y_0), *i.e.* $F(t_0) = Y_0$.

Proposition 45. La fonction F est solution du problème de Cauchy (35) sur l'intervalle J contenant t_0 si, et seulement si, elle est continue sur J, telle que l'application $t \mapsto \big(t, F(t)\big)$ envoie J dans U, et solution de l'équation intégrale :

$$\forall t \in J\ ,\ F(t) = Y_0 + \int_{t_0}^{t} \Phi\big(s, F(s)\big)\,\mathrm{d}s. \tag{36}$$

Démonstration. C'est la même que celle du lemme 23 de la page 826. ∎

Les définitions précédentes sont parfois étendues de la façon suivante. Tout d'abord, on peut évidemment considérer des fonctions Φ qui ne sont pas continues. Ensuite, on admet parfois des solutions *continues et de classe $\mathcal{C}^1$ par morceaux*. Dans le cas d'un segment, cela signifie que l'intervalle $J := [a, b]$ admet une subdivision $a = t_0 < t_1 < \cdots < t_N = b$ telle que la fonction F est de classe $\mathcal{C}^1$ sur chacun des intervalles $[t_i, t_{i+1}]$ (pour $0 \leqslant i \leqslant n-1$) ; la fonction F est donc automatiquement continue sur J. On requiert alors que la condition (34) soit satisfaite sur chacun de ces intervalles. L'égalité fondamentale $F(b) - F(a) = \displaystyle\int_a^b F'(s)\,\mathrm{d}s$ s'étend sans difficulté à une telle fonction, en attribuant une valeur arbitraire aux $F(t_i)$: elle est vérifiée sur chaque $[t_i, t_{i+1}]$ et l'on applique la relation de Chasles. Il en découle que l'inégalité des accroissements finis est également vérifiée.

Lorsque J est un intervalle semi-ouvert ou ouvert, on requiert les mêmes propriétés sur chaque segment de J.

Exemple. L'équation $y' = E(y)$ (partie entière de y) admet la solution continue et de classe $\mathcal{C}^1$ par morceaux définie par : $f(t) = t + 1$ sur $[0, 1]$, $f(t) = 2t$ sur $[1, 3/2]$, $f(t) = 3t - 3/2$ sur $[3/2, 11/6]$.

3.1.2 Cylindre de sécurité

Nous conservons ici les notations de la définition 12 de la page 850. Le *cylindre de longueur $2T > 0$ et de diamètre $2r > 0$ centré en $(t_0, Y_0) \in U$* est :

$$C := [t_0 - T, t_0 + T] \times \overline{B}(Y_0, r),$$

où $\overline{B}(Y_0, r)$ désigne la boule fermée de centre Y_0 et de rayon r dans $\mathbb{R}^n$. Puisque U est ouvert, il existe toujours un tel cylindre tel que $C \subset U$.

Définition 14. Le cylindre $C \subset U$ est un *cylindre de sécurité* de l'équation (33) si, pour toute solution F du problème de Cauchy (35) sur un intervalle $J \subset [t_0 - T, t_0 + T]$, la trajectoire de F est incluse dans C (on dit encore improprement que « la solution F reste contenue dans $\overline{B}(Y_0, r)$ »).

Proposition 46. Gardons les notations précédentes. Soit $M \in \mathbb{R}$ le maximum de la fonction continue $\|\Phi\|$ sur le compact C. On pose $T_0 := \min\left(T, \frac{r}{M}\right)$ (donc T si $M = 0$). Alors le cylindre :

$$C_0 := [t_0 - T_0, t_0 + T_0] \times \overline{B}(Y_0, r),$$

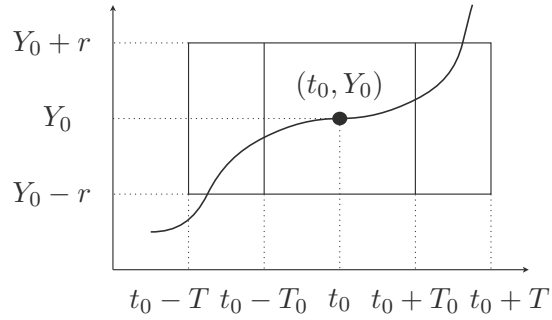
est un cylindre de sécurité pour l'équation (34). En particulier, on peut toujours construire un cylindre de sécurité.

Démonstration. Soit F une solution sur $J \subset [t_0 - T, t_0 + T]$ du problème de Cauchy (35). Supposons que sa trajectoire sorte de C_0, donc qu'il existe $t' \in J$ tel que $\|F(t') - Y_0\| > r$. Nous supposerons que $t' > t_0$, le cas où $t' < t_0$ se traitant de manière similaire. Puisque F est continue, il existe $t_1 \in]t_0, t'[$ tel que $\|F(t_1) - Y_0\| = r$ et $\|F - Y_0\| < r$ sur $[t_0, t_1[$. La trajectoire de F sur $[t_0, t_1]$ reste donc contenue dans C, et l'on a donc $\|F'(t)\| = \big\|\Phi\big(t, F(t)\big)\big\| \leqslant M$ pour $t \in [t_0, t_1]$. D'après le théorème des accroissements finis, on en déduit que $\|F(t_1) - F(t_0)\| \leqslant M(t_1 - t_0)$, donc que :

$$t_1 - t_0 \geqslant \frac{\|F(t_1) - F(t_0)\|}{M} = \frac{r}{M} \geqslant T_0.$$

On a donc $t_1 \geqslant t_0 + T_0$, donc $t' > t_0 + T_0$, ce qui contredit l'hypothèse $t' \in J$. ∎

La figure suivante représente cette situation lorsque $n := 1$.



Une solution de l'équation ne peut donc sortir du cylindre de sécurité C_0 que par la droite : $t = t_0 + T_0$ ou par la gauche : $t = t_0 - T_0$.

Exemple. L'équation $y' = 2ty^2$ avec la condition initiale $y(0) = 1$ admet la solution $f(t) := \dfrac{1}{1 - t^2}$. Prenons $T := 1$ et $r > 0$ arbitraire : la solution f s'échappe du cylindre $C := [-1, 1] \times [1 - r, 1 + r]$ par le haut en $t := \pm\sqrt{\dfrac{r}{r+1}} \in]-1, 1[$.

On a ici $M = 2(1 + r)^2$, d'où $T_0 = \dfrac{r}{2(1 + r)^2}$.

Sur l'intervalle $[-T_0, T_0]$, on a $1 \leqslant f(t) \leqslant \dfrac{1}{1 - T_0^2}$, et le lecteur en déduira sans peine que la restriction de f à $[-T_0, T_0]$ reste confinée dans le cylindre de sécurité $C_0 = [-T_0, T_0] \times [1 - r, 1 + r]$.

Exercice 27.
Étudier de même le cas de $y' = y^2$ sur le cylindre $C := [-1, 1] \times [1 - r, 1 + r]$.

Solution. La solution correspondant à la condition initiale $(0, 1)$ est $f(t) := \dfrac{1}{1-t}$, qui s'échappe de C en $t := \dfrac{r}{r+1}$. On a ici $M = (1 + r)^2$, $T_0 = \dfrac{r}{(1+r)^2}$ et $C_0 = [-T_0, T_0] \times [1 - r, 1 + r]$.

3.1.3 Solutions approchées

Les équations non linéaires sont plus difficiles à résoudre et même l'existence de solutions ne va pas de soi. Nous allons, avec des idées « numériques » proches de celles du module II.8, construire des fonctions qui sont presque solution. Soit donc F une fonction continue et $\mathcal{C}^1$ par morceaux sur le segment $J := [a, b]$ et soit $a = t_0 < t_1 < \cdots < t_N = b$ une subdivision telle que la fonction F de J dans $\mathbb{R}^n$ est de classe $\mathcal{C}^1$ sur chacun des intervalles $[t_i, t_{i+1}]$ (pour $0 \leqslant i \leqslant N - 1$).

Définition 15. Soit $\varepsilon > 0$. On dit que F est une *solution approchée à ε près* de l'équation (33) sur l'intervalle J si l'image de J par l'application $t \mapsto \big(t, F(t)\big)$ est contenue dans U et si, en tout point t de l'un des intervalles ouverts $]t_i, t_{i+1}[$, on a :

$$\big\|F'(t) - \Phi\big(t, F(t)\big)\big\| \leqslant \varepsilon. \tag{37}$$

En prenant $\varepsilon := 0$, on retrouverait la notion de solution (exacte) telle qu'elle a été étendue aux fonctions continues de classe $\mathcal{C}^1$ par morceaux.

La définition ci-dessus n'affirme pas l'existence d'une solution exacte dont F est proche. Sous des hypothèses convenables, cela découlera du théorème 50 de la page 857 et de la section 3.2.1.

Nous reprenons les notations de la section 3.1.2, en particulier le cylindre $C := [t_0 - T, t_0 + T] \times \overline{B}(Y_0, r)$, le maximum M de $\|\Phi\|$ sur C et le cylindre de sécurité $C_0 = [t_0 - T_0, t_0 + T_0] \times \overline{B}(Y_0, r)$, où $T_0 := \min\left(T, \dfrac{r}{M}\right)$.

Proposition 47. Soit $T_\varepsilon := \min\left(T, \dfrac{r}{M + \varepsilon}\right)$.

Le cylindre $C_\varepsilon = [t_0 - T_\varepsilon, t_0 + T_\varepsilon] \times \overline{B}(Y_0, r)$ est un cylindre de sécurité pour les solutions approchées à ε près : toute telle solution sur un intervalle contenu dans $[t_0 - T_\varepsilon, t_0 + T_\varepsilon]$ reste confinée dans C_ε.

Démonstration. Il suffit d'adapter la preuve de la proposition 46 de la page précédente : si $t_1 > t_0$ est le premier point de sortie du cylindre, on a, pour $t \in [t_0, t_1]$:

$$\big\|F'(t)\big\| \leqslant \big\|\Phi\big(t, F(t)\big)\big\| + \varepsilon \leqslant M + \varepsilon$$

(par l'inégalité du triangle), et l'on achève comme auparavant. L'utilisation de l'inégalité des accroissements finis a été justifiée page 852. ∎

Nous allons construire par, la *méthode d'Euler*, une solution approchée à ε près sur $[t_0, t_0+T_0]$ (le côté gauche $[t_0-T_0, t_0]$ se traite de la même manière). Pour cela, on se donne une subdivision $t_0 < t_1 < \cdots < t_N = t_0+T_0$ et l'on va définir par récurrence $Y_0 = F(t_0), \ldots, Y_N = F(t_N)$; puis on reliera les points (t_i, Y_i) par des segments : la ligne brisée obtenue sera la trajectoire de la solution approchée F. Autrement dit, F est une fonction affine sur chacun des segments $[t_i, t_{i+1}]$.

La définition des Y_i est donnée par les formules suivantes :

$$\forall i \in [\![0, N-1]\!]\ ,\ Y_{i+1} := Y_i + (t_{i+1}-t_i)\Phi(t_i, Y_i). \tag{38}$$

Géométriquement, la trajectoire part du point (t_i, Y_i) dans la direction qui serait celle de la tangente d'une solution exacte.

Théorème 48. (i) Les points (t_i, Y_i) sont bien définis pour $0 \leqslant i \leqslant N$. La trajectoire de F est contenue dans C_0.
(ii) Si le *pas* $h := \max\limits_{0\leqslant i\leqslant N-1}(t_{i+1}-t_i)$ de la subdivision est assez petit, on obtient une solution approchée à ε près.

Démonstration. (i) Montrons par récurrence que $\|Y_i - Y_0\| \leqslant M(t_i - t_0)$.
Puisque $M(t_i - t_0) \leqslant MT_0 \leqslant r$ (par construction de C_0), on en déduira que $(t_i, Y_i) \in C_0 \subset U$, donc que $\Phi(t_i, Y_i)$ est bien défini, et donc que le calcul de Y_{i+1} est possible. L'inégalité voulue est évidente pour $i := 0$. Si l'on a $\|Y_i - Y_0\| \leqslant M(t_i - t_0)$, puisque $(t_i, Y_i) \in C_0 \Rightarrow \|\Phi(t_i, Y_i)\| \leqslant M$, on trouve que $\|Y_{i+1} - Y_i\| = \|(t_{i+1}-t_i)\Phi(t_i, Y_i)\| \leqslant M(t_{i+1}-t_i)$, d'où $\|Y_{i+1} - Y_0\| \leqslant M(t_{i+1}-t_0)$ par l'inégalité du triangle. Puisque tous les points (t_i, Y_i) appartiennent à l'ensemble convexe C_0, les segments qui les relient sont contenus dans C_0.
(ii) Pour estimer l'erreur commise, introduisons des notations. Soient $h_i := t_{i+1} - t_i$ un *pas* de la subdivision et $Y_i' := \Phi(t_i, Y_i)$ la pente correspondante. Sur chaque segment $[t_i, t_{i+1}]$, la solution F construite est affine, donc de classe $\mathcal{C}^1$ et de dérivée $\Phi(t_i, Y_i)$, et l'on a l'inégalité :

$$\begin{aligned}\big\|F'(t) - \Phi\big(t, F(t)\big)\big\| &\leqslant \max_{0\leqslant u\leqslant h_i} \big\|\Phi(t_i, Y_i) - \Phi(t_i+u, Y_i + uY_i')\big\| \\ &\leqslant \max_{\substack{0\leqslant u\leqslant h \\ 0\leqslant \|Z\|\leqslant hM}} \|\Phi(t_i, Y_i) - \Phi(t_i+u, Y_i+Z)\|\,.\end{aligned}$$

La fonction Φ étant continue sur le compact C_0, elle est uniformément continue. Il existe donc un réel $\eta > 0$ tel que, sur C_0 :

$$\big|\overline{t} - t\big| + \big\|\overline{Y} - Y\big\| < \eta \Longrightarrow \big\|\Phi(\overline{t}, \overline{Y}) - \Phi(t, Y)\big\| < \varepsilon.$$

On en conclut que, si $h < \dfrac{\eta}{1+M}$, alors F est une solution approchée à ε près. ∎

On déduit de cette démonstration une estimation de l'erreur dans la méthode d'Euler. Notons $\|(t,Y)\| := |t| + \|Y\|$ (c'est une norme sur l'espace des phases élargi $\mathbb{R}\times\mathbb{R}^n$) et :

$$\omega(\eta) := \max_{\|(\overline{t},\overline{Y})-(t,Y)\|\leqslant\eta} \big\|\Phi(\overline{t},\overline{Y}) - \Phi(t,Y)\big\| \tag{39}$$

le *module de continuité* de Φ sur C_0 ; son existence et sa finitude découlent de la continuité de Φ et de la compacité de C_0.
L'uniforme continuité de Φ sur C_0 se traduit par la relation : $\lim_{\eta\to 0}\omega(\eta) = 0$.

Corollaire 49. La fonction F construite par la méthode d'Euler pour une subdivision de pas h est une solution approchée à $\omega\,(h(1+M))$ près.

Exemples. La fonction exponentielle est solution du problème de Cauchy $y' = y, y(0) = 1$. On va construire à la fois les t_i et les y_i pour obtenir une solution approchée en imposant que, t_i étant trouvé, t_{i+1} soit le plus grand possible. On pose d'abord $t_0 := 0$ et $f(0) := y_0 := 1$. Sur $[t_i, t_{i+1}]$, on a $f(t) = y_i + (t-t_i)y_i$ et la condition $|f'(t) - f(t)| \leqslant \varepsilon$ entraîne $|(t-t_i)y_i| \leqslant \varepsilon$. On prendra donc $t_{i+1} = t_i + \dfrac{\varepsilon}{y_i}$ et $y_{i+1} = y_i + \varepsilon$. Ainsi, $y_i = 1 + i\varepsilon$ et $t_i = \sum_{j=0}^{i-1} \dfrac{\varepsilon}{1+(j-1)\varepsilon}\cdot$ Comme la série $\sum \dfrac{\varepsilon}{1+(j-1)\varepsilon}$ diverge, on obtient ainsi une solution approchée sur $\mathbb{R}_+$.

Exercice 28.
Appliquer la même méthode à l'équation autonome $y' = \varphi(y)$ dans laquelle φ est strictement croissante et strictement positive, avec la condition initiale $t_0 := 0$, $y_0 > 0$. Examiner le cas particulier $y' = y^\alpha, y(0) = 1$, avec $\alpha > 0$: la solution approchée obtenue s'étend-elle à $\mathbb{R}_+$?

Solution. On doit trouver $t_{i+1} = t_i + h_i$ et $y_{i+1} = y_i + h_i\varphi(y_i)$, avec la contrainte que, pour $0 \leqslant u \leqslant h_i$, on ait $\big|\varphi(y_i) - \varphi\big(y_i + u\varphi(y_i)\big)\big| \leqslant \varepsilon$. La plus grande valeur possible du pas h_i est celle telle que $\varphi(y_{i+1}) = \varphi(y_i) + \varepsilon$. On a donc $\varphi(y_i) = \varphi(y_0) + i\varepsilon$ (sous réserve que $\varphi(y_0) + i\varepsilon$ appartienne à $\operatorname{Im}\varphi$) et $t_{i+1} := t_i + \dfrac{y_{i+1} - y_i}{\varphi(y_i)}$.
Si $\varphi(y) := y^\alpha$, on obtient ainsi $y_i = (1+i\varepsilon)^{1/\alpha}$ et :

$$h_i := t_{i+1} - t_i = \frac{(1+(i+1)\varepsilon)^{1/\alpha} - (1+i\varepsilon)^{1/\alpha}}{1+i\varepsilon}\cdot$$

Le lecteur vérifiera que, lorsque i tend vers $+\infty$, $h_i \sim \dfrac{\varepsilon^{1/\alpha-1}}{\alpha} i^{1/\alpha-2}$. Pour que la solution s'étende à $\mathbb{R}_+$, il faut, et il suffit, que t_i tende vers $+\infty$, c'est-à-dire que la série $\sum h_i$ diverge, ce qui équivaut à $1/\alpha - 2 \leqslant -1$, donc à $\alpha \leqslant 1$. Par exemple, si $\alpha := 1/2$, on trouve $y_i = (1+i\varepsilon)^2$ puis $h_i = 2\varepsilon + \dfrac{\varepsilon^2}{1+i\varepsilon}\cdot$ On en déduit sans peine que $t_i \sim 2i\varepsilon$ et donc que $y_i \sim t_i^2/4$, ce qui est raisonnable puisque la seule solution est ici $(1+t/2)^2$.

En revanche, si $\alpha > 1$, la série converge et la solution n'est définie qu'à distance finie. On retrouve le fait déjà constaté que les solutions non triviales de $y' = y^2$ « explosent ».

Nous allons maintenant *comparer* des solutions approchées de l'équation (33), en supposant de plus que *les conditions initiales sont elles mêmes approchées.* Pour cela, nous supposerons que la fonction $\Phi(t, Y)$ est *k-lipschitzienne en Y* sur le cylindre de sécurité C_0. Autrement dit, $k > 0$ et :

$$\forall (\overline{t}, \overline{Y}), (t, Y) \in C_0 \ , \ \left\|\Phi(\overline{t}, \overline{Y}) - \Phi(t, Y)\right\| \leqslant k \left\|\overline{Y} - Y\right\| .$$

L'existence d'un tel k est en particulier assurée si Φ admet sur C_0 des dérivées partielles continues en Y.

Théorème 50 (Lemme de Gronwall). Soient F_1 et F_2 des solutions approchées respectivement à ε_1 et à ε_2 près, définies sur un même intervalle J contenant t_0, et soit $\varepsilon := \varepsilon_1 + \varepsilon_2$. On suppose de plus que $\|F_1(t_0) - F_2(t_0)\| \leqslant \delta$, où $\delta > 0$. Alors :

$$\forall t \in J \ , \ \|F_1(t) - F_2(t)\| \leqslant \delta e^{k|t-t_0|} + \frac{\varepsilon}{k}\big(e^{k|t-t_0|} - 1\big). \tag{40}$$

Notons que le premier terme reflète le seul effet des conditions initiales différentes alors que le deuxième terme reflète l'effet de l'approximation différentielle.

Démonstration. Nous étudierons la fonction continue positive $g(t) := \|F_1(t) - F_2(t)\|$. On déduit de l'inégalité du triangle et de l'hypothèse sur les conditions initiales que $g(t) \leqslant \delta + \big\|\big(F_1(t) - F_2(t)\big) - \big(F_1(t_0) - F_2(t_0)\big)\big\|$. Pour la suite, nous supposerons $t \geqslant t_0$, le cas où $t \leqslant t_0$ se traitant exactement de la même manière. On invoquera l'égalité $F(b) - F(a) = \displaystyle\int_a^b F'(s)\,\mathrm{d}s$, qui a un sens et qui est vérifiée pour toute fonction continue de classe $\mathcal{C}^1$ par morceaux sur un intervalle $[a, b]$ (page 852), et l'inégalité de la moyenne :

$$\begin{aligned}
g(t) &\leqslant \delta + \int_{t_0}^t \left\|F_1'(s) - F_2'(s)\right\| \mathrm{d}s \\
&\leqslant \delta + \int_{t_0}^t \Big(\varepsilon + \big\|\Phi\big(s, F_1(s)\big) - \Phi\big(s, F_2(s)\big)\big\|\Big)\,\mathrm{d}s \\
&\leqslant \delta + \varepsilon(t - t_0) + \int_{t_0}^t k\,\|F_1(s) - F_2(s)\|\,\mathrm{d}s \\
&\leqslant \delta + \varepsilon(t - t_0) + k\int_{t_0}^t g(s)\,\mathrm{d}s.
\end{aligned}$$

Posons $G(t) := \displaystyle\int_{t_0}^t g(s)\,\mathrm{d}s$: on a prouvé que $g(t) = G'(t) \leqslant \delta + \varepsilon(t - t_0) + kG(t)$ (pour $t \geqslant t_0$). Cette inégalité se réécrit :

$$\big(e^{-k(t-t_0)}G(t)\big)' \leqslant e^{-k(t-t_0)}\big(\delta + \varepsilon(t - t_0)\big),$$

d'où, puisque $G(t_0) = 0$:

$$e^{-k(t-t_0)}G(t) \leqslant \int_{t_0}^{t} e^{-k(s-t_0)}\big(\delta + \varepsilon(s-t_0)\big)\,\mathrm{d}s$$
$$= \delta\frac{1-e^{-k(t-t_0)}}{k} + \varepsilon\frac{1-\big(1+k(t-t_0)\big)e^{-k(t-t_0)}}{k^2}\cdot$$

On en déduit d'abord que $G(t) \leqslant \delta\dfrac{e^{k(t-t_0)}-1}{k} + \varepsilon\dfrac{e^{k(t-t_0)}-\big(1+k(t-t_0)\big)}{k^2}$, puis que :

$$g(t) \leqslant \delta + \varepsilon(t-t_0) + kG(t) \leqslant \delta e^{k|t-t_0|} + \frac{\varepsilon}{k}\big(e^{k|t-t_0|}-1\big).$$

Corollaire 51. Soient F une solution exacte et G une solution approchée à ε près de l'équation (33), satisfaisant la même condition initiale $F(t_0) = G(t_0) = Y_0$ et toutes deux définies sur $[t_0 - T_0, t_0 + T_0]$. Alors, sur ce segment, $\|F - G\| \leqslant \dfrac{\varepsilon}{k}\big(e^{kT_0}-1\big)$.

En particulier, s'il existe une solution, la méthode d'Euler en donne une approximation uniforme sur tout segment.

3.2 Existence et comportement des solutions

3.2.1 Théorèmes d'existence

Nous allons énoncer sans preuve les théorèmes fondamentaux, qui justifieront en partie les exemples étudiés aux sections 3.2.3 et 3.3. Le théorème d'existence le plus général est dû à Cauchy, Peano et Arzela. Il suppose seulement la fonction Φ continue et affirme l'existence d'une solution locale au voisinage de toute condition initiale $(t_0, Y_0) \in U$.

Exercice 29.
Démontrer que l'intervalle de définition d'une solution maximale est ouvert (*théorème des bouts*).

Solution. Soit J l'intervalle de définition d'une solution maximale F. Soit $b \in \mathbb{R}$ une borne de J : par exemple, $J =]-\infty, b]$. Il existe une solution G avec condition initiale $G(b) = F(b)$ sur un intervalle $]b', b''[$, où $b' < b < b''$. La fonction obtenue par recollement de F et de la restriction de G à $]b, b''[$ est une extension stricte de F, ce qui contredit la maximalité.

Le défaut du théorème de Cauchy-Peano-Arzela est qu'il ne garantit aucune sorte d'unicité, ce qui le rend difficile à manier. Pour aller plus loin, il faut une hypothèse plus forte sur Φ. *Nous supposerons désormais que* Φ *est localement lipschitzienne sur* U. Cela équivaut à la propriété suivante : pour tout $(t_0, Y_0) \in U$, il existe un cylindre C centré en (t_0, Y_0) et un réel $k > 0$ tels que Φ est k-lipschitzienne sur C. Cette condition est en particulier vérifiée si $\Phi(t, Y)$ admet sur U des dérivées partielles continues en (t, Y).

Théorème 52 (Théorème de Cauchy-Lipschitz). Pour tout $(t_0, Y_0) \in U$ et pour tout cylindre de sécurité $C := [t_0 - T, t_0 + T] \times \overline{B}(Y_0, r)$ (voir la proposition 46 de la page 853), le problème de Cauchy (35) admet une unique solution F sur $J := [t_0 - T, t_0 + T]$.

Comme annoncé, nous ne prouverons pas le théorème de Cauchy-Lipschitz. L'unicité est facile à établir (exercice II.7.34 de la page 878.) Nous allons toutefois donner une idée des deux principales méthodes de démonstration de l'existence.

Un peu d'Histoire

La *méthode de Cauchy-Lipschitz* consiste à fixer une suite de réels strictement positifs $(\varepsilon_p)_{p\in\mathbb{N}}$ de limite 0 et à construire, pour tout p, une solution approchée F_p à ε_p près, par exemple par la méthode d'Euler. Il faut ensuite démontrer que la suite de fonctions $(F_p)_{p\in\mathbb{N}}$ admet une limite uniforme F ; puis en déduire (en passant par l'équation intégrale) que F est une solution de l'équation. Cette méthode a été inventée par Euler à la fin du XVIII$^{\text{e}}$ siècle pour obtenir des solutions approchées. En 1820, Cauchy l'a utilisée pour démontrer l'existence de solutions lorsque Φ est de classe $\mathcal{C}^1$. Lipschitz a retrouvé cette méthode en 1868, avec les conditions plus générales qui portent son nom.
La *méthode de Picard* consiste à considérer l'équation intégrale (36) comme une équation au point fixe, et à la résoudre par itération. C'est celle que nous avons utilisée dans le cas des équations linéaires, mais le cas général est bien plus délicat (voir l'exercice II.7.35 de la page 879, qui en donne une illustration). Cette méthode a été employée par Liouville en 1837 dans un cas particulier, puis généralisée et systématisée par Picard en 1890.

Les conséquences suivantes sont alors aisées à démontrer :

1. Si, pour tout p, F_p est une solution approchée à ε_p près (avec condition initiale exacte) sur J, et si $\lim\limits_{p\to+\infty} \varepsilon_p = 0$, alors les F_p convergent uniformément vers F sur J.
2. Deux solutions du problème de Cauchy en (t_0, Y_0) sur un même intervalle contenant t_0 coïncident.
3. Le problème de Cauchy en (t_0, Y_0) admet une unique solution maximale.

Exemples.

1. Soit $y_0 > 0$. La fonction $f(t) := \dfrac{y_0}{1 - y_0 t}$ est solution sur $]-\infty, 1/y_0[$ du problème de Cauchy $y' = y^2$, $y(0) = y_0$, et elle ne peut être prolongée. C'est donc l'unique solution maximale.
2. Soit de plus $\gamma > 0$. La fonction $g(t) := \gamma \tan \gamma(t + \theta)$, où $\theta := \dfrac{1}{\gamma} \arctan \dfrac{y_0}{\gamma}$, est solution sur $\left]\dfrac{-\pi}{2\gamma} - \theta, \dfrac{\pi}{2\gamma} - \theta\right[$ du problème de Cauchy $y' = y^2 + \gamma^2$, $y(0) = y_0$, et elle ne peut être prolongée. C'est donc l'unique solution maximale.

Exercice 30.
Déterminer la solution maximale du problème de Cauchy $y' = 2ty^2$, $y(0) = y_0$.
Solution. Si $y_0 = 0$, la fonction nulle est solution sur $\mathbb{R}$, donc non prolongeable. C'est donc l'unique solution maximale (et elle est globale).
Supposons $y_0 < 0$. La fonction $f(t) := \dfrac{y_0}{1 - y_0 t^2}$ est solution sur $\mathbb{R}$, donc non prolongeable. C'est donc l'unique solution maximale (et elle est globale).
Supposons $y_0 > 0$. La fonction $f(t) := \dfrac{y_0}{1 - y_0 t^2}$ est solution sur $\left]-y_0^{-1/2}, y_0^{-1/2}\right[$ et non prolongeable. C'est donc l'unique solution maximale.

Un problème mathématique et physique important est celui de la *dépendance des conditions initiales*. Le lemme de Gronwall suggère que, lorsque Y_0 varie peu, la solution du problème de Cauchy en (t_0, Y_0) sur un intervalle donné varie peu. On peut en déduire l'énoncé suivant, *que nous admettrons*.

Théorème 53 (Continuité des solutions en fonction des conditions initiales). On suppose que $U = I \times V$, où I est un intervalle ouvert de $\mathbb{R}$ et V un ouvert de $\mathbb{R}^n$, et que $\Phi(t, Y)$ est localement lipschitzienne en Y sur U. Notons $F(t, t_0, Y_0)$ la solution maximale du problème de Cauchy en (t_0, Y_0). Il existe un cylindre $J \times W \subset U$ centré en (t_0, Y_0) tel que la fonction $F(t, t_0, Y_0)$ est définie et uniformément continue sur $J \times J \times W$.

Exercice 31.
Calculer $f(t, t_0, y_0)$ et son domaine de définition dans le cas de l'équation $y' = y^2$.
Solution. La solution maximale du du problème de Cauchy $y' = y^2$, $y(t_0) = y_0$ est nulle sur $\mathbb{R}$ si $y_0 = 0$, et égale à $f(t) := \dfrac{y_0}{1 - y_0(t - t_0)}$ si $y_0 \neq 0$. Dans ce dernier cas, elle est définie sur celle des deux demi-droites $\left]-\infty, t_0 + \dfrac{1}{y_0}\right[$, $\left]t_0 + \dfrac{1}{y_0}, +\infty\right[$ qui contient 0. On a donc $f(t, t_0, y_0) = \dfrac{y_0}{1 - y_0(t - t_0)}$ dans tous les cas, fonction définie sur le complémentaire dans $\mathbb{R}^3$ de la surface $y_0(t - t_0) = 1$.

3.2.2 Explosion des solutions

Il s'agit ici de comprendre pourquoi et comment les solutions maximales des équations différentielles non linéaires ne sont pas globales. L'exemple caractéristique est celui de l'équation $y' = y^2$: au bord de son intervalle de définition, la solution $\dfrac{1}{c - t}$ diverge.

Théorème 54 (Lemme de sortie des compacts). Soit F une solution maximale de l'équation (33), définie sur l'intervalle $J :=]a, b[$. Supposons $b < +\infty$. Alors, pour tout compact $K \subset U$, il existe $b' \in J$ tel que la trajectoire de la solution restreinte à $]b', b[$ soit extérieure à K :

$$\forall t \in \left]b', b\right[\ , \ \big(t, F(t)\big) \in U \setminus K. \tag{41}$$

On dit qu'au voisinage de b, *la trajectoire sort de tout compact.* De même, si $a > -\infty$, alors, au voisinage de a, la trajectoire sort de tout compact.

Démonstration. Supposons qu'il existe $K \subset U$ compact tel que, pour tout $b' \in J$, la trajectoire de la solution restreinte à $]b', b[$ rencontre K. On a donc une suite $(t_n)_{n\in\mathbb{N}}$ dans J telle que $\lim\limits_{n\to+\infty} t_n = b$ et que, pour tout n, $\big(t_n, F(t_n)\big) \in K$. Quitte à remplacer la suite $(t_n)_{n\in\mathbb{N}}$ par une sous-suite, on trouve que la suite des $\big(t_n, F(t_n)\big)$ admet une limite $(b, Y_0) \in K \subset U$. Soit $C := [b-\varepsilon, b+\varepsilon] \times \overline{B}(Y_0, r)$ un cylindre de sécurité centré en (b, Y_0). Fixons n tel que $|t_n - b| < \varepsilon/3$ et $\|F(t_n) - Y_0\| < r/3$. Le cylindre $C' := [t_n - 2\varepsilon/3, t_n + 2\varepsilon/3] \times \overline{B}\big(F(t_n), 2r/3\big)$ est inclus dans C. C'est un cylindre de sécurité en $\big(t_n, F(t_n)\big)$. D'après le théorème 52 de la page 859, notre équation possède une unique solution G définie sur $J' :=]t_n - 2\varepsilon/3, t_n + 2\varepsilon/3[$ et telle que $G(b) = Y_0$. Les solutions F et G coïncident donc sur $J' \cap J$. En les recollant, on peut prolonger F en une solution sur $J \cup J'$, ce qui contredit la maximalité. ∎

Géométriquement, lorsque t s'approche de la borne finie b, soit la trajectoire diverge vers l'infini (« explosion »), soit elle adhère à la frontière de U.

Corollaire 55. Supposons $U = I \times \mathbb{R}^n$. Si la solution maximale F définie sur $]a, b[$ reste bornée lorsque t tend vers b (resp. lorsque t tend vers a), alors $b = +\infty$ (resp. $a = -\infty$).

Exemple. Considérons le système $\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} = \begin{pmatrix}y - x^3\\ -x\end{pmatrix}$.

Pour toute solution maximale $F(t) := \begin{pmatrix}f(t)\\ g(t)\end{pmatrix}$ définie sur $]a, b[$, l'énergie $E := f^2 + g^2$ vérifie $\dot{E} = 2f\dot{f} + 2g\dot{g} = -f^4 \leqslant 0$. Elle est donc décroissante. Ainsi, F reste bornée lorsque t tend vers b, et $b = +\infty$.

On peut parfois *quantifier* l'explosion par comparaison avec des équations dont la solution est connue. Nous allons le voir dans le cas d'équations scalaires d'ordre 1. Soient $\varphi(t, y)$ et $\psi(t, y)$ deux fonctions continues et localement lipschitziennes en y sur l'ouvert $U \subset \mathbb{R}^2$.

Théorème 56. Soient f, g les solutions maximales respectives des problèmes de Cauchy en $(t_0, y_0) \in U$ associés aux équations différentielles $y' = \varphi(t, y)$ et $y' = \psi(t, y)$. Soit $]a, b[$ leur intervalle commun de définition.
(i) Supposons que $\varphi(t) < \psi(t)$ pour tout $t \in]a, b[$. Alors $f > g$ sur $]a, t_0[$ et $f < g$ sur $]t_0, b[$.
(ii) Supposons seulement que $\varphi \leqslant \psi$. Alors $f \geqslant g$ sur $]a, t_0[$ et $f \leqslant g$ sur $]t_0, b[$.

Démonstration. (i) Nous le démontrerons du côté droit. Puisque $f(t_0) = y_0 = g(t_0)$ et que $f'(t_0) = \varphi(t_0, y_0) < \psi(t_0, y_0) = g'(t_0)$, on a $f < g$ sur un intervalle non vide $]t_0, t_0'[$. Supposons qu'il existe $t \in]t_0', b[$ tel que $f(t) \geqslant g(t)$. Soit t_1 le plus petit tel t : donc $f(t_1) = g(t_1)$ et $f < g$ sur l'intervalle non vide $]t_0, t_1[$. On a donc $f'(t_1) \geqslant g'(t_1)$. Mais d'autre part, $f'(t_1) = \varphi\big(t_1, f(t_1)\big) < \psi\big(t_1, f(t_1)\big) = g'(t_1)$, contradiction.

(ii) Encore une fois, considérons le côté droit. Soient $\psi_n := \psi + \dfrac{1}{n}$ (pour $n \in \mathbb{N}^*$ et g_n la solution maximale du problème de Cauchy associé en (t_0, y_0). D'après (i), $f < g_n$ à droite de t_0 sur leur domaine commun d'existence. Mais g_n tend uniformément vers g d'après les « conséquences aisées » du théorème de Cauchy-Lipschitz (ou, directement, d'après le lemme de Gronwall). ∎

La première partie du théorème reste valable sans l'hypothèse que φ et ψ sont localement lipschitziennes, mais il faut alors considérer « des » solutions quelconques f, g (voir également l'exercice II.7.39 de la page 879).

Exemple. Prenons $\varphi(t, y) := y^2$, $\psi(t, y) := y^2 + 1$, $t_0 := 0$ et $y_0 := 1$. Les solutions maximales respectives sont $f(t) := \dfrac{1}{1-t}$ sur $]-\infty, 1[$ et $g(t) := \tan(t + \pi/4)$, sur $\left]\dfrac{-3\pi}{4}, \dfrac{\pi}{4}\right[$. Comme $f < g$ à droite de 0, on voit que g explose à gauche de $t = 1$, donc que $\pi/4 \leqslant 1$, ce que l'on savait déjà, bien entendu ! On a donc $\dfrac{1}{1-t} < \tan(t + \pi/4)$ sur $\left]0, \dfrac{\pi}{4}\right[$.

3.2.3 Équations de Riccati

On appelle *équation de Riccati* une équation de la forme $y' = ay^2 + by + c$, où a, b, c sont continues sur un intervalle ouvert I de $\mathbb{R}$. La fonction $\varphi(t, y) := ay^2 + by + c$ étant continue et localement lipschitzienne en y sur $U := I \times \mathbb{R}$, tous les théorèmes généraux lui sont applicables. Si de plus a est de classe $\mathcal{C}^1$, et ne s'annule pas, le changement de fonction inconnue $z := ay$ nous ramène à l'équation $z' = z^2 + pz + q$, où $p := b + a'/a$ et $q := ac$ sont continues.

Exercice 32.
Peut-on se ramener à une équation de la forme $x' = x^2 + r$?
Solution. Si p est de classe $\mathcal{C}^1$, le changement de fonction inconnue $x := z + p/2$ nous ramène à l'équation $x' = x^2 + r$ avec $r := q - p^2/4 - p'/2$ continue.

Lien avec les équations linéaires du second ordre.
Soit u une solution de l'équation $y' + y^2 + py + q = 0$ sur un intervalle J. Les fonctions f telles que $f' = uf$ sur J forment un $\mathbb{R}$-espace vectoriel de dimension 1. Toute telle fonction vérifie :

$$f'' = uf' + u'f = (u^2 + u')f = (-pu - q)f = -pf' - qf, \text{ c'est-à-dire } f'' + pf' + qf = 0.$$

Réciproquement, si $f''+pf'+qf = 0$, sur tout domaine où f ne s'annule pas, la fonction $u := f'/f$ est une solution de l'équation $y' + y^2 + py + q = 0$. On a donc une application surjective $f \mapsto f'/f$ de l'ensemble des solutions non triviales de l'équation linéaire $y'' + py' + qy = 0$ sur un certain ensemble de solutions de l'équation $y' + y^2 + py + q = 0$. (Nous ne chercherons pas à être plus précis en ce qui concerne les *domaines de définition* de ces solutions.) De plus, deux fonctions f et g ont même image si, et seulement si, elles sont proportionnelles.
Notant E le $\mathbb{R}$-espace vectoriel des solutions de $y'' + py' + qy = 0$, on voit que l'on a identifié l'ensemble S des solutions de l'équation $y'+y^2+py+q = 0$ à l'espace projectif $\mathrm{Pr}(E)$ (définition 3 de la page 270).
Soit $t_0 \in I$. On sait que l'application « conditions initiales » $f \mapsto \big(f(t_0), f'(t_0)\big)$ réalise un isomorphisme de E sur $\mathbb{R}^2$. Cette application est linéaire et passe donc au quotient en une bijection de $\mathrm{Pr}(E)$ sur $\mathbf{P}^1(\mathbb{R})$. Modulo l'identification ci-dessus, on voit que l'application $u \mapsto u(t_0)$ réalise une bijection de S sur $\mathbf{P}^1(\mathbb{R})$. C'est pourquoi, dans l'étude de l'équation de Riccati, il arrive que l'on prenne comme espace des conditions initiales une droite projective.

Nous allons étudier plus particulièrement l'*équation de Liouville* $y' = y^2 + t$. Cette équation est particulièrement célèbre parce que c'est la première pour laquelle il ait été rigoureusement démontré (par Liouville, justement) qu'*elle ne peut être résolue par quadratures (résolution d'équations linéaires d'ordre* 1*) à partir de fonctions élémentaires*. L'affirmation un peu vague qui précède reçoit un sens très précis (et une démonstration) dans ce que l'on appelle aujourd'hui *théorie de Galois différentielle*.

Exercice 33.

(i) Comment ramener les unes aux autres les quatre équations $y' = \pm y^2 \pm t$?
(ii) Quel lien y a-t-il entre les trajectoires des solutions des équations $y' = y^2 + t$ et $y' = y^2 - t$?
Solution. (i) Par le changement de fonction inconnue $z(t) := \pm z(\pm t)$.
(ii) Le changement de fonction inconnue qui opère ici est $z(t) := -y(-t)$. Les trajectoires d'une des deux familles s'obtiennent donc à partir de celles de l'autre famille par une symétrie centrale par rapport à l'origine $(t, y) \mapsto (-t, -y)$.

Commençons par une approche « qualitative », et même graphique. Nous utiliserons l'équation $y' = y^2 - t$. Ses trajectoires sont des courbes dans le plan des phases élargi $\mathbb{R}^2$, de coordonnées t, y. Nous appellerons *isocline de pente* 0 la parabole P d'équation $y^2 - t = 0$: toute trajectoire qui passe par un point de P y admet en effet une tangente horizontale (en un tel point, $y' = 0$). Le complémentaire de P dans $\mathbb{R}^2$ admet deux composantes connexes : $P^+ := \{(t, y) \in \mathbb{R}^2 \mid y^2 - t > 0\}$ et $P^- := \{(t, y) \in \mathbb{R}^2 \mid y^2 - t < 0\}$.

Exercice 34.

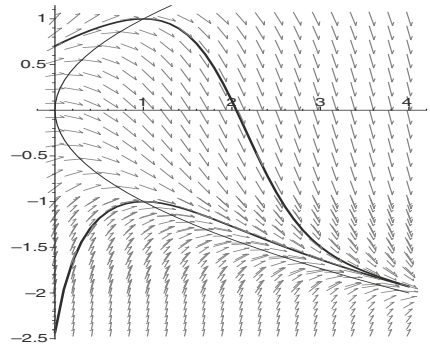
Vérifier que ce sont bien les composantes connexes.

Solution. Il n'existe aucun chemin reliant P^+ à P^- dans $\mathbb{R}^2 \setminus P$, car la fonction continue $(t, y) \mapsto y^2 - t$ y changerait de signe sans s'annuler.
Le domaine P^- est convexe (calcul facile) donc connexe par arcs.
Tout point (t, y) de P^+ peut être relié à $(-1, y)$ et ce dernier à $(-1, 0)$ par des segments contenus dans P^+, qui est donc connexe par arcs.

Ainsi, les portions des trajectoires comprises dans P^+ (resp. dans P^-) y sont croissantes (resp. décroissantes). Nous allons démontrer que *toute trajectoire qui rencontre P entre dans P^- et y reste confinée* et qu'*elle y a une durée de vie infinie*.

Pour faire dessiner la figure ci-contre à Maple, nous avons traité l'équation $y' = y^2 - t$ comme une *équation autonome d'ordre* 2 :

$$\frac{d}{dt}\begin{pmatrix} y \\ t \end{pmatrix} = \begin{pmatrix} y^2 - t \\ 1 \end{pmatrix}.$$



Soit $y = f(t)$ une solution maximale dont le graphe (trajectoire) passe par le point $(t_0, y_0) \in P$: donc $t_0 = y_0^2$ et $f'(t_0) = 0$. En t_0, la fonction $f^2(t) - t$ a pour dérivée -1, donc passe du signe $+$ au signe $-$: juste avant t_0, la trajectoire est dans P^+, juste après, elle est dans P^-. Comme toute rencontre de la trajectoire avec P est une entrée dans P^-, il n'y a pas de sortie et la trajectoire reste confinée. (Voir l'exercice II.7.40 de la page 879.)

Soit $]\alpha, \beta[$ l'intervalle de définition de f. Sur $]t_0, \beta[$, on a $f^2(t) < t$ d'après ce qui précède. Si $\beta < +\infty$, on en déduit que $\big(t, f(t)\big)$ reste borné lorsque t tend vers β, ce qui contredit le corollaire 55 de la page 861.

Exercice 35.

Démontrer que, si $y_0 > 0$, la trajectoire finit par couper l'axe $y = 0$ (étudier le signe de f'').

Solution. On a $f'' = 2ff' - 1$. Dans P^-, $f' < 0$, donc, tant que $f > 0$, $f'' < -1$. On en tire d'abord $f' < -(t - t_0)$ puis $f < y_0 - (t - t_0)^2/2$, ce qui n'est pas compatible avec $f > 0$ quand t tend vers $+\infty$.

Pour aborder l'étude quantitative, nous utiliserons plutôt l'équation $y' = y^2 + t$. Nous étudierons les trajectoires passant par les points $(t_0, 0)$ où $t_0 > 0$. Soit $]\alpha, \beta[$ l'intervalle de définition de la solution maximale f correspondant à cette condition initiale.

Nous allons d'abord montrer que la trajectoire admet une asymptote verticale en $+\infty$, et encadrer β. Soit $\gamma > 0$ tel que $\gamma < \sqrt{t_0}$. Sur $\big]\gamma^2, \beta\big[$, on a $y^2 + \gamma^2 < y^2 + t$. D'après le théorème 56 de la page 862, sur $]t_0, \beta[$, on a $f > \gamma \tan \gamma(t - t_0)$. La solution f explose donc au plus tard en $t_0 + \dfrac{\pi}{2\gamma}$, et l'on a donc $\beta < t_0 + \dfrac{\pi}{2\gamma}$.

Comme c'est vrai pour tout $\gamma < \sqrt{t_0}$, on en déduit la majoration $\beta \leqslant \beta_0 := t_0 + \dfrac{\pi}{2\sqrt{t_0}}$.

Sur $]\alpha, \beta_0[$, on a $y^2 + t < y^2 + \beta_0$. Du théorème 56 de la page 862, on tire que, sur $]t_0, \beta[$, $f < \sqrt{\beta_0} \tan \sqrt{\beta_0}(t - t_0)$. Il ne peut donc y avoir explosion à gauche de $\beta_1 := t_0 + \dfrac{\pi}{2\sqrt{\beta_0}}$. D'après le lemme de sortie des compacts, on a donc $\beta_1 \leqslant \beta$.

Pour étudier α, on choisit d'abord $\gamma > \sqrt{t_0}$ tel que $\gamma < \sqrt{\beta}$. On a, sur $\left]\alpha, \gamma^2\right[$, $y^2 + t < y^2 + \gamma^2$, donc, sur $]\alpha, t_0[$, $f > \gamma \tan \gamma(t - t_0)$. Il n'y a donc pas d'explosion à droite de $t_0 - \dfrac{\pi}{2\gamma}$, et l'on a donc $\alpha < t_0 - \dfrac{\pi}{2\gamma}$. Comme c'est vrai pour tout $\gamma > \sqrt{t_0}$, on en déduit la minoration $\alpha \leqslant \alpha_1 := t_0 - \dfrac{\pi}{2\sqrt{t_0}}$.

Il est plus difficile de minorer α ; en fait, selon la valeur de t_0, il y a ou non explosion à distance finie (voir là dessus l'exercice II.7.41 de la page 880).

3.3 Systèmes différentiels autonomes

3.3.1 Théorie

Lorsque la fonction $\Phi(t, Y)$ ne dépend pas de t, le système est dit *autonome*. Nous avons déjà rencontré les exemples des systèmes linéaires à coefficients constants, des équations scalaires $y' = y^2$ et $y' = 2\sqrt{|y|}$ et du système de l'exemple de la page 861. Pour tout le reste de ce chapitre, nous considérons donc une fonction continue $\Phi(Y)$ sur un ouvert $V \subset \mathbb{R}^n$. Les trajectoires des solutions sont contenues dans l'ouvert $U := \mathbb{R} \times V$ (espace des phases élargi), mais il sera plus intéressant d'étudier les courbes intégrales, qui sont contenues dans l'espace des phases V.

> ***Remarque***. On peut toujours ramener l'étude d'un système $Z' = \Psi(t, Z)$, où Ψ est définie sur l'espace des phases élargi U, à celle d'un système autonome $Y' = \Phi(Y)$ dont l'espace des phases est U : il suffit pour cela de poser $Y := (t, Z)$ et $\Phi(Y) = \big(1, \Psi(t, Z)\big) = \big(1, \Psi(Y)\big)$ (et donc d'*augmenter la dimension*), et il est clair que les deux systèmes sont équivalents.

Nous supposerons toujours que Φ est localement lipschitzienne. Cela sera en particulier vrai si Φ est de classe $\mathcal{C}^1$, ce qui sera le cas dans tous les exemples que nous traiterons. Nous pouvons donc appliquer les résultats généraux de la section 3.1. Pour tout $Y_0 \in V$, soit $t \mapsto F(t, Y_0)$ l'unique solution maximale du système autonome $Y' = \Phi(Y)$ dont la trajectoire passe par $(0, Y_0) \in U$. Pour tout $t_0 \in \mathbb{R}$, la fonction $t \mapsto F(t - t_0, Y_0)$ est une solution de $Y' = \Phi(Y)$ (l'argument est le même que dans la preuve du théorème 35 de la page 837), dont la trajectoire passe par (t_0, Y_0) et qui ne peut être étendue : c'est donc la solution maximale correspondante. On retrouve en grande partie les propriétés des équations linéaires autonomes :

1. Les courbes intégrales forment une partition de l'espace des phases V (le *portrait de phases*). Deux solutions dont la courbe intégrale passe par un même Y_0 aux temps t_0 et t_1 sont décalées dans le temps de $t_1 - t_0$ l'une par rapport à l'autre et leurs intervalles de définition sont translatés l'un de l'autre par $t_1 - t_0$.
2. L'application $(t, Y) \mapsto F(t, Y)$ est une action *partielle* de $\mathbb{R}$ sur V : si $Y_1 := F(t, Y_0)$ est défini et si $Y_2 := F(u, Y_1)$ est défini, alors $F(t + u, Y_0)$ est défini et égal à Y_2.

3. Le *flot* $(t, Y) \mapsto F(t, Y)$ est une application continue. Cela découle du théorème 53 de la page 860.

Dans le cas d'un système d'ordre 2, l'application $\Phi : V \to \mathbb{R}^2$ peut être considérée comme un *champ de vecteurs plan* (page 838). La tangente en $Y_0 \in V$ à la courbe intégrale passant par un point Y_0 a pour direction $\Phi(Y_0)$. Naturellement, les champs de vecteurs peuvent être définis en toutes dimensions, mais leur représentation est plus difficile.

Théorème et définition 57. On dit que $Y_0 \in V$ est un *point critique* du système autonome $Y' = \Phi(Y)$ si $\Phi(Y_0) = 0$. Pour que Y_0 soit un point critique, il faut, et il suffit, que $\{Y_0\}$ soit une courbe intégrale.

Démonstration. Si $\Phi(Y_0) = 0$, la fonction constante $F(t) := Y_0$ est solution du système $Y' = \Phi(Y)$; cette solution est globale, c'est donc la solution maximale pour toute condition initiale (t_0, Y_0). La courbe intégrale passant par Y_0 est donc $\{Y_0\}$.
Si $\{Y_0\}$ est une courbe intégrale, on a $F(t, Y_0) = Y_0$ pour tout t dans un voisinage ouvert de 0 ; la dérivée en 0 de cette fonction est donc 0, et c'est $\Phi(Y_0)$. ∎

Exemple. L'équation autonome $y' = y^2$ admet pour seul point critique 0. Si $y_0 \neq 0$, on a $F(t, y_0) = \dfrac{y_0}{1 - ty_0}$, qui est défini sur la demi-droite $]-\infty, 1/y_0[$ si $y_0 > 0$ et sur la demi-droite $]1/y_0, +\infty,[$ si $y_0 < 0$. Le flot correspond ici à l'action par homographies du sous groupe de $GL_2(\mathbb{R})$ formé des matrices de la forme $\begin{pmatrix} 1 & 0 \\ -t & 1 \end{pmatrix}$. L'équation étant de Riccati, il est naturel de considérer son espace des phases comme étant la droite projective $\mathbf{P}^1(\mathbb{R})$ et d'y prolonger le flot, puisque ces homographies agissent aussi sur $\infty \in \mathbf{P}^1(\mathbb{R})$; il y a alors deux orbites (ou courbes intégrales) qui sont $\{0\}$ et l'ouvert affine $\mathbf{P}^1(\mathbb{R}) \setminus \{0\}$.
Le changement de fonction inconnue $z := 1/y$ nous ramène au système autonome $z' = 1$. Le point critique 0 est envoyé à l'infini et la courbe intégrale « régulière » est envoyée sur $\mathbb{R}$. le flot est alors l'action de $\mathbb{R}$ sur lui-même par translation.

La notion d'*intégrale première* de la définition 10 de la page 842 s'étend telle quelle ici, ainsi que le raisonnement de la page 843 sur le lien avec le portrait de phases.

Exercice 36.

Déterminer les points critiques du système autonome $\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} y - x^3 \\ -x \end{pmatrix}$ de l'exemple de la page 861. On rappelle que les solutions ont une vie future infinie. À l'aide d'un passage en coordonnées polaires, décrire les portions correspondantes des courbes intégrales. En déduire que ce système n'admet pas d'intégrale première continue non triviale sur $\mathbb{R}^2$.

Solution. Les points critiques sont ceux tels que $y - x^3 = -x = 0$, il n'y a donc que $(0, 0)$. Sur $\mathbb{R}^2 \setminus \{(0, 0)\}$, on peut écrire $x = \rho \cos\theta, y = \rho \sin\theta$ avec $(\rho, \theta) \in \mathbb{R}_+^* \times \mathbb{R}$ et l'on trouve que le système de départ équivaut à :

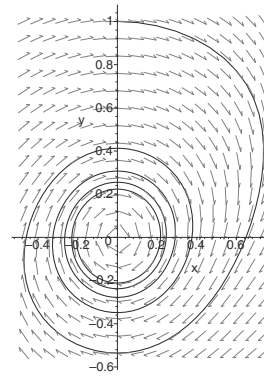
$$\begin{cases} \dot{\rho} = -\rho^3 \cos^4\theta, \\ \dot{\theta} = -1 + \rho^2 \cos^3\theta \sin\theta. \end{cases}$$

On retrouve le fait que ρ décroît (donc est borné, d'où le futur infini). Comme $\rho > 0$, cette fonction admet une limite ρ_0 quand $t \to +\infty$.

D'autre part, $\ddot{\rho} = -3\rho^2\dot{\rho}\cos^4\theta + 4\rho^3\cos^3\theta(\sin\theta)\dot{\theta}$ est borné. D'après l'exercice II.7.43 de la page 880, $\dot{\rho}$ tend donc vers 0. On en déduit d'abord que $\rho\cos\theta$ tend vers 0, donc que $\dot{\theta}$ tend vers -1, donc que θ tend vers $-\infty$; puis que $\rho_0 = 0$, c'est-à-dire que ρ tend vers 0.

Les portions futures des courbes intégrales du système de départ sont donc des spirales qui s'enroulent autour de l'origine (qui est donc un *foyer attractif*), parcourues dans le sens inverse du sens trigonométrique. Puisqu'elles adhèrent à l'origine, toute intégrale première continue a même valeur partout qu'en $(0,0)$, donc est constante.

Nous montrons ci-contre la courbe intégrale passant par $(0,1)$. On remarquera qu'il n'est nullement évident, numériquement ou graphiquement, que l'origine est bien un foyer.



3.3.2 Systèmes non dissipatifs à un degré de liberté

L'équation $\ddot{X} = F(t, X)$ apparaît en mécanique Newtonienne, où elle régit le mouvement d'un point matériel soumis à un champ de forces $F(t, X)$ dépendant du temps t et de la position X. Nous supposerons que le champ ne dépend pas du temps (le système est donc autonome) et nous restreindrons au cas où la position dépend d'une seule coordonnée $x \in \mathbb{R}$ (l'abscisse dans le cas d'un ressort, l'angle dans le cas d'un pendule). On peut donc écrire $\ddot{x} = f(x) = -U'(x)$, où U est une primitive arbitraire de $-f$, appelée *potentiel* (le potentiel est donc défini à une constante près). Nous supposerons de plus que $U(x)$ est une fonction de classe $\mathcal{C}^2$ sur un ouvert $V \subset \mathbb{R}$. Par vectorialisation, on obtient le système autonome :

$$\begin{cases} \dot{x} = y, \\ \dot{y} = -U'(x). \end{cases} \tag{42}$$

La fonction $\Phi(x, y) := \big(y, -U'(x)\big)$ est de classe $\mathcal{C}^1$ sur l'espace des phases $V \times \mathbb{R}$. Notons que les coordonnées x, y sur l'espace des phases représentent respectivement la position et la vitesse du point matériel.

Exercice 37.

Déterminer et interpréter physiquement les points critiques de ce système.

Solution. Il s'agit des couples $(x_0, 0)$ tels que $U'(x_0) = 0$. Un tel point représente l'état d'un point matériel de position x_0 où le champ de forces s'annule et de vitesse 0. D'après le théorème de Cauchy-Lipschitz, la solution constante correspondante est la seule possible : le point est stationnaire, c'est une position d'équilibre.

Bien que très facile, le théorème suivant est très utile.

Théorème 58. La fonction $E(x,y) := U(x) + \dfrac{1}{2}y^2$ est une intégrale première.

Démonstration. Si $\big(u(t), v(t)\big)$ est une solution, la dérivée de $E\big(u(t), v(t)\big)$ est :

$$U'\big(u(t)\big)u'(t) + v(t)v'(t) = U'\big(u(t)\big)v(t) - v(t)U'\big(u(t)\big) = 0.$$

Il s'agit évidemment de la loi de conservation de l'énergie et, pour cette raison, on dit que le système est *non dissipatif*.

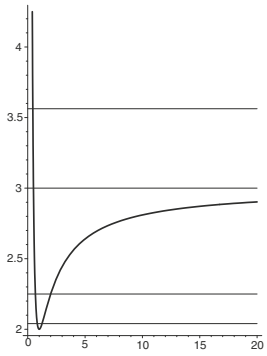
Comme déjà remarqué, l'existence d'une intégrale première permet d'intégrer le système par quadratures. Sur un intervalle ouvert où $\dot{x} > 0$ (par exemple), on a $\dot{x} = \sqrt{2\big(E_0 - U(x)\big)}$, où E_0 est une constante fixée par les conditions initiales. Puisque x est de classe $\mathcal{C}^1$ et strictement croissante, on peut l'inverser et écrire $t = \tau(x)$, où :

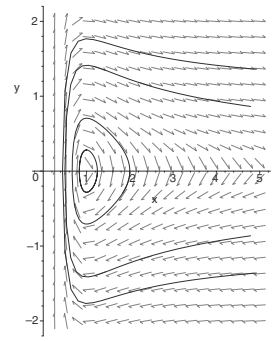
$$\tau'(x) = \frac{1}{\sqrt{2\big(E_0 - U(x)\big)}}\cdot \tag{43}$$

Exemple. Prenons $U(x) := \dfrac{\omega^2}{2}x^2$ ($\omega > 0$). L'équation correspondante est $\ddot{x} + \omega^2 x = 0$ (oscillateur harmonique). Pour des conditions initiales $x(0) = x_0$, $\dot{x}(0) = 0$ (position x_0, vitesse nulle), on a $E_0 = \dfrac{1}{2}\omega^2 x_0^2$ et l'on trouve : $\tau'(x) = \dfrac{1}{\omega\sqrt{x_0^2 - x^2}}\cdot$

Potentiel et portrait de phase

Puisque les courbes intégrales sont incluses dans les courbes de niveau de l'intégrale première $E(x,y)$, c'est-à-dire dans les courbes d'équation $y = \pm\sqrt{2\big(E_0 - U(x)\big)}$, le graphe de la fonction U donne une idée du portrait de phases.





Nous allons justifier en partie l'usage heuristique de ces dessins. Soient $a < x_0 < b$ tels que $U' < 0$ sur $[a, x_0[$ et $U' > 0$ sur $]x_0, b]$. Ainsi, $(x_0, 0)$ est un point critique d'énergie $E_0 := U(x_0)$. Soit E un niveau d'énergie proche de E_0 : $E_0 < E < \min\big(U(a), U(b)\big)$.

Il y a donc un unique couple (a_E, b_E) de points tels que $a < a_E < x_0 < b_E < b$ tels que $U(a_E) = U(b_E) = E$.

La fonction $x \mapsto \sqrt{2\big(E - U(x)\big)}$ croit de 0 à $\sqrt{2\big(E - E_0\big)}$ sur $[a_E, x_0]$ et décroît de $\sqrt{2\big(E - E_0\big)}$ à 0 sur $[x_0, b_E]$, avec tangentes verticales aux deux extrémités. La courbe $y = \pm\sqrt{2\big(E_0 - U(x)\big)}$, avec $x \in [a_E, b_E]$, a donc l'allure d'une ellipse. En particulier, elle est compacte. Nous allons voir que c'est une courbe intégrale.

Prenons les conditions initiales $x(t_0) = a_E$, $y(t_0) = 0$ et notons (u, v) la solution maximale correspondante. On a $\dot{y}(t_0) = -U'(a_E) > 0$, et le mouvement part donc vers le haut, sur le graphe décrit ci-dessus. Tant que $y = v(t) > 0$, on a $\dot{x} > 0$ et l'on va vers la droite. Le mouvement ne peut s'arrêter avant d'atteindre $(b_E, 0)$ (lemme de sortie des compacts). Celui-ci est atteint en un temps t_1 tel que :

$$t_1 - t_0 = \int_{a_E}^{b_E} \tau'(x)\,\mathrm{d}x = \int_{a_E}^{b_E} \frac{\mathrm{d}x}{\sqrt{2\big(E - U(x)\big)}}\cdot$$

On montre de même que le point $(x, y) = (u(t), v(t))$ repart de $(b_E, 0)$ vers $(a_E, 0)$, qu'il atteint en un temps t_2 tel que :

$$t_2 - t_1 = \int_{b_E}^{a_E} \tau'(x)\,\mathrm{d}x = \int_{b_E}^{a_E} \frac{-\mathrm{d}x}{\sqrt{2\big(E - U(x)\big)}}\cdot$$

En effet, sur ce parcours, $\dot{x} < 0$ et la formule pour $\tau(x)$ est l'opposé de la précédente.

Théorème 59. La solution maximale d'énergie E est définie sur $\mathbb{R}$. Elle est périodique de période :

$$T_E := 2\int_{a_E}^{b_E} \frac{\mathrm{d}x}{\sqrt{2\big(E - U(x)\big)}}\cdot \tag{44}$$

Démonstration. On a ici $t_2 - t_1 = t_1 - t_0$ et le T_E donné est leur somme. La périodicité vient de ce que la solution de condition initiale $(a_E, 0)$ en $t_2 = t_0 + T_E$ s'obtient par décalage temporel à partir de la solution trouvée. ∎

Exercice 38.

Faire le calcul dans le cas de l'oscillateur harmonique.

Solution. Prenons $x_0 := 0$ (le résultat ne dépend pas de ce choix). Alors $a_E := -\sqrt{\dfrac{2E}{\omega}}$, $b_E := \sqrt{\dfrac{2E}{\omega}}$ et : $T_E = 2\displaystyle\int_{a_E}^{b_E} \frac{\mathrm{d}x}{\sqrt{2E - \omega^2 x}} = \frac{2\pi}{\omega}\cdot$

Dans ce cas, la période ne dépend pas du niveau d'énergie (voir également l'exercice II.7.44).

Pendule non linéaire non amorti

L'équation du pendule non amorti est $\ddot{\theta} = -\sin\theta$. Elle correspond au potentiel $U(\theta) = 1 - \cos\theta$ (par exemple), qui est compris entre 0 ($\theta \in 2\pi\mathbb{Z}$, équilibres stables) et 2 ($\theta \in \pi + 2\pi\mathbb{Z}$, équilibres instables).

De l'étude précédente, on déduit le portrait de phases. Nous écrirons la condition initiale sous la forme $\theta(0) = 0$, $\dot{\theta}(0) = \alpha$ avec $\alpha \neq 0$. Nous prendrons $\alpha > 0$ (le cas $\alpha < 0$ s'en déduit en inversant le sens de parcours des trajectoires). L'énergie est donc $E = \dfrac{1}{2}\alpha^2$ et la loi de conservation (intégrale première) s'écrit :

$$\frac{\dot{\theta}^2}{2} + 1 - \cos\theta = \frac{1}{2}\alpha^2.$$

Si $\alpha < 2$, on peut appliquer l'étude précédente et le théorème 59 de la page précédente. Notons $\theta_1 := \arccos(1 - \alpha^2/2) \in]0, \pi[$. L'angle θ oscille périodiquement entre $-\theta_1$ et θ_1 avec la période :

$$T_E = 2\int_{-\theta_1}^{\theta_1} \frac{\mathrm{d}\theta}{\sqrt{2(\cos\theta - \cos\theta_1)}}.$$

Résumons brièvement les autres cas. Si $\alpha = 2$, l'énergie est 1. Il y a un plus grand intervalle $]t_1, t_2[$ contenant 0 sur lequel $\dot{\theta} > 0$, pendant lequel θ parcourt l'intervalle $]\theta_1, \theta_2[$ en croissant. Aux bornes, $\dot{\theta}$ s'annule (sinon, on pourrait élargir l'intervalle) et l'on a donc $\theta_1 = -\pi$, $\theta_2 = \pi$ et :

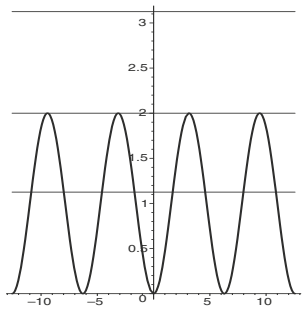
$$t_2 = \int_0^{\pi} \frac{\mathrm{d}\theta}{\sqrt{2(\cos\theta + 1)}} \quad \text{et} \quad t_1 = \int_0^{-\pi} \frac{\mathrm{d}\theta}{\sqrt{2(\cos\theta + 1)}},$$

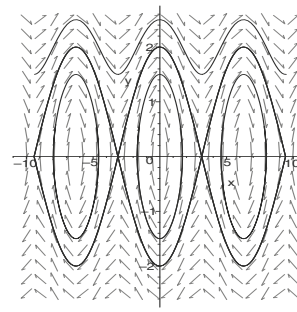
donc $t_1 = -\infty$ et $t_2 = +\infty$ (calcul laissé au lecteur). Le pendule parcourt donc un cercle complet, sauf le point haut, qui n'est qu'asymptotiquement approché aux temps $\pm\infty$.

Si $\alpha > 2$, on a $E > 2$ qui ne peut être atteint par $U(\theta)$, donc $\dot{\theta}$ ne s'annule jamais, donc reste strictement positif. Le mouvement est défini pour tout $t \in \mathbb{R}$ et tout $\theta \in \mathbb{R}$ et l'on a une bijection :

$$t = \tau(\theta) = \int_0^{\theta} \frac{\mathrm{d}\varphi}{\sqrt{\alpha^2 - 2(1 - \cos\varphi)}}.$$

Le pendule tourne indéfiniment.





3.3.3 Lois de Kepler

Le mouvement d'une planète M dans le système solaire est, en première approximation, déterminé par l'attraction solaire. L'accélération est donc dirigée vers le centre O (le soleil) et, en valeur absolue, proportionnelle à $\frac{1}{r^2}$, où $r := \left\|\overrightarrow{OM}\right\|$ (norme euclidienne) est la distance de la planète au soleil. Nous noterons en lettres grasses les vecteurs : $\mathbf{r} := \overrightarrow{OM}$ le vecteur indiquant la position, $\dot{\mathbf{r}} := \frac{d\mathbf{r}}{dt}$ le vecteur vitesse et $\ddot{\mathbf{r}} := \frac{d^2\mathbf{r}}{dt^2} = \frac{d\dot{\mathbf{r}}}{dt}$ l'accélération. La loi de Newton se traduit donc par l'équation :

$$\ddot{\mathbf{r}} = -\frac{c}{r^3}\mathbf{r}, \tag{45}$$

où $c > 0$ est déterminée par la masse du soleil et la constante d'accélération de la pesanteur. Après choix d'un repère orthonormé centré en O, on trouve le système différentiel autonome :

$$\begin{cases} \ddot{x} = -\dfrac{cx}{r^3}, \\ \ddot{y} = -\dfrac{cy}{r^3}, \\ \ddot{z} = -\dfrac{cz}{r^3}, \end{cases} \qquad \text{où} \quad r := \sqrt{x^2+y^2+z^2}. \tag{46}$$

Ce système autonome est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^3 \setminus \{O\}$, c'est-à-dire tant que la planète n'est pas tombée dans le soleil, ce que nous supposerons. C'est un système à trois degrés de liberté, mais nous allons réduire ce nombre par des principes généraux issus de la mécanique.

Accélération centrale implique mouvement plan. On dérive le produit vectoriel $\mathbf{r} \wedge \dot{\mathbf{r}}$, ce qui donne :

$$\frac{d}{dt}\mathbf{r} \wedge \dot{\mathbf{r}} = \dot{\mathbf{r}} \wedge \dot{\mathbf{r}} + \mathbf{r} \wedge \ddot{\mathbf{r}} = \vec{0},$$

puisque $\mathbf{r}$ et $\ddot{\mathbf{r}}$ sont colinéaires (champ central). Ainsi, le vecteur $\overrightarrow{W} := \mathbf{r} \wedge \dot{\mathbf{r}}$ est constant. Nous laissons au lecteur le soin de démontrer que, si $\overrightarrow{W} = \vec{0}$, le mouvement est rectiligne. Si $\overrightarrow{W} \neq \vec{0}$, le point M reste dans le plan affine orthogonal à $\overrightarrow{W}$ et passant par O. On a donc un mouvement plan, et, si l'on choisit bien le repère orthonormé, $z = 0$. On est ramené à un système à deux degrés de liberté :

$$\begin{cases} \ddot{x} = -\dfrac{cx}{r^3}, \\ \ddot{y} = -\dfrac{cy}{r^3}, \end{cases} \qquad \text{où } r := \sqrt{x^2+y^2}. \tag{47}$$

Pour un mouvement central plan, on a une loi de conservation de plus. La fonction $A(t) := x\dot{y} - y\dot{x}$ a pour dérivée $A'(t) = x\ddot{y} - y\ddot{x} = 0$, donc est une constante J. Cette loi admet deux interprétations. Avec les notations $x = r\cos\theta$, $y = r\sin\theta$ (coordonnées polaires), on a $A(t) = r^2\dot{\theta}$. Les physiciens appellent $mA(t)$ (où m est la masse de la planète) *moment cinétique* et nous avons donc prouvé la conservation du moment cinétique dans un

champ central. Les mathématiciens appellent $\frac{1}{2}A(t)$ *vitesse aréolaire*. En effet, lorsque M se déplace dans le plan de la position $M_1 := M(t_1)$ à la position $M_2 := M(t_2)$, l'aire du triangle balayé par $\mathbf{r}$ est :

$$\int_{M_1}^{M_2} \frac{1}{2} r^2 \, \mathrm{d}\theta = \int_{t_1}^{t_2} \frac{1}{2} A(t) \, \mathrm{d}t.$$

Nous avons prouvé que cette intégrale est proportionnelle à $t_2 - t_1$: *l'aire balayée par le rayon vecteur est proportionnelle au temps*, c'est la *deuxième loi de Kepler.*

Champ radial implique conservation de l'énergie. On dit que le champ est radial parce que l'on peut écrire $\ddot{\mathbf{r}} = \dfrac{f(r)}{r}\,\mathbf{r}$. On en tire :

$$\frac{\mathrm{d}}{\mathrm{d}t}\dot{\mathbf{r}}^2 = 2\dot{\mathbf{r}}\ddot{\mathbf{r}} = 2\frac{f(r)}{r}\mathbf{r}\dot{\mathbf{r}} = \frac{f(r)}{r}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{r}^2 = 2f(r)\dot{r},$$

puisque $\mathbf{r}^2 = r^2$. Notons $U(r)$ une primitive de $-f(r)$ (*potentiel*). Dans notre cas, $f(r) = -c/r^2$ et $U(r) = -c/r$ (par exemple). On obtient la loi de conservation de l'énergie :

$$\frac{1}{2}\dot{\mathbf{r}}^2 + U(r) = E,$$

où E est constante.

On se ramène à un degré de liberté. En coordonnées polaires, $\dot{\mathbf{r}}^2 = \dot{r}^2 + r^2\dot{\theta}^2 = \dot{r}^2 + \dfrac{J^2}{r^2}$, d'après la loi de conservation du moment cinétique. La dérivée de cette expression est $2\dot{r}\ddot{r} - 2\dfrac{J^2}{r^3}\dot{r}$, et elle vaut également $2\dot{\mathbf{r}}\ddot{\mathbf{r}} = 2f(r)\dot{r}$ d'après le calcul précédent. On en déduit :

$$\ddot{r} = f(r) + \frac{J^2}{r^3}, \tag{48}$$

un système autonome à un degré de liberté. Le potentiel associé à ce système est $\overline{U}(r) = U(r) + \dfrac{J^2}{2r^2}\cdot$ La loi de conservation de l'énergie prend alors la forme :

$$\frac{\dot{r}^2}{2} + \overline{U}(r) = E,$$

où E est constante. On aurait également obtenir cette loi à partir de la précédente en y remplaçant directement $\dot{\mathbf{r}}^2$ par $\dot{r}^2 + r^2\dot{\theta}^2$ et $\dot{\theta}$ par $\dfrac{J}{r^2}\cdot$ Cette loi permet distinguer les orbites compactes (planètes, astéroïdes, comètes périodiques) des orbites infinies (comètes exceptionnelles, etc). Dans le cas où $J = 0$, donc où θ est constant, le mouvement est alors complètement déterminé (exercice II.7.47 de la page 880).

Détermination des orbites par élimination du temps. Si $J \neq 0$, la loi de conservation $\dot{\theta} = \dfrac{J}{r^2}$ entraîne que $\dot{\theta}$ est de signe constant, donc que la trajectoire peut être paramétrée

par θ. Nous noterons $r' := \dfrac{dr}{d\theta}\cdot$ On a donc $\dot r = r'\dot\theta = \dfrac{Jr'}{r^2}$ (dérivée d'une fonction composée) et la deuxième version de la loi de conservation de l'énergie fournit l'équation différentielle :

$$r'^2 = \frac{2E}{J^2}r^4 + \frac{2c}{J^2}r^3 - r^2.$$

En posant $u := 1/r$, on la ramène à :

$$u'^2 = \frac{2E}{J^2} + \frac{2c}{J^2}u - u^2 = \frac{2EJ^2 + c^2}{J^4} - \left(u - \frac{c}{J^2}\right)^2.$$

Si $2EJ^2 + c^2 < 0$, il n'y a aucune solution. Si $2EJ^2 + c^2 = 0$, la seule solution est $u = \dfrac{c}{J^2}$, donc $r = \dfrac{J^2}{c}$: l'orbite est circulaire, et la relation $\dot\theta = \dfrac{J}{r^2} = \dfrac{c^2}{J^3}$ montre que la vitesse angulaire est constante. Si $2EJ^2 + c^2 > 0$, en posant $u - \dfrac{c}{J^2} = \sqrt{\dfrac{2EJ^2 + c^2}{J^4}}v$, on se ramène à $v'^2 = 1 - v^2$. Quitte à changer θ en $\pm(\theta - \theta_0)$, c'est-à-dire à choisir le sens de rotation et l'origine de la mesure de l'angle, on a donc $v = \cos\theta$. On trouve en fin de compte l'équation polaire :

$$r = \frac{p}{1 + e\cos\theta},$$

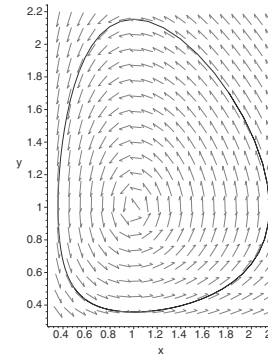
où $p := \dfrac{J^2}{c}$ et $e := \sqrt{1 + \dfrac{2EJ^2}{c^2}}\cdot$ C'est l'équation focale d'une conique d'excentricité e (module de géométrie affine de [L1]) : on a prouvé la *première loi de Kepler*.

Si $-\dfrac{c^2}{2J} < E < 0$, on a $0 < e < 1$ et l'orbite est une ellipse. Si $E = 0$, on a $e = 1$ et l'orbite est une parabole. Si $E > 0$, on a $e > 1$ et l'orbite est une hyperbole.

3.3.4 Exemples de systèmes admettant un cycle limite

Un *cycle* est une courbe intégrale fermée parcourue périodiquement. Un *cycle limite* est un cycle autour duquel viennent s'enrouler des courbes intégrales qui spiralent. Nous ne donnerons ni définition rigoureuse ni théorèmes, mais des exemples classiques.

Lotka-Volterra



Le système suivant, appelé *équation de Lotka-Volterra*, modélise la croissance (et la décroissance) de deux populations animales dont l'une est la proie et l'autre le prédateur :

$$\begin{cases} \dot x = x - xy, \\ \dot y = -y + xy. \end{cases}$$

Ses points critiques sont $(0,0)$ et $(1,1)$. Si (x,y) est une solution, $\dfrac{\mathrm{d}}{\mathrm{d}t}xye^{-(x+y)}(t) = 0$ (calcul facile) et $xye^{-(x+y)}$ est donc une intégrale première. Le signe de xy est donc constant sur toute courbe intégrale, et la région $x, y > 0$ est stable par le flot. Dans cette région, les courbes intégrales sont des courbes fermées parcourues périodiquement (exercice II.7.49 de la page 880).

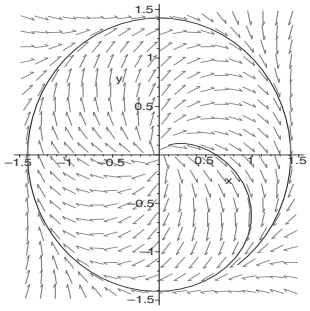
La figure représente la courbe de conditions initiales $(1/2, 1/2)$.

Le système suivant est considéré comme une modélisation plus précise du système proie-prédateur :

$$\begin{cases} \dot{x} = \lambda x + y - x(x^2+y^2), \\ \dot{y} = -x + \lambda y - y(x^2+y^2), \end{cases} \quad \text{où } \lambda > 0.$$

Son seul point critique est l'origine. On se place sur le complémentaire de l'origine dans le plan des phases. Le système s'écrit alors en coordonnées polaires :

$$\begin{cases} \dot{\rho} = -\rho^3 + \lambda\rho, \\ \dot{\theta} = -1, \end{cases} \quad \text{où } \lambda > 0.$$



Le cercle $\rho = \sqrt{\lambda}$ est donc une courbe intégrale, parcourue à vitesse angulaire constante. On peut déterminer les autres courbes, intégrales et l'on prouve (exercice II.7.50 de la page 880) que toutes s'enroulent autour du cercle $\rho = \sqrt{\lambda}$, qui est donc un cycle limite.
La figure représente la courbe de conditions initiales $(1/10, 1/10)$.

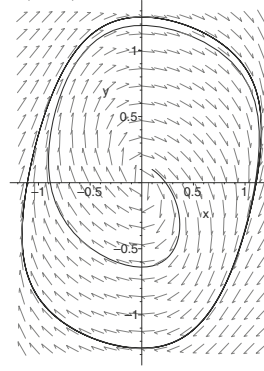
L'équation de Van der Pol

Le système suivant modélise le comportement de certains dispositifs électroniques (*équation de Van der Pol*).

$$\begin{cases} \dot{x} = y - (x^3 - x), \\ \dot{y} = -x. \end{cases}$$

Son unique point critique est l'origine. Nous allons effleurer l'étude de ses autres courbes intégrales.
Soit (x, y) une solution maximale définie sur l'intervalle $]\alpha, \beta[$.

Notant $u := x^2 + y^2$, on a $\dot{u} = 2x\dot{x} + 2y\dot{y} = 2x^2 - 2x^4 \leqslant 2u$, d'où $\dfrac{d}{dt}\big(u(t)e^{-2t}\big) \leqslant 0$. Si β était fini, $u(t)e^{-2t}$ serait bornée au voisinage de β, donc également u, donc (f, g). Ainsi, $\beta = +\infty$: les solutions maximales ont un futur infini.
Notons $D := \{(x, y) \in \mathbb{R}^2 \mid x > 0, y > 0, y > x^3 - x\}$.
Soit (f, g) la solution maximale correspondant à la condition initiale $(x, y)(0) = (0, y_0)$, où $y_0 > 0$. La trajectoire entre dans D juste après le temps 0 puisqu'elle part vers la droite. De plus, la trajectoire sort de D : en effet, tant qu'elle est dans D, elle reste confinée dans le compact $D \cap \{(x, y) \in \mathbb{R}^2 \mid y \leqslant y_0\}$. Après avoir deviné par où la trajectoire sort de D, le lecteur poursuivra cette étude par l'exercice II.7.51 de la page 880. La conclusion est celle-ci : il existe un cycle limite autour duquel les autres courbes s'enroulent.
La figure représente la courbe de conditions initiales $(1/10, 1/10)$.



Exercice type corrigé

II.7.0 (i) Soient r et s deux fonctions continues sur l'intervalle ouvert I telles que $r \leqslant s$. Soient f une solution non triviale de l'équation (E_r) $x'' + rx = 0$ et g une solution non triviale de l'équation (E_s) $x'' + sx = 0$. Soient $t_1 < t_2$ deux zéros consécutifs de f (exercice 15 de la page 831). Démontrer que, soit g s'annule en un point de $]t_1, t_2[$, soit g y est proportionnelle à f. (Étudier le comportement du wronskien de f et g sur cet intervalle.)
(ii) Montrer que les zéros de deux solutions indépendantes de (E_r) sont entrelacés. (Appliquer la question (i) avec $r = s$.) En déduire que le nombre $N_{a,b}(f)$ des zéros d'une telle solution sur $[a, b] \subset I$ dépend faiblement de la solution f de (E_r).
(iii) On suppose que $r \leqslant 0$. Montrer qu'une solution non triviale de (E_r) admet au plus un zéro. (Appliquer la question (i) avec $s := 0$.)
(iv) On suppose que $r \leqslant \mu^2$, où $\mu \in \mathbb{R}_+^*$. Soient $t_1 < t_2$ deux zéros consécutifs d'une solution f non triviale de (E_r). Démontrer que $t_2 - t_1 \geqslant \dfrac{\pi}{\mu}$. (Appliquer la question (i) avec $g := \sin(\mu t + \varphi)$.)
(v) On suppose que $r \geqslant \lambda^2$, où $\lambda \in \mathbb{R}_+$. Démontrer que, sur tout intervalle $]a, b[\subset I$ de longueur $\dfrac{\pi}{\lambda}$, toute solution de (E_r) s'annule au moins une fois.

Solution. (i) Supposons que g ne s'annule pas sur $]t_1, t_2[$, par exemple $g > 0$; d'après l'exercice 3 de la page 806, il s'agit de démontrer que le wronskien $w := fg' - f'g$ de f et de g est nul sur $]t_1, t_2[$. Puisque f est nulle en t_1 et t_2 et que l'on peut la supposer positive sur $]t_1, t_2[$, on a $f'(t_1) \geqslant 0$ et $f'(t_2) \leqslant 0$. On en déduit que $w(t_1) = -f'(t_1)g(t_1) \leqslant 0$ et $w(t_2) = -f'(t_2)g(t_2) \geqslant 0$.
D'autre part, $w' = fg'' + f'g' - f'g' - f''g = fg'' - f''g = (r - s)fg \leqslant 0$ sur $]t_1, t_2[$, et la fonction continue w y est donc décroissante (au sens large) : comme $w(t_1) \leqslant 0 \leqslant w(t_2)$, elle est nulle.
(ii) Si l'on prend $s := r$, la question (i) peut être appliquée symétriquement à f et à g : on en déduit qu'entre deux zéros consécutifs de f il y a un zéro de g et réciproquement. Sur tout segment $[a, b] \subset I$, le nombre de zéros de f et celui de g sont finis (exercice 15 de la page 831). Comme ces zéros sont entrelacés, $N_{a,b}(f) - N_{a,b}(g)$ vaut 0 ou ± 1.
(iii) Si f avait deux zéros consécutifs $t_1 < t_2$, toute solution de $x'' = 0$, c'est-à-dire toute fonction affine $\alpha t + \beta$, s'annulerait entre les deux : c'est évidemment impossible. (Si une solution affine g était proportionnelle à f, celle-ci serait affine non triviale et n'aurait qu'un zéro.)
(iv) Soient $t_1 < t_2$ deux zéros consécutifs de f. La solution $\sin \mu(t - t_1)$ de $x'' + \mu^2 x$ est proportionnelle à f $\left(\text{et dans ce cas, } t_2 = t_1 + \dfrac{\pi}{\mu}\right)$ ou bien s'annule sur $]t_1, t_2[$ (et dans ce cas, $t_2 > t_1 + \dfrac{\pi}{\mu}$).
(v) Toujours par application de la question (i), on voit que, pour tout a, f doit être proportionnelle à $\sin \lambda(t - a)$ ou bien s'annuler sur $\left]a, a + \dfrac{\pi}{\lambda}\right[$.

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.7.1 Soit $x(t)$ une solution de classe $\mathcal{C}^1$ de l'équation différentielle $x^2 + \dot{x}^2 = E_0$. Démontrer que, sur tout intervalle ouvert où $\dot{x}$ ne s'annule pas, $x(t)$ est de la forme $\varepsilon\sqrt{E_0}\sin(t - t_0)$, où $\varepsilon := \pm 1$.

II.7.2 Montrer que les solutions obtenues à l'exercice précédent peuvent être « recollées » à l'aide de fonctions constantes $\pm\sqrt{E_0}$. Y a-t-il unicité de la solution à conditions initiales données ?

II.7.3 Trouver des fonctions f et g de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$ qui ne sont pas liées et dont le wronskien est nul.

II.7.4 ∗∗ Soient $f_1, \dots, f_n$ des fonctions de classe $\mathcal{C}^n$ sur l'intervalle ouvert I. On suppose que leur wronskien est identiquement nul. Soit $J \subset I$ un intervalle ouvert non vide. Démontrer que J contient un intervalle ouvert non vide J' tel que les restrictions de $f_1, \dots, f_n$ à J' soient liées.

II.7.5 Reprendre l'exercice 5 de la page 810 en utilisant comme système fondamental de solutions $(e^{\mathrm{i}t}, e^{-\mathrm{i}t})$

II.7.6 Reprendre l'exercice 7 de la page 816 lorsque $\omega = \pm 1$.

II.7.7 ∗ Si $k \in \mathbb{R}_+^*$, $l \in \mathbb{R}$, l'équation $y' - ky = l$ admet une et une seule solution bornée.

II.7.8 ∗ Si $\lim\limits_{+\infty}(f + f') = 0$, alors $\lim\limits_{+\infty} f = 0$.

II.7.9 ∗ Dans l'exercice 10 de la page 821, préciser si tous les cas possibles ont été examinés.

II.7.10 Résoudre l'exercice 11 de la page 824 par la méthode de la section 1.2.1 et par celle de la section 1.2.3.

II.7.11 Résoudre l'exercice 6 de la page 814 par la méthode de la section 1.3.3 et par celle de la section 1.2.3.

Exercices sur la section 2

II.7.12 Soit $A(t) := \begin{pmatrix} 0 & t \\ 0 & i \end{pmatrix}$. A-t-on $R(t,0) = \exp\left(\displaystyle\int_0^t A(s)\,\mathrm{d}s\right)$?

II.7.13 ∗ On suppose que toutes les valeurs de $A(t)$ commutent entre elles.

Vérifier que $R(t,t_0) = \exp\left(\displaystyle\int_{t_0}^t A(s)\,\mathrm{d}s\right)$.

II.7.14 $*$ Soit B une fonction matricielle de classe $\mathcal{C}^1$ sur l'intervalle ouvert I à valeurs dans $M_n(\mathbb{C})$. Montrer que, si toutes les valeurs de B commutent entre elles, alors $BB' = B'B$ (identiquement sur I). Que dire de la réciproque ?

II.7.15 $**$ (i) Soit B une fonction matricielle de classe $\mathcal{C}^1$ sur l'intervalle ouvert I à valeurs dans $M_n(\mathbb{C})$. On suppose que $B = PDP^{-1}$, où $P \in GL_n(\mathbb{C})$ est constante et où D est diagonale (mais pas nécessairement constante). Montrer que toutes les valeurs de B commutent entre elles, et que donc $BB' = B'B$.

(ii) On considère la matrice de fonctions $\begin{pmatrix} t^2 & t^3 & t^4 \\ -2t & -2t^2 & -2t^3 \\ 1 & t & t^2 \end{pmatrix}$ de classe $\mathcal{C}^1$ sur l'intervalle ouvert $I = \mathbb{R}$. Calculer B^2, BB' et $B'B$ et vérifier que $B(0)B(t) \neq B(t)B(0)$ lorsque t est assez grand. Qu'en déduire à propos de la réciproque des questions précédentes ?

(iii) On suppose que $BB' = B'B$ et que, pour tout t, les valeurs propres de B sont simples. On a alors vu (exercice II.5.9 de la page 758) que ce sont des fonctions $\lambda_1(t), \ldots, \lambda_n(t)$ de classe $\mathcal{C}^1$ sur I. Montrer que, pour tout $t_0 \in I$, il y a, au voisinage de t_0, une base de vecteurs propres $U_1(t), \ldots, U_n(t)$ de classe $\mathcal{C}^1$.

(iv) Montrer que, pour tout i, $B'U_i$ est un multiple de U_i, puis que U_i' est un multiple de U_i. En déduire qu'il existe, au voisinage de t_0, une base de vecteurs propres constants.

(v) Montrer que $B = PDP^{-1}$, où $P \in GL_n(\mathbb{C})$ est constante et où D est diagonale (mais pas nécessairement constante).

II.7.16 $**$ On suppose que l'intervalle ouvert I n'est pas majoré et que la fonction matricielle $A(t)$ est absolument intégrable sur I. Montrer que $R(t, t_0)$ admet une limite lorsque t tend vers $+\infty$ et que cette limite est un élément de $GL_n(K)$.

II.7.17 $*$ Trouver toutes les intégrales premières de classe $\mathcal{C}^1$ de l'équation $\ddot{x} - x = 0$.

II.7.18 Résoudre l'équation $-u''(x)+u(x) = 1$ avec conditions aux limites $u(0) = u(1) = 0$.

II.7.19 L'intensité du courant dans un circuit RLC alimenté par une force électromotrice $V(t) := a\cos(\omega t + \varphi)$, vérifie l'équation différentielle : $L\ddot{I} + R\dot{I} + CI = V$. Reprendre dans ce cas l'exemple de la page 839.

II.7.20 Reprendre l'exemple de la page 839 lorsque $R < 0$ (« amplificateur opérationnel »). Y a-t-il des intégrales premières ?

II.7.21 $*$ Démontrer que l'équation à coefficients constants (E_0) est stable si, et seulement si, toutes les racines du polynôme $P := X^n + a_1X^{n-1} + \cdots + a_n \in K[X]$ (section 1.2) sont soit de partie réelle strictement négative, soit de partie réelle nulle et de multiplicité 1. Démontrer que (E_0) est asymptotiquement stable si, et seulement si, toutes les racines de P sont de partie réelle strictement négative.

II.7.22 Soient p une fonction de classe $\mathcal{C}^1$ et q une fonction continue sur I. Montrer que l'équation $x'' + px' + qx = 0$ peut se ramener à une équation de la forme $y'' + ry = 0$ par un changement de fonction inconnue $x = uy$.

II.7.23 $*$ Avec les notations de l'exercice II.7.0 de la page 875, question (ii), montrer qu'il existe un entier k tel que, pour toute solution f non triviale de (E_r), on ait $N_{a,b}(f) \in [\![k, k+1]\!]$.

II.7.24 Préciser comme suit la question (iv) de l'exercice II.7.0 de la page 875 : si r n'est pas la fonction μ^2, alors $t_2 - t_1 > \dfrac{\pi}{\mu}$.

II.7.25 Si $f > 0$ sur $[a, a+\pi]$ et si f est de classe $\mathcal{C}^2$, montrer qu'il existe $\xi \in]a, a+\pi[$ tel que $(f + f'')(\xi) > 0$.

II.7.26 $*$ Soit q une fonction de classe $\mathcal{C}^1$ et telle que $q, q' > 0$ sur $I :=]a, +\infty[$. Montrer que toute solution sur I de $x'' + qx = 0$ est bornée lorsque $t \to +\infty$.

II.7.27 Résoudre l'équation $x'' + \dfrac{5}{4t^2}x = 0$ sur $\mathbb{R}_+^*$. Montrer que les zéros de toute solution non triviale f s'accumulent en 0.

II.7.28 $*$ Si aucune valeur propre de $A \in M_n(\mathbb{C})$ n'est dans $2\mathrm{i}\pi\mathbb{Z}$ et si B est 1-périodique, alors le système $Y' = AY + B$ admet exactement une solution 1-périodique.

II.7.29 Exprimer la condition de stabilité d'un système à coefficients périodiques en termes de ses nombres caractéristiques.

II.7.30 $*$ Soient A et B deux matrices de période T et soit $C_\varepsilon(t_0$ l'application de Floquet du système de matrice $A + \varepsilon B$. Montrer que, lorsque $\varepsilon \to 0$, on a : $\|C_\varepsilon(t_0) - C(t_0)\| = O(\varepsilon)$.

Exercices sur la section 3

II.7.31 Démontrer rigoureusement que la solution de l'exercice 26 de la page 851 mentionne bien toutes les solutions maximales.

II.7.32 $*$ Soient A, B, g des fonctions continues positives sur un intervalle J telles que :

$$\forall t \in J\ ,\ g(t) \leqslant A(t) + \int_{t_0}^{t} B(s)g(s)\,\mathrm{d}s,$$

où $t_0 \in J$. Démontrer l'inégalité :

$$\forall t \in J\ ,\ g(t) \leqslant A(t) + \int_{t_0}^{t} B(s)A(s)\Big(\exp\int_{s}^{t} B(u)\,\mathrm{d}u\Big)\,\mathrm{d}s.$$

Comment appliquer ce résultat à la preuve du lemme de Gronwall ?

II.7.33 $*$ À l'aide du lemme de Zorn, déduire du théorème de Cauchy-Peano-Arzela l'existence d'une solution maximale.

II.7.34 Démontrer l'unicité dans le théorème 52 de la page 859.

II.7.35 $**$ On va employer la méthode de Picard pour résoudre le problème de Cauchy $y' = y^2$, $y(0) = 1$. On pose $f_0 := 0$ et, par récurrence : $f_{i+1}(t) := 1 + \displaystyle\int_0^t f_i^2(s)\,\mathrm{d}s$.

Démontrer par récurrence que, sur $[0, 1[$, la suite $(f_n)_{n\in\mathbb{N}}$ est croissante et majorée par $\dfrac{1}{1-t}$. Elle admet donc une limite g. Démontrer que g est intégrable et vérifie l'équation :

$$\forall t \in [0, 1[\ ,\ g(t) := 1 + \int_0^t g^2(s)\,\mathrm{d}s.$$

Conclure.

II.7.36 Calculer $f(t, t_0, y_0)$ (théorème 53 de la page 860) et son domaine de définition dans le cas de l'équation $y' = 2ty^2$.

II.7.37 $*$ Déterminer toutes les solutions maximales de l'équation $y' = |y - t|$.

II.7.38 $*$ Soit $\varphi(t, y)$ une fonction bornée de classe $\mathcal{C}^1$ sur $\mathbb{R}^2$. Démontrer que toute solution maximale de l'équation $y' = \varphi(t, y)$ est globale.

II.7.39 $**$ (i) Cet exercice propose une variante du critère de comparaison du théorème 56 de la page 862. Soient $\varphi(t, y)$ une fonction continue sur l'ouvert U de $\mathbb{R}^2$ et g une fonction dérivable sur un intervalle I et telle que : $\forall t \in I\ ,\ g'(t) > \varphi\big(t, g(t)\big)$ (*barrière supérieure stricte*). Soient f une solution de l'équation $y' = \varphi(t, y)$ sur l'intervalle I et $t_0 \in I$ tel que $f(t_0) \leqslant g(t_0)$. Démontrer que, pour $t \in I$, $t > t_0$, $f(t) < g(t)$. Formuler et démontrer un résultat analogue valable à gauche de t_0. Étudier le cas d'une *barrière inférieure stricte* (*i.e.* une fonction h dérivable sur un intervalle I et telle que : $\forall t \in I\ ,\ h'(t) < \varphi\big(t, h(t)\big)$).

(ii) On suppose maintenant que φ est lipschitzienne en y. Étendre les résultats ci-dessus aux barrières non strictes : $\forall t \in I\ ,\ g'(t) \geqslant \varphi\big(t, g(t)\big)$ et $\forall t \in I\ ,\ h'(t) \leqslant \varphi\big(t, h(t)\big)$. En déduire le théorème 56 de la page 862 comme cas particulier.

II.7.40 $**$ (i) Rédiger *rigoureusement et de manière détaillée* le raisonnement de la section 3.2.3 de la page 862 selon lequel une trajectoire qui rencontre P reste confinée dans P^-.

(ii) Prolonger l'étude qualitative géométrique de la section 3.2.3 comme suit. Soit f une une solution maximale de l'équation $y' = y^2 + t$ dont la trajectoire passe par le point (t_0, y_0) où $y_0 < 0$ et $y_0^2 + t_0 = 0$, et soit $]\alpha, \beta[$ son intervalle de définition. Vérifier que $\alpha = -\infty$.

(iii) Démontrer qu'à droite de t_0, la trajectoire reste dans la zone $y^2 + t > 0$. Démontrer que la trajectoire finit par rencontrer l'axe $t = 0$ en un point $(0, y_1)$ tel que $y_0 < y_1 < 0$.

(iv) Démontrer que la suite de la trajectoire finit par rencontrer l'axe $y = 0$ en un point $(t_1, 0)$ tel que $0 < t_1 < -1/y_0$.

(v) Montrer que l'application $u : y_0 \mapsto t_1$ de $\mathbb{R}_-^*$ dans $\mathbb{R}_+^*$ est strictement croissante.

(vi) Montrer qu'elle est continue et la prolonger par continuité en 0. Montrer qu'il existe un réel $a > 0$ tel que, si $0 < t_0 < a$, la solution maximale de l'équation $y' = y^2 + t$ dont la trajectoire passe par le point $(t_0, 0)$ est définie jusqu'en $-\infty$.

II.7.41 ∗∗ Soit f une solution maximale de l'équation $y' = y^2 + t$ dont la trajectoire passe par le point $(t_0, 0)$ où $t_0 > 0$ et soit $]\alpha, \beta[$ son intervalle de définition.
(i) Déterminer la valeur minimum d de $\gamma^2 + \dfrac{\pi}{2\gamma}$ pour $\gamma > 0$ et la valeur c de γ où ce minimum est atteint. Nous supposons maintenant que $t_0 > d$. Vérifier qu'il existe un unique $\gamma_0 > c$ tel que $t_0 = \gamma_0^2 + \dfrac{\pi}{2\gamma_0}$. Soit γ tel que $c < \gamma < \gamma_0$. Vérifier que $\gamma^2 < t_0 - \dfrac{\pi}{2\gamma}$. Montrer que, si l'on avait $\gamma^2 \geqslant \alpha$, on aurait $f < \gamma \tan \gamma(t - t_0)$ sur $\left]\gamma^2, t_0\right[$, d'où explosion en $t_0 - \dfrac{\pi}{2\gamma}$. En déduire que $\gamma^2 < \alpha$, puis que $\alpha \geqslant \gamma_0^2 = t_0 - \dfrac{\pi}{2\gamma_0}$.
(ii) Déduire de l'exercice II.7.40 de la page précédente que, pour $0 < t_0 < a$, on a $\alpha = -\infty$.

II.7.42 ∗ Soit φ une fonction *quelconque* (*i.e.* pas nécessairement continue) sur un intervalle J de $\mathbb{R}$. Alors toute solution sur un intervalle I et à valeurs dans J de l'équation autonome $\dot{x} = \varphi(x)$ est monotone sur I.

II.7.43 ∗ Soit ρ une fonction positive décroissante de classe $\mathcal{C}^2$ sur $[a, +\infty[$ et telle que $\ddot{\rho}$ est borné. Démontrer que $\lim\limits_{t \to +\infty} \dot{\rho}(t) = 0$.

II.7.44 ∗∗ Avec les notations du théorème 59 de la page 869, on suppose que $\alpha := U''(x_0) > 0$. Démontrer la formule suivante, donnant la *période des petites oscillations* :

$$\lim_{E \to E_0} T_E = \frac{2\pi}{\sqrt{\alpha}}.$$

II.7.45 ∗ Étudier le portrait de phases du pendule amorti $\ddot{\theta} = -C\dot{\theta} - \sin\theta$, où $C > 0$. Y a-t-il des intégrales premières continues non triviales ?

II.7.46 ∗∗ Comparer le pendule linéaire et le pendule non linéaire à l'aide du lemme de Gronwall.

II.7.47 ∗ Achever l'étude du mouvement des planètes lorsque $J := 0$.

II.7.48 ∗∗ Démontrer la *troisième loi de Kepler* : pour les orbites elliptiques, les carrés des périodes de rotation sont proportionnels aux cubes des demi-grands axes.

II.7.49 ∗∗ Montrer que les courbes intégrales du système de Lotka-Volterra (page 873) dans la région $x, y > 0$ sont fermées et parcourues périodiquement.

II.7.50 ∗ Déterminer les courbes intégrales du système proie-prédateur amélioré (page 874) et prouver que le cercle $\rho = \sqrt{\lambda}$ est un cycle limite.

II.7.51 ∗∗ (i) Poursuivre l'étude de l'équation de Van der Pol (page 874) en montrant que la trajectoire passe successivement par les quatre quadrants et arrive en un point $(0, y_1)$ tel que $y_1 > 0$. On définit ainsi une application $\sigma : y_0 \mapsto y_1$ de $\mathbb{R}_+^*$ dans lui-même. Montrer que σ est croissante.
(ii) Étudier, selon le signe de $\sigma(y_0) - y_0$, les variations de la suite des $y_n := \sigma^n(y_0)$ et en déduire l'allure de la courbe intégrale selon le cas.

Méthodes numériques

II.8

Le calcul numérique n'est pas né avec les ordinateurs, ni même avec la calculatrice mécanique de Pascal. Mettant à profit la puissance du calcul infinitésimal, des mathématiciens ont réalisé à la main des prouesses qui laissent pantois. Newton a inventé la méthode de résolution approchée d'équations qui porte son nom. Euler, à lui seul, a initié et mis en œuvre des méthodes efficaces pour évaluer numériquement des fonctions et des intégrales, et pour approcher les solutions d'équations différentielles.

Est-il nécessaire de s'initier aux méthodes numériques à une époque où des calculettes de 50 grammes font tout ? Il en va comme du reste de l'algorithmique : d'abord, on ne peut intelligemment utiliser un outil que si l'on sait quelque chose des principes qu'il met en jeu ; ensuite, de nouveaux algorithmes numériques sont nécessaires, qui ne peuvent être mis au point que si l'on est familier des algorithmes élémentaires ; enfin, les nombres réels sont intéressants et même leur manipulation effective est une voie d'approche vers l'analyse (voir également l'exercice 22 de la page 932).

L'application professionnelle de telles méthodes demande cependant une technicité qui n'est pas du ressort d'un ouvrage généraliste. Nous ne proposons ici qu'une première approche motivée et illustrée par des exemples (obtenus à l'aide de Maple). Cette approche est suivie dans [MPA1] de compléments sur les équations différentielles (au-delà de la méthode d'Euler) et sur l'analyse matricielle (recherche d'éléments propres et optimisation). Le lecteur désireux de préciser ces points trouvera son bonheur dans [Knuth] pour ce qui concerne la section 1 (voir aussi l'ouvrage plus concret [Muller]) ; dans le livre [Demailly] pour ce qui concerne les sections 2, 3 et 4 ; et dans le livre [Ciarlet] pour ce qui concerne la section 5.

1 Calcul numérique

Tout ordinateur contient une unité de « calcul flottant » à laquelle font appel toutes les routines numériques plus évoluées (dont celles qui occuperont les sections suivantes). Les algorithmes mis en œuvre à ce niveau posent des problèmes difficiles de gestion de la mémoire (ce que nous n'aborderons quasiment pas), de coût en temps (ce que nous détaillerons dans des cas simples) et de précision et gestion des erreurs (ce que nous effleurerons). Pour une magnifique et substantielle exposition de ces questions, l'ouvrage [Knuth] est chaudement conseillé.

1.1 Rappels et compléments sur le calcul en base b

Soit $b \geqslant 2$ un entier naturel. La numération en base b a été décrite dans les modules d'arithmétique et d'algorithmique de [L1]. Tout entier naturel n admet une représentation : $n = c_0 + \cdots + c_k b^k, c_i \in [\![0, b-1]\!]$. Si $n > 0$ et si l'on impose $c_k \neq 0$, cette représentation est unique, et l'on note : $n = (c_k \cdots c_0)_b$. La taille de n dans cette représentation est le nombre $|n|_b := k+1$ de chiffres, qui est équivalent à $\log_b n$, donc du même ordre de grandeur que le logarithme de n dans n'importe quelle base : $|n|_b = \Theta(\log n)$. Rappelons que la notation $u_n = \Theta(v_n)$ signifie que $u_n = \mathrm{O}(v_n)$ et $v_n = \mathrm{O}(u_n)$. Les estimations des coûts des algorithmes qui vont suivre seront donc largement indépendantes du choix de la base. On a donné, dans [L1], un algorithme d'addition de deux entiers de taille k qui opère en temps $\mathrm{O}(k)$. Naturellement, si les entiers sont de tailles respectives k et l, il suffit de compléter le plus court par des zéros pour obtenir une addition en temps $\mathrm{O}\big(\max(k,l)\big)$.

Exercice 1.

Décrire pareillement la soustraction $m - n$ d'entiers de même taille tels que $n \leqslant m$.

Solution. On reprend les notations de l'algorithme d'addition : m et n sont respectivement représentés par les tableaux **`C[0..k]`** et **`D[0..k]`** et le résultat est stocké dans **`E[0..k]`**.

```
r := 0 ; pour i de 0 a k faire
(x := C[i] - D[i] - r ;
 si x < 0 alors (E[i] := x+b;r := 1)
          sinon (E[i] := x ; r := 0)) ;;
```

La dernière valeur (inutilisée) de r est 0, puisque $n \leqslant m$.

Voici la description (abrégée) d'un algorithme de multiplication. Notons **`C[0..k]`** et **`D[0..l]`** les tableaux représentant m et n ; quitte à permuter m et n, nous supposons que $l+1 = |n|_b \leqslant k+1 = |m|_b$:

```
prod := 0;
pour j de 0 a l faire prod := prod + D[j] * m * b^j ;;
```

Chaque multiplication $d_j m$ se fait en temps $\Theta(|m|_b)$, car c'est une multiplication par un nombre à un chiffre, *i.e.* une succession de $(k+1)$ multiplications de deux nombres à un chiffre et de propagations de retenues. On considère que chaque multiplication par b^j se fait en temps constant $\Theta(1)$, car c'est un simple décalage sur les chiffres, directement géré par les « couches basses » du calculateur. Enfin, on additionne successivement des nombres de tailles $|m|_b + 1$, $|m|_b + 2, \ldots$, jusqu'à $|m|_b + |n|_b$. De la formule $\sum_{i=1}^{|n|_b} (|m|_b + i) = |m|_b |n|_b + \dfrac{|n|_b(|n|_b+1)}{2}$, on déduit que le coût est $\Theta(|m|_b |n|_b)$.

Une amélioration substantielle est possible en s'inspirant de la *multiplication dichotomique des polynômes* de Karatsuba ([L1]). Les nombres m et n sont écrits en base $b := 2$. On les suppose d'abord de même taille paire $k = 2k'$. On peut alors écrire $n = 2^{k'} n_0 + n_1$ et $m = 2^{k'} m_0 + m_1$. La formule clé est la suivante :

$$mn = m_1 n_1 2^k + \big((m_1+m_0)(n_1+n_0) - m_1 n_1 - m_0 n_0\big) 2^{k'} + m_0 n_0.$$

Il y a donc trois multiplications à effectuer, d'entiers de taille k' ou $k'+1$; et un certain nombre d'additions, de soustractions et de décalages (les multiplications par une puissance de 2). On obtient la formule suivante pour les coûts : $C(k) = 3C(k') + \mathrm{O}(k')$. Si $k = 2^p$, on en tire immédiatement : $C(k) = \mathrm{O}(3^p)$. Enfin, le même type d'argument que pour la multiplication dichotomique des polynômes fournit l'estimation suivante : le coût de la multiplication de deux entiers de taille k est $\mathrm{O}(k^{\log_3 2})$, ce qui est un gain net par rapport au $\mathrm{O}(k^2)$ prévisible, puisque $\log_3(2) < 2$.

Il existe évidemment des algorithmes de division généralisant ceux appris à l'école primaire, mais leur formulation précise est assez compliquée (voir là dessus [Knuth]). Nous allons plutôt examiner un algorithme *itératif* d'inversion approchée. Remarquons d'abord que, pour obtenir l'estimation $\dfrac{1}{3} \approx 0{,}33$, on se ramène à la relation entre nombres entiers $\dfrac{100}{3} \approx 33$ qui signifie ici que l'égalité a lieu à une unité près. Nous anticipons donc le fait que le calcul en représentation flottante à une certaine précision se ramène au calcul sur des entiers de longueur convenable.

Nous partons donc d'un entier naturel non nul $n = (c_k \cdots c_0)_b$: donc $b^k \leqslant n < b^{k+1}$, mais nous pouvons aussi bien supposer que $b^k < n < b^{k+1}$ (si $n = b^k$, le problème est trivial). Nous choisissons $B := b^N$, où N est un entier assez grand. Notre but est d'approcher $\dfrac{B}{n}$ par un entier. Il s'agit donc de résoudre approximativement l'équation $f(x) = 0$, où l'on a posé $f(x) := n - \dfrac{B}{x}\cdot$ La *méthode de Newton* (que nous étudierons à la section 2) suggère de choisir un point de départ x_0 convenable, puis de calculer itérativement :

$$x_{i+1} := x_i - \frac{f(x_i)}{f'(x_i)} = 2x_i - \frac{nx_i^2}{B}\cdot$$

Nous allons d'abord analyser cet algorithme en supposant les calculs exacts (et les x_i réels), puis nous l'adapterons pour ne manipuler que des entiers. Notons $x := \dfrac{B}{n}$ la valeur exacte que l'on cherche à approcher. L'erreur relative commise à la $i^{\,ème}$ itération est :

$$\varepsilon_i := \frac{x - x_i}{x} = 1 - \frac{nx_i}{B}\cdot$$

Un point de départ simple est $x_0 = b^{N-k-1}$, qui donne au départ l'encadrement $x_0 < x < bx_0$ et l'erreur relative $\varepsilon_0 = 1 - \dfrac{n}{b^{k+1}} \in \left]0, 1 - \dfrac{1}{b}\right[$. Un calcul direct montre la relation : $1 - \dfrac{nx_{i+1}}{B} = \left(1 - \dfrac{nx_i}{B}\right)^2$. On en déduit :

$$\forall i \in \mathbb{N}\ ,\ 0 < \varepsilon_i < \left(1 - \frac{1}{b}\right)^{2^i} = b^{-\beta 2^i},$$

où l'on a posé $\beta := -\log_b\left(1 - \dfrac{1}{b}\right)$. C'est un réel positif, qui vaut 1 pour $b := 2$ et

qui vaut approximativement $0{,}046$ pour $b := 10$ et $\dfrac{1}{b \ln b}$ lorsque b est grand. Autrement dit, l'erreur relative décroît très vite vers 0. À la $i^{ème}$ itération, elle a environ $\beta 2^i$ chiffres nuls après la virgule, et x_i a approximativement $\beta 2^i$ chiffres corrects : ce nombre double à chaque étape ! Nous verrons que l'on parle de *convergence quadratique*. On en déduit également une majoration du nombre d'étapes nécessaire pour arriver à une précision inférieure à 1 : il faut avoir $\varepsilon_i x < 1$, autrement dit, il suffit d'obtenir :

$$b^{-\beta 2^i} < \frac{n}{B} \iff -\beta 2^i < \log_b n - N \iff 2^i > \frac{N - \log_b n}{\beta}.$$

On aura donc au pire environ $\log_2 \dfrac{N-k}{\beta}$ étapes : cela concorde avec notre estimation de la vitesse de production des chiffres significatifs.

Le calcul de $\dfrac{x - x_i}{x}$ montre que la suite est croissante et approche x par dessous. On ne peut donc qu'accélérer le processus en remplaçant chaque x_i par son *plafond* (c'est-à-dire l'entier supérieur le plus proche) $\lceil x_i \rceil$. On garde le même point de départ et l'on itère selon la formule modifiée :

$$x_{i+1} = 2x_i - \left\lfloor \frac{n x_i^2}{B} \right\rfloor$$

(où l'on note $\lfloor y \rfloor$ la partie entière de y). Le test d'arrêt est le suivant : si $x_{i+1} = x_i$, on a trouvé la valeur exacte ; si $x_{i+1} = x_i + 1$, la valeur approchée à une unité près par défaut de x est x_i. Ces calculs sont très faciles : soustractions, multiplications, décalages. On a déjà estimé le nombre d'étapes. On peut remarquer que tous les x_i ont la même taille $N-k$: les itérations ont donc toutes même coût.

Exemples. Prenons d'abord $b := 2$, $n := 3$ et $N := 10$. On trouve : $x_0 = (100000000)_2$, $x_1 = (101000000)_2$, $x_2 = (101010100)_2$, $x_3 = (101010101)_2$. Pour ce dernier, qui représente l'arrêt, on a gardé la valeur par défaut (on a $x_3 < x < x_3 + 1$).
Prenons ensuite $b := 10$, $n := 7$ et $N := 12$.
On trouve : $x_0 = 10^{11} = 100000000000$ (il s'agit naturellement d'écriture décimale), $x_1 = 130000000000$, $x_2 = 141700000000$, $x_3 = 142847770000$, $x_4 = 142857135243$ et la valeur finale (par défaut) $x_5 = 142857142857$.

1.2 Calcul en représentation flottante

Nous allons d'abord étendre à une base $b \geqslant 2$ quelconque les résultats sur les développements décimaux du module sur les nombres réels de [L1] (pour les justifications, voir l'exercice II.8.1 de la page 951).
Tout réel positif x se décompose en $x = \lfloor x \rfloor + M(x)$, où $x \in \mathbb{N}^*$ et $M(x) \in [0, 1[$. On écrit $\lfloor x \rfloor = (c_k \cdots c_0)_b$ (ou 0, selon le cas) et $M(x) = \sum\limits_{j \geqslant 1} d_j b^{-j}$. En général, cette écriture est unique. Les exceptions sont les nombres de la forme $\dfrac{a}{b^p}$, qui admettent deux écritures :

$$y = 0, d_1 \cdots d_p = 0, d_1 \cdots d_{p-1} d'_p d'_{p+1} \cdots ,$$

où $d'_p := d_p - 1$ et $d'_j = b-1$ pour $j > p$. On dira que cette dernière écriture est impropre et on ne l'utilisera pas (penser à la double écriture $\frac{1}{4} = 0{,}25 = 0{,}24999\cdots$). On note alors : $M(x) = (0, d_1 \cdots)_b$ et $x = (c_k \cdots c_0, d_1 \cdots)_b$, ou bien $x = (0, d_1 \cdots)_b$ si $x < 1$. Tout réel non nul admet une unique écriture *flottante normalisée* :

$$x = \varepsilon y b^e,$$

où $\varepsilon \in \{\pm 1\}$ est le signe de x (on l'omet souvent s'il est positif), où $e \in \mathbb{Z}$ est appelé l'*exposant* et où $y \in \left[\frac{1}{b}, 1\right[$ est appelé la *mantisse* (noter cependant qu'il existe une autre convention dans laquelle on prend la mantisse dans $[1, b[$).
Par exemple $x = (c_k \cdots c_0, d_1 \cdots)_b$ (avec $c_k \neq 0$) admet pour exposant $k+1$ et pour mantisse $(0, c_k \cdots c_0 d_1 \cdots)_b$.

Exemples.
En numération décimale, la constante de Planck s'écrit $h = +0{,}10545\cdots 10^{-26}$; le nombre d'Avogadro s'écrit $N = +0{,}602252\cdots 10^{24}$; on a également :

$$\pi = 0{,}31415926535\cdots 10^1 \qquad e = 0{,}27181828\cdots 10^1 \qquad \sqrt{2} = 0{,}141421\cdots 10^1$$

$$\sqrt{3} = 0{,}173205\cdots 10^1 \qquad \ln 0{,}5 = -0{,}69314718\cdots 10^0 \quad \text{etc.}$$

Dans la pratique, deux limitations sont imposées par le matériel (qu'il s'agisse de la mémoire humaine, du papier ou du hardware) : le nombre de chiffres effectivement écrits de la mantisse est borné, ainsi que l'exposant. Discutons d'abord ce dernier point. En règle générale, il y a un entier e_{max} tel que l'on exige : $-e_{max} < e \leqslant e_{max}$. La dissymétrie provient d'un problème de représentation informatique : si par exemple 8 bits sont utilisés pour coder l'exposant, celui-ci peut prendre $2^8 = 256$ valeurs distinctes, et l'on conviendra que ce sont les entiers de -127 à 128. Ainsi, les réels représentables sont de taille strictement comprise entre $b^{-e_{max}}$ et $b^{e_{max}}$. Par exemple, dans beaucoup de systèmes (langage **C** sur PDP11, Turbo-Pascal …) les limites sont $10^{\pm 38}$. Le calcul de $\frac{N}{h} = \frac{+0{,}602252\cdots 10^{24}}{+0{,}10545\cdots 10^{-26}}$ est alors impossible (« overflow ») et celui de $\frac{h}{N} = \frac{+0{,}10545\cdots 10^{-26}}{+0{,}602252\cdots 10^{24}}$ est abusivement forcé à 0 (« underflow »). Nous ne discuterons plus de cette limitation.

Exemple. Le projet de norme **IEEE 754** prévoit le codage binaire des flottants en « simple précision » sur trente-deux bits. Le signe S est codé sur un bit (0 pour $+$). La mantisse est prise dans l'intervalle $[1, 2[$. On l'écrit $1 + M$ et $M \in [0, 1[$ est codé sur vingt-trois bits, qui représentent les chiffres après la virgule. L'exposant est codé sur 8 bits, et il est compris entre -127 et 128 : l'entier binaire E codé est compris entre 0 et 255 et l'exposant est égal à $E - exces$, où $exces := 127$. Par exemple, le nombre décimal $9{,}75 = 39/4$ s'écrit en binaire $1001{,}11$, en « notation scientifique » (celle des calculettes) $1{,}00111E{+}11$ (ici, en binaire) et il est codé par le bit S de signe **0**, la partie fractionnaire M de la mantisse **001 1100 0000 0000 0000 0000** et l'exposant $E = (11)_2 + exces$ soit **1000 0010**.

La taille autorisée pour représenter la mantisse est également rationnée. On remplace donc $y = (0, d_1 \cdots)_b$ par une valeur *arrondie* :

$$\overline{y} = (0, d_1 \cdots d_N)_b.$$

Le nombre N est appelé *nombre de chiffres significatifs*[1]. De même, $x = \varepsilon y b^e$ est arrondi en $\overline{x} = \varepsilon \overline{y} b^e$. Par exemple, avec trois chiffres décimaux significatifs, on écrira : $\pi \approx 0{,}314.10^1$. L'erreur relative commise en arrondissant ainsi est facilement estimée :

$$\frac{\Delta x}{x} := \frac{|\overline{x} - x|}{|x|} = \frac{(0, 0 \cdots 0 d_{N+1} \cdots)_b}{(0, d_1 \cdots d_N d_{N+1} \cdots)_b} < \frac{b}{b^N} = b^{1-N}.$$

Naturellement, si l'on a droit à 5 chiffres significatifs, il est plus judicieux d'arrondir $\pi = 0{,}31415926535 \cdots 10^1$ en $\pi \approx 0{,}31416 \cdots 10^1$ qu'en $\pi \approx 0{,}31415 \cdots 10^1$. Cela ne modifie pas l'estimation ci-dessus de l'erreur relative.

1.2.1 Gestion des erreurs d'arrondi

Pour simplifier l'exposition, nous prenons ici la base $b := 10$ et nous ne l'indiquons plus dans les notations.

Lorsque $d_{N+1} \in [\![0, 4]\!]$, il est naturel d'arrondir $y := (0, d_1 \cdots d_N d_{N+1} \cdots)$ à la valeur $\overline{y} := (0, d_1 \cdots d_N) = (d_1 \cdots d_N) 10^{-N}$. Lorsque $d_{N+1} \in [\![6, 9]\!]$, il est tout aussi naturel d'arrondir y en :

$$\overline{y} := (c'_0, d'_1 \cdots d'_N) = ((d_1 \cdots d_N) + 1)\, 10^{-N}.$$

Si $d_N < 9$, on a simplement $c'_0 := 0$, $d'_j := d_j$ pour $1 \leqslant i \leqslant N-1$ et $d'_N := d_N + 1$; mais, si $d_N = 9$, il faut tenir compte des retenues ; par exemple $0{,}314159.10^1$ (avec $N := 5$) s'arrondit en $0{,}31416.10^1$ et $0{,}999999.10^0$ en $0{,}10000.10^1$. Lorsque $d_N := 5$, on est à équidistance du haut et du bas et le choix ne va pas de soi. Deux raisons suggèrent d'arrondir vers le haut : d'abord, cela donnerait cinq chiffres (de 0 à 4) qui vont vers le bas et cinq (de 5 à 9) qui vont vers le haut ; ensuite, le chiffre $d_N = 5$ est très probablement suivi d'un chiffre non nul (même très loin) et l'on est donc plus près du haut que du bas. L'intérêt de traiter à égalité l'arrondi vers le haut et celui vers le bas est l'espoir que les erreurs commises seront indépendantes et auront tendance à se compenser en moyenne.

Exercice 2.

En numération décimale, avec $N := 3$, écrire sous forme flottante normalisée $x := 1$ et $y := 0{,}555$. Calculer ensuite : $x + y$ et son arrondi s ; $s - y$ et son arrondi x' ; $x' + y$ et son arrondi s' ; $s' - y$ et son arrondi x''. Qu'observe-t-on ?

Solution. On a $x = 0{,}100.10^1$ et $y = 0{,}555.10^0$ (aucun arrondi n'est nécessaire). Nous écrirons maintenant les résultats des calculs sous forme usuelle avant de les normaliser. On a évidemment $x + y = 1{,}555 = 0{,}1555.10^1$, d'où $s = 0{,}156.10^1$ (arrondi vers le haut puisque le chiffre supprimé est un 5). Puisque $s = 1{,}56$, on a $s - y = 1{,}005 = 0{,}1005.10^1$, d'où $x' = 0{,}101.10^1$. On trouve de même (en écriture usuelle) $x' + y = 1{,}565$, $s' = 1{,}57$, $s' - y = 1{,}015$ et $x'' = 1{,}02$. À chaque addition suivie

[1] Dorénavant, la lettre N n'aura plus rien à voir avec le nombre d'Avogadro, dont nous ne parlerons plus.

d'une soustraction (de y), la valeur arrondie dérive vers le haut. Les erreurs d'arrondi se cumulent. Une stratégie surnommée « round to even » a été étudiée, dans laquelle la règle est d'arrondir vers un chiffre pair : par exemple, $1{,}555$ en $1{,}56$ et $1{,}565$ en $1{,}56$ aussi (plutôt que $1{,}57$, car 7 est impair). On démontre que l'effet cumulatif observé dans l'exercice ne se produit plus.

Ainsi, après arrondi, $(x+y)-y \neq x$ et l'associativité est mise en défaut. Cet exemple conduit à une morale :
Les règles de calcul arrondi devraient définir des opérations qui vérifient, autant que possible, les lois algébriques usuelles.
Une fois la règle d'arrondi choisie, il faut décider comment mener les calculs.
Par exemple, avec trois chiffres significatifs, $\sqrt{2} = 1{,}41421\cdots$ et $\sqrt{3} = 1{,}73205\cdots$ sont arrondis en $0{,}141.10^1$ et $0{,}173.10^1$, dont la somme est $0{,}314.10^1$, alors que $\sqrt{2}+\sqrt{3} = 3{,}1462\cdots$ s'arrondit en $0{,}315.10^2$. Les ouvrages spécialisés disent quelle précision est nécessaire sur les opérandes pour obtenir la précision voulue sur le résultat.
Pour résumer :
Il convient d'utiliser plus de précision pour les calculs que pour la représentation des résultats.

Exercice 3.

On pose $v_0 := \dfrac{11}{2}$, $v_1 := \dfrac{61}{11}$ et $v_{n+2} := 111 - \dfrac{1130}{v_{n+1}} + \dfrac{3000}{v_n v_{n+1}}$.
Calculer $\lim\limits_{n\to+\infty} v_n$ et le vérifier numériquement.

On pourra poser $v_n = \dfrac{u_{n+1}}{u_n}$, où la suite $(u_n)_{n\in\mathbb{N}}$ vérifie une récurrence linéaire à trois pas.

Solution.
Posons $u_0 := 2$, $u_1 := 11$, $u_2 := 61$ et $u_{n+3} := 111u_{n+2} - 1130u_{n+1} + 3000u_n$. Il découlera de l'étude ci-dessous que la suite $(u_n)_{n\in\mathbb{N}}$ ne s'annule pas. La suite des $w_n := \dfrac{u_{n+1}}{u_n}$ vérifie $w_0 := \dfrac{11}{2}$, $w_1 := \dfrac{61}{11}$ et $w_{n+2} := 111 - \dfrac{1130}{w_{n+1}} + \dfrac{3000}{w_n w_{n+1}}$ (c'est immédiat). On a donc $v_n = w_n = \dfrac{u_{n+1}}{u_n}$ pour tout n.
Pour expliciter $(u_n)_{n\in\mathbb{N}}$ selon les méthodes de la proposition 37 de la page 246, on résout l'équation caractéristique $x^3 - 111x^2 + 1130x - 3000 = 0$. Les racines (« évidentes ») sont 5, 6 et 100. Il existe donc des scalaires λ, μ, ν tels que $u_n = \lambda 5^n + \mu 6^n + \nu 100^n$. Par identification pour $n := 0, 1, 2$, on trouve $\lambda = \mu = 1$ et $\nu = 0$. En particulier, $u_n = 5^n + 6^n$ ne s'annule pas, ce qui justifie le calcul de $\dfrac{u_{n+1}}{u_n}$. Enfin :

$$\lim_{n\to+\infty} v_n = \lim_{n\to+\infty} \frac{5^{n+1}+6^{n+1}}{5^n+6^n} = 6.$$

Le calcul de trente termes avec Maple (par défaut, dix chiffres significatifs sont utilisés pour les calculs) donne un résultat étrange : la suite semble approcher de 6, puis prend des valeurs chaotiques, puis part vers 100. Le passage à vingt-quatre chiffres significatifs

semble améliorer le début, puis l'on voit :

```
 5.5000000000000000000000,    5.5454545454545454545545,
 5.5901639344262295081967 7,  5.6334310850439882697955 9,
 5.6746486205101509630555 0,  5.7133290523805155493038 2,
 5.7491209197026380484575 5,  5.7818109204856156621028 4,
 5.8113142382939971519711 2,  5.8376565489587365406449 9,
 5.8609515225165529124103 7,  5.8813772158485991496976 3,
 5.8991539059121318042222 8,  5.9145249527478654841319 5,
 5.9277414427515640874922 7,  5.9390510754086489953841 9,
 5.9486974248261790076883 4,  5.9570376841240098152233 1,
 5.9666011188496243261144 2,  6.0166143026762772951326 9,
 6.7551292508656833035283 3,  17.533142662125916594158 20,
71.8802003848420755407662 2,  97.659810733073828914429 30,
99.8565840837022152038079 9,  99.991400715775645883567 60,
99.9994849001192729606876 7,  99.999969138925449190524 68,
99.9999815058883072153101,    99.999999891480215538744 5,
```

Ce résultat peut être interprété ainsi. Si l'on ne tient pas compte des conditions initiales v_0, v_1 et donc de u_0, u_1, u_2, la relation de récurrence implique une égalité de la forme :

$$v_n = \frac{\lambda 5^{n+1} + \mu 6^{n+1} + \nu 100^{n+1}}{\lambda 5^n + \mu 6^n + \nu 100^n},$$

où λ, μ, ν sont quelconques. Pour un choix *générique* de ces coefficients, donc des conditions initiales, on a donc $\lim_{n\to+\infty} v_n = 100$. Mais, sur un sous-espace particulier, $\nu = 0$ et la limite est 6 (et même 5 sur un sous-espace encore plus petit). Cependant, si l'on perturbe un tout petit peu la suite, *ce qui est l'effet d'une erreur d'arrondi*, cela revient à repartir de conditions initiales qui ne sont plus dans le sous-espace spécial. On n'a plus $\nu = 0$. Comme on n'en est pas loin, ν est petit, mais c'est quand même le terme en $\nu 100^n$ qui finira par dominer et imposer la limite 100.

On peut remarquer que les v_n sont tous rationnels et les calculer exactement. Voici les dixième, vingtième et trentième termes, calculés avec Maple :

```
70231801 / 12030821, 3751525871703601 / 628433226338621,
22200524229534883641540 1 / 37031917801711988686421,
```

Ce dernier nombre vaut $5{,}99497016287824\cdots$, il n'y a donc pas eu dérive. Nous en tirons une leçon :

Il peut être utile de combiner le calcul numérique avec du calcul formel.

Cependant, le calcul formel présente ses propres inconvénients, par exemple la *taille non bornée des données*. Dans l'exemple, les numérateurs et dénominateurs croissent en taille. Le trentième terme, avec une assez faible précision (trois chiffres significatifs), utilise des entiers de vingt-trois chiffres (le cinquantième : trente-huit chiffres, le centième : soixante dix-sept chiffres). Il est ainsi reconnu que l'algorithme du pivot exécuté sur des rationnels

peut donner lieu à une croissance incontrôlée des entiers qui y apparaissent (*cf.* la remarque de la page 935). Voici donc notre dernière leçon :
Un peu de théorie doit guider les calculs.

1.3 Performances des méthodes itératives

Mettons de côté les erreurs d'arrondi. Les problèmes que l'on peut résoudre en un nombre fini de calculs exacts sont rares : lorsque c'est le cas, on parle de *méthodes directes* : par exemple, la méthode du pivot pour résoudre un système linéaire. Dans *presque tous les cas*, aucun calcul exact ne donne le résultat souhaité. On ne peut qu'approcher celui-ci : après un nombre suffisant d'étapes, on arrive à toute précision souhaitée. On parle alors de *méthodes itératives*. Il faut d'ailleurs noter que, même lorsque des méthodes directes existent, des méthodes itératives peuvent s'avérer préférables (voir la section 5.2). Par définition, le résultat est obtenu avec une précision limitée, donc avec une *erreur de troncature* (on a « tronqué » le calcul après mille itérations pour ne pas y passer la vie ...) et il faudra tenir compte de cette erreur de troncature lorsque le résultat sera réinjecté dans d'autres calculs.

Exemple. Calculons $z := (1+x)^n = e^{n\ln(1+x)}$ à l'aide des développements en série du logarithme et de l'exponentielle. On suppose $x > 0$ et « petit ». On approche $y := \ln(1+x)$ par $\overline{y} := \sum_{i=1}^{p} \frac{(-1)^{i-1}}{i} x^i$, avec une erreur de l'ordre de $\frac{x^{p+1}}{p+1}$ (puisqu'il s'agit d'une somme partielle d'une série alternée). Le lecteur calculera le nombre d'opérations arithmétiques qui ont été nécessaires pour obtenir cette précision. On approche ensuite $z := e^y$ par $\overline{z} = \sum_{i=0}^{q} \frac{1}{i!} y^i$, avec une erreur de l'ordre de $\frac{y^{q+1}}{(q+1)!} e^y$, donc de l'ordre de $\frac{y^{q+1}}{(q+1)!}$ puisque y est petit. Le lecteur calculera encore le coût de cette précision. Mais *cette précision est fictive* : on ne dispose pas de y, seulement de la valeur tronquée $\overline{y}$ et c'est donc une valeur approchée de $e^{\overline{y}}$ que l'on calculera.

Exercice 4.
Voyez-vous une autre objection au calcul ci-dessus ?
Solution. Il est aberrant de composer deux fonctions transcendantes pour calculer une puissance $n^{ème}$! Pourtant, un étudiant de première année sur trois envisage de faire appel aux fonctions spéciales la première fois qu'il doit programmer le calcul de x^n. Ce phénomène est certainement dû à l'addiction aux calculettes.

On évalue l'intérêt, ou la performance, d'une méthode itérative en estimant le *coût* des calculs nécessaires pour obtenir une *précision* donnée. En général, on fait l'hypothèse que les itérations ont toutes le même coût : par exemple, le calcul de chaque terme lorsque l'on évalue la somme d'une série ; ou bien le passage d'un terme au suivant lorsque l'on calcule la limite d'une suite. Notons donc ε_p l'erreur de troncature après p itérations.
On dit que l'on a une *convergence linéaire* s'il existe un réel $k \in]0,1[$ tel que $\varepsilon_p = \mathrm{O}(k^p)$. Si l'on note $\beta := -\log_b k$, cela signifie qu'après p itérations, on a à peu près βp chiffres exacts. C'est cela qui justifie le terme « linéaire » et, dans la pratique, on qualifie de linéaire toute convergence qui a ce comportement, c'est à dire telle que $-\log \varepsilon_p$ est à peu près linéaire en p. Le calcul de $\ln(1+x)$ par sa série de Taylor converge donc linéairement.

Exercice 5.
Que dire du calcul de e^y ?
Solution. On a vu en exercice du module d'algorithmique de [L1] que $\ln p! \sim p \ln p$ (ce qui est d'ailleurs un corollaire de la *formule de Stirling*, théorème 4 de la page 892). Le nombre de chiffres exacts est donc $\mathrm{O}(p \ln p)$: on pourrait parler de convergence quasi-linéaire.

Lorsque $-\log \varepsilon_p$ est de l'ordre de grandeur de α^p pour un certain $\alpha > 1$, on dit que l'on a une *convergence d'ordre* α. Si $\alpha = 2$, c'est une *convergence quadratique* : dans ce cas, le nombre de chiffres exacts est à peu près doublé à chaque itération. C'est le cas du calcul de B/n par la suite $x_{i+1} := 2x_i - \dfrac{nx_i^2}{B}$.

Exemples. (i) La fonction $\varphi(x) := \dfrac{2x+2}{x+2}$ est strictement croissante de $\mathbb{R}_+$ dans $\mathbb{R}_+$, et l'unique $x \in \mathbb{R}_+$ tel que $\varphi(x) = x$ est $\sqrt{2}$. Posons $u_0 := 1$ et $u_{n+1} := \varphi(u_n)$. Cette suite ne prend que des valeurs positives. Comme $u_1 = 4/3$, on a $u_0 < u_1 < \sqrt{2}$, d'où, par récurrence, une suite strictement croissante et majorée par $\sqrt{2}$. Elle admet donc une limite, qui, par continuité de φ, doit être un point fixe, donc $\sqrt{2}$. De l'égalité :

$$u_{n+1} - \sqrt{2} = \frac{(2-\sqrt{2})(u_n - \sqrt{2})}{u_n + 2}, \quad \text{on déduit que } {}_{n+1} - \sqrt{2} \sim \frac{2-\sqrt{2}}{2+\sqrt{2}}(u_n - \sqrt{2}),$$

donc que la convergence est linéaire (exercice II.8.5 de la page 951). C'est un cas particulier de la méthode de la sécante (section 2.2.1).

(ii) La fonction $\psi(x) := \dfrac{1}{2}\left(x + \dfrac{2}{x}\right)$ est strictement croissante sur $\left[\sqrt{2}, +\infty\right[$ et l'unique $x \in \mathbb{R}_+$ tel que $\psi(x) = x$ est $\sqrt{2}$. Posons $v_0 := 2$ et $v_{n+1} := \psi(v_n)$. Cette suite ne prend que des valeurs positives. Comme $v_1 = 3/2$, on a $v_0 > v_1 > \sqrt{2}$, d'où, par récurrence, une suite strictement décroissante et minorée par $\sqrt{2}$. Elle admet donc une limite, qui, par continuité de ψ, doit être un point fixe, donc $\sqrt{2}$. De l'égalité :

$$v_{n+1} - \sqrt{2} = \frac{(v_n - \sqrt{2})^2}{2v_n}, \quad \text{on déduit que } v_{n+1} - \sqrt{2} \sim \frac{1}{2\sqrt{2}}(v_n - \sqrt{2})^2,$$

donc que la convergence est quadratique (exercice II.8.5 de la page 951). C'est un cas particulier de la méthode de Newton (section 2.2.3).

1.4 La formule de Stirling

Il s'agit d'une formule extrêmement utile. Sous sa forme générale (que nous donnons en [MPA1]) elle fournit un développement asymptotique de la fonction Gamma (*cf.* l'exercice 16 de la page 614). Le cas particulier d'un équivalent de $n!$ est facile à prouver et suffisant pour nos besoins dans la suite de ce module.

Lemme 1. Il existe une constante $C > 0$ telle que, lorsque n tend vers $+\infty$:

$$n! \sim C \left(\frac{n}{e}\right)^n \sqrt{n}. \tag{1}$$

Démonstration. Notons $u_n := \ln \dfrac{n!}{\left(\dfrac{n}{e}\right)^n \sqrt{n}}$.

On a donc $u_n = \ln n! - \left(n + \dfrac{1}{2}\right) \ln n + n$ et $u_{n+1} - u_n = 1 - \left(n + \dfrac{1}{2}\right) \ln \dfrac{n+1}{n}$.
De la formule de Taylor-Young $\ln(1+x) = x - \dfrac{x^2}{2} + \mathrm{O}(x^3)$, on déduit alors que $u_{n+1} - u_n = \mathrm{O}\left(\dfrac{1}{n^2}\right)$. La série de terme général $u_{n+1} - u_n$ est donc absolument convergente, la suite $(u_n)_{n\in\mathbb{N}}$ admet donc une limite $c \in \mathbb{R}$ et $\dfrac{n!}{\left(\dfrac{n}{e}\right)^n \sqrt{n}}$ tend vers $C := e^c$. ∎
Introduisons maintenant les *intégrales de Wallis* :

$$\forall n \in \mathbb{N}\ ,\ I_n := \int_0^{\pi/2} \cos^n x \,\mathrm{d}x.$$

On trouve immédiatement $I_0 = \dfrac{\pi}{2}$ et $I_1 = 1$. Par intégrations par parties, on trouve une relation de récurrence à deux pas :

$$\begin{aligned} I_{n+2} &= \int_0^{\pi/2} \cos^{n+1} x \,\sin' x \,\mathrm{d}x \\ &= \left[\cos^{n+1} x \,\sin x\right]_0^{\pi/2} - \int_0^{\pi/2} \left(\cos^{n+1}\right)' x \sin x \,\mathrm{d}x \\ &= (n+1) \int_0^{\pi/2} \cos^n x \sin^2 x \,\mathrm{d}x \\ &= (n+1) \int_0^{\pi/2} \cos^n x (1 - \cos^2 x) \,\mathrm{d}x \\ &= (n+1)(I_n - I_{n+2}), \end{aligned}$$

d'où la relation annoncée : $I_{n+2} = \dfrac{n+1}{n+2} I_n$.
On en tire d'abord $(n+2) I_{n+1} I_{n+2} = (n+1) I_n I_{n+1}$, qui ne dépend donc pas de n, d'où ensuite $(n+1) I_n I_{n+1} = 1 I_1 I_0 = \dfrac{\pi}{2}$.

Lemme 2. Lorsque n tend vers $+\infty$, on a :

$$I_n \sim \sqrt{\frac{\pi}{2n}}. \tag{2}$$

Démonstration. En intégrant l'inégalité : $\cos^{n+1} x \leqslant \cos^n x$ (valable sur $[0, \pi/2]$), on trouve : $I_{n+1} \leqslant I_n$. Combinée avec la relation de récurrence, cette inégalité donne l'encadrement :

$$\forall n \in \mathbb{N}\ ,\ \frac{n+1}{n+2} \leqslant \frac{I_{n+1}}{I_n} \leqslant 1.$$

Ainsi, $I_{n+1} \sim I_n$, donc $\dfrac{\pi}{2} = (n+1) I_n I_{n+1} \sim n I_n^2$.
Comme $I_n > 0$, la conclusion s'ensuit. ∎

Proposition 3. Lorsque n tend vers $+\infty$, on a l'équivalence :

$$\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n}}. \tag{3}$$

Démonstration. En itérant la relation de récurrence $I_{2p+2} = \dfrac{2p+1}{2p+2} I_{2p}$, on obtient :

$$I_{2p} = \frac{(2p-1)\cdots 1}{2p\cdots 2} I_0 = \frac{(2p)!}{(2p\cdots 2)^2}\frac{\pi}{2} = \frac{\pi}{2} 4^{-p}\binom{2p}{p}.$$

En combinant avec l'équivalence $I_{2p} \sim \sqrt{\dfrac{\pi}{4p}}$ du lemme, on obtient la formule indiquée.

Théorème 4 (Formule de Stirling).
Lorsque n tend vers $+\infty$, on a l'équivalence :

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}. \tag{4}$$

Démonstration.
En reportant la relation (1) (du lemme 1) dans la relation (3) ci-dessus, on trouve $C = \sqrt{2\pi}$.

2 Résolution approchée de l'équation $f(x) = 0$

De nombreux théorèmes d'algèbre ou d'analyse affirment : « il existe un réel (ou un complexe) x tel que $f(x) = 0$ », où f désigne une fonction. Nous dirons que x est *un zéro* ou *une racine* de f (le terme « racine » s'emploie surtout dans le cas des polynômes, mais ce n'est pas une règle absolue). Les archétypes sont le théorème des valeurs intermédiaires (et la fonction f est alors continue sur un intervalle de $\mathbb{R}$) et le théorème de d'Alembert-Gauß (et la fonction f est alors un polynôme). Un autre exemple est le théorème de Rolle, qui affirme l'existence d'un zéro de f'. Nous aborderons ici le cas des fonctions d'une variable réelle. Il existe des extensions au cas de fonctions de plusieurs variables (par exemple, de la méthode de Newton) mais elles sont nettement plus complexes (voir [MPA1]).

2.1 Principes généraux

La première étape est, dans presque tous les cas, de déterminer un intervalle $[a, b] \subset \mathbb{R}$ sur lequel la fonction f change de signe : $f(a)f(b) < 0$. Si f est continue, le théorème des valeurs intermédiaires garantit alors l'existence d'un zéro de f sur l'intervalle et le but est de l'approcher. Mais ce projet est ambigu : il pourrait en effet y avoir *plusieurs* zéros de f et l'algorithme utilisé pourrait en approcher un sans déceler la présence des autres. On considère donc comme première étape la détermination du nombre de zéros de f et d'intervalles qui en contiennent chacun un ou aucun.

Exemple. Soit $f(x) := x^3 + px + q$, où $p, q \in \mathbb{R}$. Puisque f est continue sur $\mathbb{R}$ et que ses limites aux bornes sont $-\infty$ et $+\infty$, elle admet au moins un zéro réel (comme toute fonction polynomiale réelle de degré impair). Le nombre de ces zéros est, au plus, son degré 3. Si $p \geqslant 0$, la fonction dérivée $f'(x) = 3x^2 + p$ est strictement positive partout sauf peut-être en 0, et f est strictement croissante sur $\mathbb{R}$: elle admet donc un unique zéro. Si $p < 0$, notons $x' := -\sqrt{-p/3}$ et $x'' := \sqrt{-p/3}$ les deux zéros de f'. La fonction f est croissante, puis décroissante, puis croissante, avec un maximum local en x' et un minimum local en x''. L'examen de son graphe montre que le critère discriminant est alors l'existence d'un changement de signe entre ces deux points. On vérifie sans peine que :

$$f(x')f(x'') = \frac{\Delta}{27}, \text{ où } \Delta := 4p^3 + 27q^2$$

(voir la section 2.2 de la page 338 du module I.7). On a alors les cas suivants :

- Si $\Delta > 0$, il y a un seul zéro réel.
- Si $\Delta < 0$, il y a un zéro réel sur chacun des intervalles $]-\infty, x'[$, $]x', x''[$ et $]x'', +\infty[$.
- Si $\Delta = 0$, il y a un zéro double et un zéro simple réels.

Naturellement, ces informations ne sont pas tout à fait suffisantes pour attaquer la recherche d'un zéro.

Exercice 6.

Tout zéro (réel ou complexe) de $f(x) := x^3 + px + q$ vérifie : $|x| \leqslant \max(1, |p| + |q|)$.

Solution. Si x est un zéro tel que $|x| > 1$, alors $x = -\dfrac{p}{x} - \dfrac{q}{x^2}$, d'où, par l'inégalité du triangle :

$$|x| \leqslant \left|\frac{p}{x}\right| + \left|\frac{q}{x^2}\right| \leqslant |p| + |q|\,.$$

(Cette preuve s'applique encore lorsque $p, q \in \mathbb{C}$.)

Exercice 7.

Dans le cas où $p < 0$ et $\Delta > 0$, comment placer le zéro réel ?
Indication : étudier le signe de $f(x') + f(x'')$.

Solution. Puisque $\Delta > 0$, $f(x')$ et $f(x'')$ sont de même signe. On a $f(x') + f(x'') = 2q$, qui n'est pas nul car $27q^2 > -4p^3 > 0$. Si $q > 0$, $f(x'), f(x'') > 0$ et le zéro est à gauche de x'. Si $q < 0$, $f(x'), f(x'') < 0$ et le zéro est à droite de x''.

L'utilisation du *tableau de variations* de f est donc un outil de base pour l'étape de *localisation des racines*. Notons que, dans le cas des fonctions polynomiales, il existe des algorithmes complets et efficaces pour cette étape : nous abordons ce sujet un peu plus loin (section 2.1.1).

Une fois déterminé un intervalle $[a, b] \subset \mathbb{R}$ sur lequel la fonction f change de signe, la méthode la plus simple est la *dichotomie* : on coupe l'intervalle de recherche $]a, b[$ en deux et l'on calcule la valeur au milieu $c := \dfrac{a+b}{2}$. Si $f(c) = 0$, on a trouvé le zéro. Si $f(c)$ est du même signe que $f(a)$, on remplace l'intervalle de recherche $]a, b[$ par $]c, b[$. Si $f(c)$ est

du même signe que $f(b)$, on remplace l'intervalle de recherche par $]a, c[$. C'est d'ailleurs la méthode qui a servi à *prouver* l'existence d'un zéro dans le module « Fonctions réelles » de [L1].

Exemple.
L'équation $x^3 - x - 1 = 0$ a un seul zéro réel car $p = -1 < 0$ et $\Delta = 23 > 0$. D'après les critères vus plus haut, ce zéro est supérieur à $\sqrt{1/3}$ et inférieur à 2. En fait, $f(1) = -1$ et $f(2) = 5$, on cherche donc entre 1 et 2. On a $f(3/2) = 7/8$, donc $x \in]1, 3/2]$; puis $f(5/4) = -19/64$, donc $x \in]5/4, 3/2]$; puis $f(11/8) = 115/512$, donc $x \in]5/4, 11/8]$, etc.

Quand arrête-t-on le calcul ? Quand on a trouvé le zéro, bien entendu, mais peut-être ne le trouvera-t-on pas, ce qui est le plus probable : en effet, les valeurs en lesquelles on calcule f sont toutes de la forme $\dfrac{ma + (2^n - m)b}{2^n}$, où n est un entier positif (le nombre d'étapes de calcul) et m un entier compris entre 0 et 2^n. Le lecteur vérifiera d'ailleurs que n'importe quel tel nombre peut être atteint par une succession de choix gauche-droite. En tous cas, l'ensemble de ces valeurs-tests est dénombrable, donc il n'épuise pas l'intervalle.
En revanche, il est *dense* dans $]a, b[$, donc il permet d'approcher la racine avec toute précision désirée : après n étapes, l'intervalle dans lequel on a situé la racine est de longueur $\dfrac{b-a}{2^n}$, ce qui suggère le test d'arrêt suivant : si ε est la précision désirée, on s'arrête dès que l'on a trouvé (exactement) la racine ou dès que $\dfrac{b-a}{2^n} \leqslant \varepsilon$ (exercice II.8.7 de la page 952). Le nombre n d'étapes nécessaires est donc proche de $\log_2 \dfrac{b-a}{\varepsilon}$, ce que l'on écrira : $n = \Theta(-\log \varepsilon)$.
Il est important de mesurer la *performance* de cet algorithme, c'est-à-dire le coût des calculs nécessaires pour obtenir la précision voulue. Nous supposerons que les calculs $f(c)$ (et les comparaisons avec 0) se font de manière exacte et à coût constant : ce n'est pas totalement réaliste (section 1 de la page 881), mais c'est une estimation raisonnable dans bien des cas. Le coût est donc proportionnel au nombre n d'étapes. Le coût d'une précision ε est donc $\Theta(-\log \varepsilon)$, ce que l'on exprime en disant que l'on a une *convergence linéaire*.
On peut comprendre cette terminologie en évaluant le nombre de chiffres connus après n étapes. Quelle que soit la base de numération choisie (base 10 : décimales, base 2 : bits ...), ce nombre de chiffres est essentiellement proportionnel à $-\log \varepsilon$. Dans la méthode dichotomique, le nombre de chiffres obtenus est proportionnel au temps de calcul. *Ce n'est pas une bonne performance* et nous tenterons de faire mieux.
Avant de passer à des méthodes moins primaires, signalons des améliorations mineures mais faciles à envisager. La première est très intuitive. Si $f(a)$ et $f(b)$ sont de taille très différente, on peut s'attendre à trouver la racine plus près de l'un que de l'autre.

Exemple. On cherche x tel que $2^x - 1000 = 0$. On a $2^{10} - 1000 = 24$ et $2^9 - 1000 = -488$. On s'attend donc à trouver x proche de 10. Puisque 0 se trouve aux $488\ 512^{\text{èmes}}$ de

l'intervalle $[9, 10]$, on pourrait même viser $9 + \dfrac{488}{512}$ plutôt que $9{,}5$. De fait,

$$9 + \frac{488}{512} = 9{,}953125 \text{ (exact) et } \log_2 1000 \approx 9{,}97.$$

L'idée sous-jacente est que, entre 9 et 10, on peut approcher la fonction $f(x) := 2^x - 1000$ par son *interpolation linéaire* $L(x) := -488 + 512(x-9)$, et donc l'unique zéro $\log_2 1000$ de f par l'unique zéro $9 + \dfrac{488}{512}$ de L. Cette méthode sera étudiée plus loin (section 2.2.1).

À défaut d'*approcher* la fonction f par une fonction connue, on peut parfois tirer parti d'un *encadrement*.

Exercice 8.

À l'aide de l'encadrement $x - x^3/6 \leqslant \sin x \leqslant x$, valable sur $\mathbb{R}_+$, déterminer le plus petit $a > 0$ tel que $\sin a = 1/2$.

Solution. Sur $\left[0, \sqrt{2}\right]$, les trois fonctions sont croissantes.

On vérifie que $f(x) := x^3 - 6x + 3$ admet exactement un zéro b sur $\left[0, \sqrt{2}\right]$ et l'on en déduit : $1/2 \leqslant x \leqslant b$. Comme $f(1/2) = 1/8$ et $f(1) = -2$, une interpolation linéaire donnerait $b \approx 9/17 \approx 0{,}53$. En fait, nous savons bien que $a = \pi/6 \approx 0{,}52$ et le résultat n'est pas si mauvais !

2.1.1 Localisation des racines d'un polynôme réel

Proposition 5.

Soit $P := a_0 X^n + \cdots + a_n \in \mathbb{R}[X]$ un polynôme non constant de degré n.

Les racines de P sont majorées, en valeur absolue, par $\max\left(1, \left|\dfrac{a_1}{a_0}\right| + \cdots + \left|\dfrac{a_n}{a_0}\right|\right)$.

Démonstration. Si $P(x) = 0$ et $|x| > 1$, de l'égalité $x = -\sum_{i=1}^{n} \dfrac{a_i}{a_0 x^{i-1}}$, on déduit :

$$|x| \leqslant \sum_{i=1}^{n} \left|\frac{a_i}{a_0 x^{i-1}}\right| \leqslant \sum_{i=1}^{n} \left|\frac{a_i}{a_0}\right|.$$

Définition 1. Soient $f_0, \ldots, f_k$ des fonctions continues sur un intervalle I de $\mathbb{R}$. On dit que $(f_0, \ldots, f_k)$ est *une suite de Sturm sur* I si :

1. Les f_i n'ont qu'un nombre fini de zéros sur I.
2. Si f_0 s'annule en $x \in I$, il y a deux possibilités : soit $f_1(x) > 0$ et f_0 est strictement croissante dans un voisinage de x ; soit $f_1(x) < 0$ et f_0 est strictement décroissante dans un voisinage de x.
3. Pour tout $i \in [\![1, k-1]\!]$, il existe une fonction g_i telle que $f_{i-1} = g_i f_i - f_{i+1}$.
4. La fonction f_k est constante non nulle.

La condition 2 exprime que f_1 « se comporte comme la dérivée de f_0 ». Elle implique que les zéros de f_0 sont intérieurs à I et que f_0 y change de signe. Les conditions 3 et 4 entraînent que, pour $i \in [\![1, k]\!]$, les fonctions f_i et f_{i-1} n'ont pas de zéro commun. En effet, si $f_i(x) = f_{i-1}(x) = 0$, alors $f_{i+1}(x) = 0$ (par la condition 3), et, en itérant, on trouve que $f_k(x) = 0$, ce qui contredit la condition 4.
Pour toute suite finie de réels $(y_0, \dots, y_k)$, on appelle *changement de signe* un couple d'indices (i, j) tels que $0 \leqslant i < j \leqslant k$, tous les y_l avec $i < l < j$ sont nuls et $y_i y_j < 0$. Le *nombre de changements de signes* de $(y_0, \dots, y_k)$ est donc le nombre de tels couples. La remarque élémentaire suivante nous sera utile. Supposons que $y_i y_{i+2} < 0$. Alors, que y_{i+1} soit positif, nul ou négatif, il se sera produit exactement un changement de signe entre y_i et y_{i+2}.

Théorème 6. Soient $a < b$ des réels et $(f_0, \dots, f_k)$ une suite de Sturm sur $[a; b]$. En particulier, $f_0(a), f_0(b) \neq 0$. Pour tout $x \in [a; b]$, notons $w(x)$ le nombre de changements de signe de la suite $\big(f_0(x), \dots, f_k(x)\big)$. Alors le nombre de zéros distincts de f_0 sur $[a; b]$ est $w(a) - w(b)$.

Démonstration. Notons $a_1 < \cdots < a_m$ tous les points de $[a; b]$ où l'un des f_i au moins s'annule (règle 1). Le raisonnement repose sur le « tableau de variations généralisé » :

	a	$\cdots$	a_j	$\cdots$	$a_{j'}$	$\cdots$	b
f_0	$\pm$	$\cdots$	0	$\cdots$	$\pm$	$\cdots$	$\pm$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
f_i	$\pm$	$\cdots$	$\pm$	$\cdots$	0	$\cdots$	$\pm$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
f_k	$+$	$+$	$+$	$+$	$+$	$+$	$+$

Le choix du signe $+$ pour f_k est arbitraire, l'important est que ce signe soit le même sur toute la ligne. Sur chaque intervalle $]a_j; a_{j+1}[$, aucun f_i ne s'annule, donc tous sont de signe constant (théorème des valeurs intermédiaires : il s'applique puisque les f_i ont été supposées continues). Sur ces intervalles, la fonction $w(x)$ est donc constante. Il reste à voir comment elle se comporte au voisinage des a_j. Si a_j est un zéro de f_0, le haut du tableau de variations au voisinage de a_j a l'une des allures suivantes :

	$\cdots$	a_j^-	a_j	a_j^+	$\cdots$
f_0	$\cdots$	$-$	0	$+$	$\cdots$
f_1	$\cdots$	$+$	$+$	$+$	$\cdots$

ou

	$\cdots$	a_j^-	a_j	a_j^+	$\cdots$
f_0	$\cdots$	$+$	0	$-$	$\cdots$
f_1	$\cdots$	$-$	$-$	$-$	$\cdots$

Cela ne fait que traduire la règle 2. On voit que le changement de signe $(-, +)$ ou $(+, -)$ a disparu au passage de a_j. Si a_j est un zéro de f_i avec $1 \leqslant i \leqslant k-1$, voici l'allure des lignes $i-1, i, i+1$ au voisinage des a_j :

	$\cdots$	a_j^-	a_j	a_j^+	$\cdots$
f_{i-1}	$\cdots$	$+$	$+$	$+$	$\cdots$
f_i	$\cdots$	$?$	0	$?$	$\cdots$
f_{i+1}	$\cdots$	$-$	$-$	$-$	$\cdots$

ou

	$\cdots$	a_j^-	a_j	a_j^+	$\cdots$
f_{i-1}	$\cdots$	$-$	$-$	$-$	$\cdots$
f_i	$\cdots$	$?$	0	$?$	$\cdots$
f_{i+1}	$\cdots$	$+$	$+$	$+$	$\cdots$

Cette configuration traduit exactement la règle 3. Les symboles ? n'ont aucune importance : comme remarqué précédemment, à gauche comme au milieu comme à droite, il y a un changement de signe entre $i-1$ et $i+1$. Cela s'applique également si $a_j = a$ ou b.
En conclusion, $w(x)$ diminue de 1 chaque fois que x passe de gauche à droite d'un zéro de f_0 et il reste constant ailleurs. L'assertion du théorème en découle. ∎

Pour rendre ce théorème opérationnel, nous allons associer à tout polynôme $P \in \mathbb{R}[X]$ non constant une suite de Sturm. Pour cela, nous posons $f_0 := P$, $f_1 := P'$ et les suivants sont définis par divisions euclidiennes : $f_{i+1} := -f_{i-1} \mod f_i$. C'est donc, aux signes près, l'algorithme d'Euclide. Soit k l'indice tel que $f_k \neq 0$, $f_{k+1} = 0$. Remarquons que tous ces polynômes sont dans $\mathbb{R}[X]$. Si f_k est constant, c'est-à-dire si P et P' sont premiers entre eux, alors $(f_0, \ldots, f_k)$ est une suite de Sturm sur tout intervalle $I = [a; b]$ tel que $P(a) \neq 0$ et $P(b) \neq 0$. Toutes les conditions de la définition sont en effet évidentes. Si f_k n'est pas constant, c'est un pgcd de P et P'. On pose alors, pour tout $i \in [\![0, k]\!]$, $g_i := \dfrac{f_i}{f_k}$.
Ainsi g_0 a les mêmes racines que P. De plus, on a encore une suite de Sturm. Ici, c'est la condition 2 qui est la seule non évidente. On la prouve comme suit. On décompose P sur $\mathbb{C}$: $P = \prod (X - z_i)^{\mu_i}$, d'où : $\dfrac{P'}{P}(x) = \sum \dfrac{\mu_i}{x - z_i}$, qui est aussi l'expression de $\dfrac{f_1}{f_0} = \dfrac{g_1}{g_0}$. D'autre part, $g_0 = C \prod (X - z_i)$, d'où : $\dfrac{g_0'}{g_0}(x) = \sum \dfrac{1}{x - z_i}$, et, en combinant ces égalités : $\dfrac{g_1}{g_0'}(x) = \dfrac{\sum \frac{\mu_i}{x - z_i}}{\sum \frac{1}{x - z_i}}$.
Cela montre qu'au voisinage d'une racine de P, la fonction $\dfrac{g_1}{g_0'}(x)$ est équivalente à sa multiplicité de cette racine, donc strictement positive : g_1 vérifie donc bien la règle.

Exemple. Prenons $f_0 := P := X^3 + pX + q$ $(p, q \in \mathbb{R})$, donc $f_1 := P' := 3X^2 + p$ et $f_2 = -(2p/3)X - q$. On suppose p et q non nuls, d'où $f_3 = -\dfrac{\Delta}{4p^2}$, où $\Delta = 4p^3 + 27q^2$.
Si $\Delta = 0$, P a une racine multiple. Nous supposerons $\Delta \neq 0$. On choisit a strictement plus petit et b strictement plus grand que toutes les racines réelles des f_i. Le signe de $f_i(a)$ est donc le signe de f_i en $-\infty$, qui est totalement déterminé par son terme dominant (signe du coefficient dominant et parité du degré). De même, le signe de $f_i(b)$ est le signe de f_i en $+\infty$, qui est le signe de son coefficient dominant. On obtient donc :

$$w(a) = \text{ nombre de changements de signe de } -,\ +,\ \operatorname{signe}(p),\ -\operatorname{signe}(\Delta)$$
$$w(b) = \text{ nombre de changements de signe de } +,\ +,\ -\operatorname{signe}(p),\ -\operatorname{signe}(\Delta).$$

Par exemple, si $\Delta > 0$, $w(a) = 2$, $w(b) = 1$ et P a une seule racine réelle. Si $\Delta < 0$ et $p > 0$, $w(a) = 1$, $w(b) = 2$, ce qui semble bizarre : mais en fait, ce cas ne peut pas se produire puisque $4p^3 + 27q^2 < 0 \Rightarrow p < 0$. Si $\Delta < 0$ et $p < 0$, $w(a) = 3$, $w(b) = 0$ et P a trois racines réelles (distinctes).

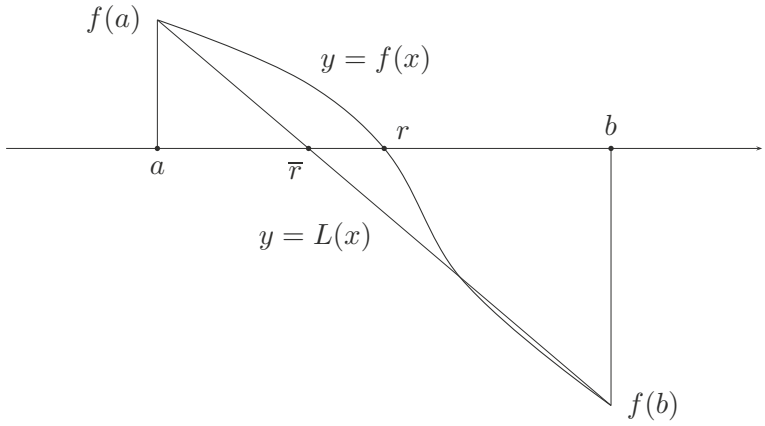
2.2 Les principales méthodes

2.2.1 Méthode d'interpolation linéaire

Nous supposons ici que f est continue et monotone sur $[a,b]$ et que $f(a)f(b)<0$. Elle s'annule donc en un unique point $r\in\,]a,b[$. Pour approcher r, nous allons interpoler f en a,b par un polynôme L de degré 1 et déterminer le point $\overline{r}$ où celui-ci s'annule : puisque $L(a)L(b)=f(a)f(b)<0$, on a $\overline{r}\in\,]a,b[$. Le calcul de L et $\overline{r}$ est immédiat :

$$L(x)=\frac{(x-a)f(b)-(x-b)f(a)}{b-a}=f(a)+\frac{f(b)-f(a)}{b-a}\,(x-a)\ \text{ et }\overline{r}=\frac{af(b)-bf(a)}{f(b)-f(a)}.$$

C'est donc le point d'intersection de la sécante menée par $\big(a,f(a)\big)$ et $\big(b,f(b)\big)$. Cette manière d'approcher r par $\overline{r}$ porte le nom de *méthode de la fausse position* ou *regula falsi*.



Nous supposerons désormais que f est de classe $\mathcal{C}^2$ sur un intervalle ouvert I contenant $[a,b]$. De plus, nous supposerons que f' ne s'annule pas sur $[a,b]$. Cela entraîne donc qu'elle est de signe constant, donc que f y est strictement monotone.

Lemme 7. Pour tout $x\in\,]a,b[$, il existe $u\in\,]a,b[$ tel que :

$$f(x)-L(x)=\frac{1}{2}f''(u)(x-a)(x-b).$$

Démonstration. C'est un cas particulier du théorème 13 de la page 906. ∎

Proposition 8. Notons encore r le zéro de f sur $]a,b[$ et $\overline{r}$ celui de L. Alors :

$$\exists u,v\in\,]a,b[\ :\ \overline{r}-r=\frac{1}{2}\frac{f''(u)}{f'(v)}(\overline{r}-a)(\overline{r}-b). \tag{5}$$

En conséquence, si l'on note $m:=\min\limits_{[a,b]}|f'|$ et $M:=\max\limits_{[a,b]}|f''|$, on a :

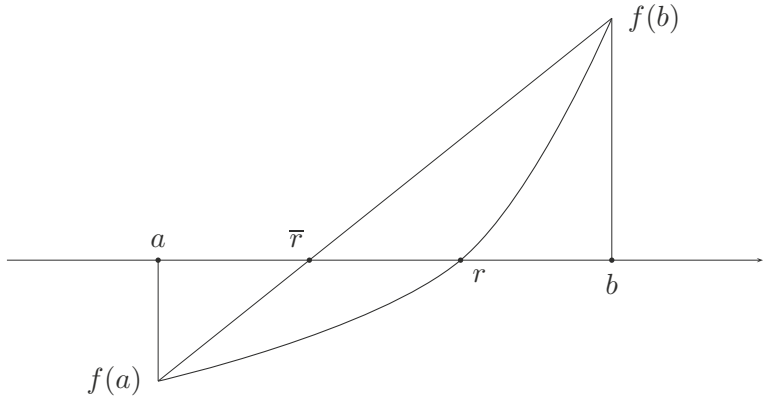
$$|\overline{r}-r|\leqslant\frac{M}{2m}\,|\overline{r}-a|\,|\overline{r}-b|\,. \tag{6}$$

Démonstration. D'après le lemme, $f(\overline{r}) = f(\overline{r}) - L(\overline{r}) = \dfrac{1}{2}f''(u)(\overline{r} - a)(\overline{r} - b)$ pour un certain $u \in]a, b[$.
D'après le théorème des accroissements finis, $f(\overline{r}) = f(\overline{r}) - f(r) = (\overline{r} - r)f'(v)$ pour un certain $v \in]a, b[$. Par hypothèse, on peut diviser par $f'(v)$, ce qui donne la première relation. La deuxième relation est alors immédiate. ∎

Une fois obtenue une évaluation de l'erreur commise, on peut décider que l'approximation est suffisante. Sinon, on réitère le procédé sur l'un des deux intervalles : $[a, \overline{r}]$ dans le cas où $f(\overline{r})$ est du signe de $f(b)$; ou $[\overline{r}, b]$ dans le cas où $f(\overline{r})$ est du signe de $f(a)$. On peut démontrer que le processus converge.

Méthode de la sécante

Nous supposerons que l'intervalle $[a, b]$ a été déterminé de telle manière que f' et f'' ne s'y annulent pas (cela relève de l'étape préliminaire de localisation des racines). Quitte à opérer des symétries, c'est-à-dire à remplacer $f(x)$ par $\pm f(\pm x)$, on peut supposer que f' et f'' sont strictement positives, donc que f est strictement croissante et strictement convexe.



De la proposition ou du dessin, on tire que $a < \overline{r} < r$. Le processus peut alors être décrit sous forme itérative :

$$x_0 := a \quad \text{et} \quad x_{n+1} := \frac{x_n f(b) - b f(x_n)}{f(b) - f(x_n)}.$$

La suite (x_n) est croissante et majorée par r. Elle converge donc vers un réel $l \leqslant r$ tel que

$$\ell = \frac{l f(b) - b f(l)}{f(b) - f(l)},$$

ce qui implique $f(\ell) = 0$ donc $\ell = r$. L'erreur peut être majorée à chaque étape par la formule :

$$|x_{n+1} - r| \leqslant \frac{M}{2m} |x_{n+1} - x_n| \, |b - x_n| ,$$

ce qui permet de choisir un test d'arrêt : on s'arrête dès que $\dfrac{M}{2m} |x_{n+1} - x_n| \, |b - x_n| < \varepsilon$.

Exemple. Soit $\alpha > 1$. Sur $[a,b] := [1,\alpha]$, la fonction $f(x) := x^2 - \alpha$ vérifie les trois hypothèses $f(a)f(b) < 0$, $f' > 0$ et $f'' > 0$. Son unique zéro est $r := \sqrt{\alpha}$. On pose donc :

$$x_0 := a = 1 \quad \text{et} \quad x_{n+1} := \frac{x_n(b^2-\alpha) - b(x_n^2-\alpha)}{(b^2-\alpha)-(x_n^2-\alpha)} = \frac{bx_n+\alpha}{b+x_n}\cdot$$

L'homographie $\varphi(x) := \dfrac{bx+\alpha}{b+x}$ admet pour points fixes $\pm\sqrt{\alpha}$ et un petit calcul donne :

$$\frac{\varphi(x)-\sqrt{\alpha}}{\varphi(x)+\sqrt{\alpha}} = \frac{b-\sqrt{\alpha}}{b+\sqrt{\alpha}}\,\frac{x-\sqrt{\alpha}}{x+\sqrt{\alpha}}\cdot$$

On en tire la formule :

$$\frac{x_n-\sqrt{\alpha}}{x_n+\sqrt{\alpha}} = \frac{x_0-\sqrt{\alpha}}{x_0+\sqrt{\alpha}}\left(\frac{b-\sqrt{\alpha}}{b+\sqrt{\alpha}}\right)^n = Cq^n,$$

où $C := \dfrac{x_0-\sqrt{\alpha}}{x_0+\sqrt{\alpha}}$ est fixé et où $q := \dfrac{b-\sqrt{\alpha}}{b+\sqrt{\alpha}} \in]0,1[$.

On en déduit que $x_n - \sqrt{\alpha} \sim 2\sqrt{\alpha}\,C\,q^n$: la convergence est donc linéaire.

Exercice 9.

Appliquer la méthode de la sécante à $f(x) := x^3$ sur $[a,b] := [-1/2, 1]$. On montrera que la suite est à valeurs strictement négatives et croissante.

Estimer la vitesse de convergence en appliquant le lemme de Cesàro à $y_{n+1}^2 - y_n^2$, où $y_n := -1/x_n$.

Solution. On a bien $f(a)f(b) < 0$ et $f' > 0$ sur cet intervalle (sauf en 0, mais cela ne change pas les conclusions) d'où la présence d'un unique zéro, mais pas $f'' > 0$. On pose

$$x_0 := a = -1/2 \text{ et } x_{n+1} := \frac{x_n b^3 - b x_n^3}{b^3 - x_n^3} = \frac{x_n^2 + x_n}{x_n^2 + x_n + 1}\cdot$$

Si $x_n \in]-1,0[$, alors $x_n^2 + x_n < 0$. Comme $x_{n+1} - x_n = -\dfrac{x_n^3}{x_n^2+x_n+1} > 0$, on obtient $-1 < x_n < x_{n+1} < 0$, ce qui donne, par récurrence, les conclusions voulues. La suite converge donc vers un réel ℓ tel que $\ell = \dfrac{\ell^2+\ell}{\ell^2+\ell+1}$, *i.e.* $\ell = 0$.

Posons $y_n := -\dfrac{1}{x_n}$: c'est donc une suite strictement positive et strictement croissante, de limite $+\infty$, et nous voulons estimer sa croissance. La relation de récurrence s'écrit : $y_0 := 2$, $y_{n+1} := \dfrac{y_n^2 - y_n + 1}{y_n - 1}\cdot$

On en tire : $y_{n+1} = y_n + \dfrac{1}{y_n - 1}$, puis $y_{n+1}^2 - y_n^2 = 2 + o(1) \sim 2$, d'où, par le lemme de Cesàro, $y_n^2 \sim 2n$, d'où $y_n \sim \sqrt{2n}$. L'estimation $x_n \sim -\dfrac{1}{\sqrt{2n}}$ signifie que la suite des x_n converge très lentement vers 0.

Théorème 9. Sous les hypothèses $f(a)f(b) < 0$, $f' > 0$ et $f'' > 0$, la méthode de la sécante converge linéairement.

Démonstration. Avec les notations précédentes, posons $\lambda := f'(r) > 0$. On a donc un développement limité : $f(x_n) = \lambda(x_n - r) + \mathrm{o}(x_n - r)$, d'où :

$$x_{n+1} - r = \frac{(x_n - r)f(b) - (b - r)f(x_n)}{f(b) - f(x_n)} = \mu(x_n - r) + \mathrm{o}(x_n - r),$$

où $\mu := \dfrac{f(b) - \lambda(b - r)}{f(b)} \in]0, 1[$, d'après les hypothèses de croissance et de convexité. On a donc $x_{n+1} - r \sim \mu(x_n - r)$, d'où la convergence linéaire (exercice II.8.5 de la page 951).

2.2.2 Rappels et compléments sur le théorème du point fixe

Soit φ une fonction k-lipschitzienne sur un intervalle $I \subset \mathbb{R}$; dans la pratique, φ sera une fonction de classe $\mathcal{C}^1$ telle que $|\varphi'| \leqslant k$. On suppose de plus que $k < 1$ (la fonction φ est donc *contractante*). Soit enfin $[a, b] \subset I$ un segment stable par φ, *i.e.* $\varphi([a, b]) \subset [a, b]$. Il découle alors du *théorème du point fixe* (théorème 39 de la page 427) que φ admet un unique point fixe $c \in [a, b]$, *i.e.* tel que $\varphi(c) = c$. De plus, pour tout $x_0 \in [a, b]$, la suite définie par la relation de récurrence $x_{n+1} := \varphi(x_n)$ converge vers c et l'on a :

$$\forall n \in \mathbb{N}\ ,\ |x_n - c| \leqslant k^n\,|x_0 - c|\,.$$

Dans l'esprit de cette section, on dispose donc d'un algorithme itératif de résolution de l'équation $\varphi(x) = x$, et sa convergence est linéaire. Pratiquement, si l'on ne connaît pas c, l'inégalité ci-dessus est inutilisable pour décider du nombre d'itérations nécessaire pour une précision donnée. Le lemme suivant pallie cet inconvénient.

Lemme 10. On a :

$$\forall n \in \mathbb{N}^*\ ,\ |x_n - c| \leqslant \frac{k^n}{1 - k}\,|x_1 - x_0|\,.$$

Démonstration. De $|x_1 - c|\,k \leqslant k\,|x_0 - c| \leqslant k\,|x_0 - x_1| + k\,|x_1 - c|$, on tire $(1 - k)\,|x_1 - c|\,k \leqslant k\,|x_0 - x_1|$, d'où l'assertion pour $n := 1$.
Comme $|x_{n+1} - c| = |\varphi(x_n) - \varphi(c)| \leqslant k\,|x_n - c|$, le cas général s'en déduit par récurrence.

Exercice 10.

Soient $\alpha > 0$ et $b > \sqrt{\alpha}$. Montrer que la fonction $\varphi(x) := \dfrac{bx + \alpha}{b + x}$ est contractante sur $[0, b]$ et que ce segment est stable. Que peut-on ainsi calculer ?

Solution.

On a $\varphi'(x) = \dfrac{b^2 - \alpha}{(b + x)^2}$, donc $0 < \varphi'(x) \leqslant k$ sur $[0, b]$, où l'on a posé $k := \dfrac{b^2 - \alpha}{b^2} \in]0, 1[$. La fonction φ est contractante. Les points fixes de φ sont $\pm\sqrt{\alpha}$, le

seul point fixe sur $[0,b]$ est donc $\sqrt{\alpha}$. On a en particulier : $|\varphi(x)-\sqrt{\alpha}| \leqslant k\,|x-\sqrt{\alpha}|$.
D'autre part, $\varphi(x)-x = \dfrac{\alpha-x^2}{b+x}$ est positif sur $[0,\sqrt{\alpha}]$ et négatif sur $[\sqrt{\alpha},b]$; sur le premier intervalle, on a donc $0 \leqslant x \leqslant \varphi(x) \leqslant \sqrt{\alpha}$ et, sur le deuxième intervalle, on a de même $b \geqslant x \geqslant \varphi(x) \geqslant \sqrt{\alpha}$. Le segment est donc stable.
Ainsi, pour tout $x_0 \in [0,\sqrt{\alpha}]$, la suite définie par la relation $x_{n+1} := \dfrac{bx_n+\alpha}{b+x_n}$ converge linéairement vers $\sqrt{\alpha}$.

Classification des points fixes

On suppose seulement φ de classe $\mathcal{C}^1$ sur I. Pour tout $x_0 \in I$, on peut considérer la suite définie par la relation de récurrence $x_{n+1} := \varphi(x_n)$ comme un *système dynamique autonome* de condition initiale x_0. Tout point fixe $c \in I$ de φ peut alors être vu comme une position d'équilibre de ce système. Les notions de stabilité introduites pour les équations différentielles ont ici des analogues.
Dans le cas où $|\varphi'(c)| < 1$, on dit que le point fixe c est *attractif*. Il existe $h_0 > 0$ tel que $|\varphi'| \leqslant 1$ sur $[c-h_0, c+h_0]$. Sur cet intervalle, l'inégalité des accroissements finis entraîne $|\varphi(x)-c| \leqslant |x-c|$, et l'intervalle est donc stable : tout système $x_{n+1} := \varphi(x_n)$ de condition initiale $x_0 \in [c-h_0, c+h_0]$ y reste cantonné. C'est l'analogue de la stabilité.
Soit k tel que $|\varphi'(c)| < k < 1$. Il existe alors $h > 0$ tel que $|\varphi'| \leqslant k$ sur $[c-h, c+h]$. Cet intervalle est également stable et φ y est contractante : tout système $x_{n+1} := \varphi(x_n)$ de condition initiale $x_0 \in [c-h, c+h]$ y reste cantonné et converge vers c. C'est l'analogue de la stabilité asymptotique. Dans le cas où $\lambda := \varphi'(c) \neq 0$, on a l'équivalence : $x_{n+1}-c \sim \lambda(x_n-c)$ et l'on trouve une convergence « exactement » linéaire.
Dans le cas où $\varphi'(c) = 0$, on a $x_{n+1}-c = \mathrm{o}(x_n-c)$ et l'on dit que le point fixe c est *super-attractif*. Si par exemple que φ est de classe $\mathcal{C}^2$ sur I, et si $|\varphi''| \leqslant M$ sur $[c-h, c+h]$, alors : $|\varphi(x)-c| \leqslant \dfrac{M}{2}(x-c)^2$. On en déduit que $\dfrac{M}{2}\,|x_n-c| \leqslant \left(\dfrac{M}{2}\,|x_0-c|\right)^{2^n}$, autrement dit, la convergence est quadratique dans ce cas.
Dans le cas où $|\varphi'(c)| > 1$, la suite ne peut converger que si elle est stationnaire ; on dit que le point fixe c est *répulsif*. Pour tout x proche de c, on a $|\varphi(x)-c| > |x-c|$. Dans le cas où $|\varphi'(c)| = 1$, il y a discussion, comme le montrent les exemples ci-dessous.

Exemples. Si $\varphi(x) := \dfrac{x}{x+1}$ sur $I :=]-1,+\infty[$, le point fixe 0 vérifie $\varphi'(0) = 1$.
Comme $\dfrac{1}{\varphi} - \dfrac{1}{x} = 1$, on voit que $x_n = \dfrac{x_0}{x_0+n}$ pour toute condition initiale : il y a stabilité asymptotique (mais convergence lente). De même, la fonction $\varphi(x) := \dfrac{x^2+x}{x^2+x+1}$ (*cf.* l'exercice 9 de la page 900) a 0 pour point fixe asymptotiquement stable avec convergence très lente.

Si $\varphi(x) := x + x^3$ sur $I := \mathbb{R}$, le point fixe 0 vérifie $\varphi'(0) = 1$. Puisque φ est impaire, les conditions initiales x_0 et $-x_0$ donnent lieu à des suites opposées. Supposons donc $x_0 > 0$. On voit d'abord que $x_n > 0$ pour tout n, puis que la suite est strictement croissante. Puisque 0 est le seul point fixe la limite est $+\infty$. Le point fixe 0 est donc instable.
Si $\varphi(x) := x + x^2$ sur $I := \mathbb{R}$, le point fixe 0 vérifie $\varphi'(0) = 1$. Pour une condition initiale $x_0 > 0$, le même argument que ci-dessus montre qu'il y a divergence vers $+\infty$. Si $x_0 < -1$, on a $x_1 > 0$ et l'on retombe sur le premier cas. Si $x_0 := -1$, $x_1 = 0$ qui est fixe et rien ne bouge plus. Si $x_0 \in]-1, 0[$, on vérifie d'abord que tous les x_n restent dans l'intervalle stable $]-1, 0[$. La suite est évidemment croissante ; elle admet donc une limite, qui ne peut être que 0. Comme $\dfrac{1}{x_{n+1}} - \dfrac{1}{x_n} = -\dfrac{1}{x_n + 1} \sim -1$, on voit (lemme de Cesàro) que $x_n \sim -\dfrac{1}{n}$: convergence lente et *conditionnelle* (dépendant de la condition initiale).

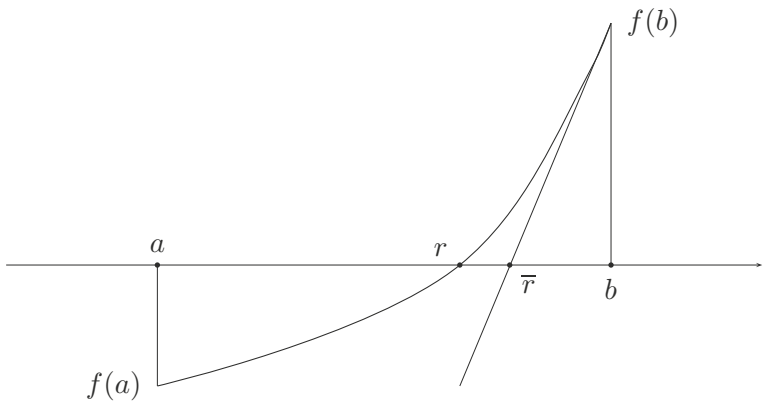
Exercice 11.

Examiner le cas des fonctions $\varphi(x) := \dfrac{xf(b) - bf(x)}{f(b) - f(x)}$ et $\varphi(x) := x - \dfrac{f(x)}{f'(x)}$.

Solution. La première fonction a été rencontrée dans l'étude de la méthode de la sécante (page 899). Sur tout intervalle I où $f(x) \neq f(b)$, tout point fixe r est un zéro de f et l'on a $\varphi'(r) = \dfrac{f(b) - (b - r)f'(r)}{f(b)}$.
Sous les hypothèses de l'étude citée, c'est le coefficient $\lambda \in]0, 1[$ et le point fixe est attractif.
La seconde fonction va intervenir dans l'étude de la méthode de Newton. Sur tout intervalle I où f' ne s'annule pas, tout point fixe r est un zéro de f et l'on a $\varphi' = \dfrac{ff''}{f'^2}$, d'où $\varphi'(r) = 0$: le point fixe est superattractif.

2.2.3 La méthode de Newton

Le théorème du point fixe permet de résoudre une équation de la forme $\varphi(x) = x$. Pour résoudre une équation de la forme $f(x) = 0$, il est donc raisonnable de poser $\varphi(x) := x + f(x)g(x)$ pour une fonction g convenable (par exemple, qui ne s'annule jamais). Soit r un zéro de f. C'est donc un point fixe de φ. L'idéal, au vu de ce qui précède, serait qu'il soit superattractif, *i.e.* $\varphi'(r) = 0$. Mais on a $\varphi'(r) = 1 + f'(r)g(r)$ (puisque $f(r) = 0$), d'où la condition $g(r) = -\dfrac{1}{f'(r)}$ (si possible). Le choix le plus simple est alors de prendre $g := -\dfrac{1}{f'}$, donc $\varphi(x) := x - \dfrac{f(x)}{f'(x)}$.
La *méthode de Newton* consiste à approcher un zéro de f en itérant la fonction φ.
Nous supposerons que f est de classe $\mathcal{C}^2$ sur un intervalle ouvert I, et que $[a, b] \subset I$ est un segment tel que $f(a)f(b) < 0$, et sur lequel $f', f'' > 0$. (Ce sont les mêmes hypothèses que pour la méthode de la sécante). Soit $r \in [a, b]$ l'unique zéro de f.



La tangente au graphe de f en $\big(b, f(b)\big)$ a pour équation

$$y = \Lambda(x) := f'(b)x + f(b) - bf'(b).$$

Elle coupe l'axe des x au point d'abscisse $b - \dfrac{f(b)}{f'(b)} = \varphi(b)$. La méthode de Newton admet donc une interprétation géométrique analogue à la méthode de la sécante : on approche f par l'équation $\Lambda(x)$ de sa tangente en un point et l'on calcule le zéro de Λ, que l'on espère être une meilleure approximation.

Théorème 11. On suppose $f(a)f(b) < 0$ et $f', f'' > 0$ sur $[a, b]$. La suite définie par $x_0 := b$ et $x_{n+1} := x_n - \dfrac{f(x_n)}{f'(x_n)}$ est strictement décroissante et de limite r.
La convergence est quadratique.

Démonstration. On a $\varphi' = \dfrac{ff''}{f'^2}$, donc φ est strictement croissante sur $]r, b]$. Par convexité, on a : $x_1 := b - \dfrac{f(b)}{f'(b)} \in]r, b[$, *i.e.* $r < x_1 < x_0$. En itérant l'application de φ, on trouve $r < x_{n+1} < x_n$, d'où une suite décroissante minorée, de limite un point fixe, qui ne peut être que r.
On a : $\varphi(x) - r = (x - r) - \dfrac{f(x)}{f'(x)}$. Soit M le maximum de f'' sur $[r, b]$. Alors $|f(x) - f(r) - (x - r)f'(x)| \leqslant \dfrac{M}{2}(x - r)^2$. Soit m le minimum de f' sur $[r, b]$. Alors $0 < \varphi(x) - r \leqslant \dfrac{M}{2m}(x - r)^2$, d'où une convergence quadratique. ∎

Exemple. Prenons $f(x) := x^2 - \alpha$. Alors $\varphi(x) = \dfrac{1}{2}\left(x + \dfrac{\alpha}{x}\right)$. Cette méthode de calcul approché de la racine carrée est appelée *algorithme de Babylone*.

Point Méthode

Lorsque la fonction f est connue par des valeurs numériques, il peut être difficile de calculer $f'(x)$. Dans ce cas, on préférera la méthode de la sécante qui n'utilise que les valeurs de f. Notons cependant que l'on peut obtenir des approximations raisonnables des dérivées de f à partir des valeurs connues, par des calculs de *différences finies*. Des exemples figurent dans l'exercice type du module sur l'approximation de [L1], ainsi que dans l'exercice II.8.0 de la page 949.
Même lorsque f' est connue, il peut être avantageux de combiner les deux méthodes : sous les hypothèses communes des théorèmes 9 de la page 901 et 11 de la page ci-contre, l'une fournit une *minoration* et l'autre une *majoration* de r.

3 Interpolation et approximation polynomiales

Nous étudions ici la possibilité d'approcher une fonction continue f sur un segment $[a,b] \subset \mathbb{R}$ par un polynôme P. Le théorème de Weierstraß garantit, en principe, la possibilité d'une approximation *uniforme*. Nous noterons donc $\|-\|_\infty$ la *norme uniforme* (ou *norme du sup*) sur $[a,b]$. Notre problème est de *calculer effectivement* de tels polynômes d'approximation P et de *contrôler l'erreur d'approximation* $\|f-P\|_\infty$.

3.1 Interpolation polynomiale

3.1.1 Rappels et compléments sur l'interpolation de Lagrange

Fixons $(n+1)$ points deux à deux distincts $a_0,\dots,a_n$ de $\mathbb{R}$. On notera, dans toute cette section :

$$\omega_A := \prod_{i=0}^{n}(X-a_i) = \prod_{a\in A}(X-a), \quad \text{où } A := \{a_0,\dots,a_n\}.$$

L'application linéaire $P \mapsto \big(P(a_0),\dots,P(a_n)\big)$ de $\mathbb{R}[X]$ dans $\mathbb{R}^{n+1}$ est surjective et que son noyau est l'idéal principal $<\omega_A>$ (voir le module sur les polynômes de [L1], section 1.4). La restriction de cette application à $\mathbb{R}_n[X]$ est bijective. Nous avons donc un isomorphisme de $\mathbb{R}^{n+1}$ sur $\mathbb{R}_n[X]$ qui, à $(f_0,\dots,f_n) \in \mathbb{R}^{n+1}$, associe l'unique polynôme P de degré inférieur ou égal à n et tel que $P(a_i) = f_i$ pour $i \in [\![0,n]\!]$. Notons, pour $i \in [\![0,n]\!]$:

$$L_i := \prod_{\substack{0\leqslant j\leqslant n\\ j\neq i}} \frac{(X-a_j)}{(a_i-a_j)} = \frac{\omega_A}{(X-a_i)\omega'_A(a_i)}.$$

Alors $L_i(a_j) = \delta_{i,j}$ (symbole de Kronecker), ce qui peut se traduire en disant que, par l'isomorphisme ci-dessus, se correspondent la base canonique de $\mathbb{R}^{n+1}$ et la base $(L_0,\dots,L_n)$ de $\mathbb{R}_n[X]$ formée des polynômes d'interpolation de Lagrange. Le polynôme P ci-dessus est donné par la formule $P = \sum_{i=0}^{n} f_i L_i$.

Nous appliquerons ces généralités au contexte suivant. Les points a_i appartiennent à un

segment non trivial fixé $[a,b]$ de $\mathbb{R}$. À toute fonction f continue sur $[a,b]$, on associe $(f_0,\dots,f_n) := \big(f(a_0),\dots,f(a_n)\big) \in \mathbb{R}^{n+1}$, et le polynôme d'interpolation $P \in \mathbb{R}_n[X]$ correspondant. On dit alors que les a_i sont les *points* ou *nœuds d'interpolation*, que les $f(a_i)$ sont les *valeurs d'interpolation* et que le polynôme P est le *polynôme d'interpolation (de Lagrange) de f aux points a_i*. Ce polynôme est noté $L[A;f]$ ou $L[a_0,\dots,a_n;f]$ (mais il faut garder à l'esprit que l'ordre des points a_i n'a aucune importance). On a donc la formule :

$$L[a_0,\dots,a_n;f] = \sum_{i=0}^{n} f(a_i)L_i.$$

Comme nous avons l'intention d'utiliser $L[A;f]$ pour *approcher* f, il sera utile d'étudier l'*erreur d'approximation*. Naturellement, si la fonction à interpoler est complètement arbitraire, c'est impossible : même la condition de continuité n'impose *aucune* contrainte à la vitesse de variation de f entre les nœuds d'interpolation. Il faudra donc une *hypothèse de régularité* sur f. Nous aurons de plus besoin d'une extension du théorème de Rolle.

Lemme 12. Soit g une fonction de classe $\mathcal{C}^n$ sur un intervalle I contenant $a_0,\dots,a_n$, et telle que $g(a_0) = \cdots = g(a_n) = 0$. Il existe alors $\xi \in I$ tel que $g^{(n)}(\xi) = 0$.

Démonstration. Lorsque $n := 0$, c'est tautologique. Lorsque $n := 1$, c'est le théorème de Rolle. Si $n \geqslant 2$, on applique le théorème de Rolle successivement sur chacun des $[a_i,a_{i+1}]$. Il existe donc des points $b_i \in \,]a_i,a_{i+1}[$ (pour $0 \leqslant i \leqslant n-1$) tels que $g'(b_i) = 0$. La fonction g' est de classe $\mathcal{C}^{n-1}$ et s'annule en les points $b_0 < \cdots < b_{n-1}$, par récurrence on a donc $(g')^{(n-1)}(\xi) = 0$. ∎

Théorème 13. On suppose que f est de classe $\mathcal{C}^{n+1}$ sur $[a,b]$. Pour tout $x \in [a,b]$, il existe $\xi_x \in [a,b]$ tel que :

$$f(x) - L[A;f](x) = \frac{\omega_A(x)}{(n+1)!} f^{(n+1)}(\xi_x). \tag{7}$$

Démonstration. Si x est l'un des a_i, la formule est triviale ; ξ_x est alors arbitraire.

Dans le cas contraire, on considère x comme fixé et l'on introduit une fonction de la variable $t \in [a,b]$:

$$g(t) := \omega_A(x)\big(f(t) - L(t)\big) - \omega_A(t)\big(f(x) - L(x)\big),$$

où l'on a noté, pour simplifier, $L := L[A;f]$. La fonction g est de classe $\mathcal{C}^{n+1}$ sur $[a,b]$ et elle s'annule aux points $a_0,\dots,a_n$ et x.

D'après le lemme, il existe $\xi \in [a,b]$ tel que $g^{(n+1)}(\xi) = 0$ (mais ce ξ dépend de x, d'où la notation ξ_x). Enfin, puisque $\deg L \leqslant n$, on a $L^{(n+1)} = 0$; et puisque ω_A est unitaire de degré $n+1$, on a $\omega_A^{(n+1)} = (n+1)!$. La formule s'ensuit. ∎

Remarques.

1. La démonstration permet en fait d'imposer que ξ_x appartienne au plus petit intervalle qui contient $A \cup \{x\}$.
2. Ces résultats s'étendent au *processus d'interpolation de Lagrange-Hermite*, dans lequel on peut affecter des multiplicités aux nœuds d'interpolation (exercice II.8.15 de la page 952).

Corollaire 14. On a la majoration suivante de l'erreur d'approximation :

$$\|f(x) - L[A; f]\|_\infty \leqslant \frac{1}{(n+1)!} \|\omega_A\|_\infty \left\|f^{(n+1)}\right\|_\infty . \tag{8}$$

La conclusion pratique de ce théorème est que l'erreur d'approximation dépend de deux facteurs indépendants : le facteur $\|\omega_A\|_\infty$ est lié au choix des nœuds d'interpolation, alors que le facteur $\left\|f^{(n+1)}\right\|_\infty$ est lié à la régularité de la fonction f.

Exercice 12.

Si f est la restriction au segment $[a, b]$ de la fonction exponentielle ou d'un polynôme trigonométrique, et si P_n est le polynôme d'interpolation de f en n points distincts arbitraires, alors, lorsque $n \to +\infty$, les P_n convergent vers f uniformément sur $[a, b]$.

Solution.

Dans les deux cas, les dérivées de f vérifient une majoration de la forme $\left\|f^{(k)}\right\|_\infty = \mathrm{O}(M^k)$ pour un certain $M > 0$. La conclusion découle alors de la formule (8).

3.1.2 Le calcul pratique des polynômes d'interpolation

Le calcul direct des coefficients des polynômes L_i et de $L[A; f]$ est assez coûteux (exercice II.8.13 de la page 952). En fait, cela cache un grand nombre de calculs redondants et des algorithmes permettent de réduire ce coût : la *méthode des différences divisées*, accompagnée d'une variante du *schéma de Horner*. Nous noterons $f[a_0, \ldots, a_n]$ le coefficient du monôme de degré n dans $L[a_0, \ldots, a_n; f]$:

$$f[a_0, \ldots, a_n] = \sum_{i=0}^{n} f(a_i) \prod_{\substack{0 \leqslant j \leqslant n \\ j \neq i}} \frac{1}{(a_i - a_j)}. \tag{9}$$

La *différence divisée* $f[a_0, \ldots, a_n]$ ne dépend donc pas de l'ordre des points a_i. Par exemple, $f[a_0] = f(a_0)$ et $f[a_0, a_1] = \dfrac{f(a_1) - f(a_0)}{a_1 - a_0}$.

L'intérêt des différences divisées pour le calcul de $L[A; f]$ repose sur la *formule de Newton* :

$$L[a_0, \ldots, a_n; f] = \sum_{i=0}^{n} f[a_0, \ldots, a_i] \prod_{j=0}^{i-1} (X - a_j). \tag{10}$$

Par convention, tout produit vide valant 1, le premier terme de cette somme est $f[a_0] = f(a_0)$. Pour $n := 0$, cette formule est évidente. Le cas général se prouve par récurrence en remarquant que $L[a_0, \ldots, a_i; f] - L[a_0, \ldots, a_{i-1}; f]$ est nul en $a_0, \ldots, a_{i-1}$, donc multiple

de $(X - a_0)\cdots(X - a_{i-1})$ avec un facteur constant à cause du degré ; et ce facteur est le coefficient du monôme de degré i de $L[a_0, \ldots, a_i; f] - L[a_0, \ldots, a_{i-1}; f]$, c'est-à-dire $f[a_0, \ldots, a_i]$.

Exercice 13.

Démontrer l'existence de ξ tel que : $f[a_0, \ldots, a_i] = \dfrac{1}{i!} f^{(i)}(\xi)$ (s'inspirer du théorème 13 de la page 906). Où peut-on prendre ξ ?

Solution. D'après le théorème invoqué, on a :

$$\begin{aligned} L[a_0, \ldots, a_i; f](a_i) - L[a_0, \ldots, a_{i-1}; f](a_i) &= f(a_i) - L[a_0, \ldots, a_{i-1}; f](a_i) \\ &= \frac{1}{i!} \prod_{j=0}^{i-1} (a_i - a_j) f^{(i)}(\xi); \end{aligned}$$

et l'on vient de voir que :

$$L[a_0, \ldots, a_i; f](a_i) - L[a_0, \ldots, a_{i-1}; f](a_i) = f[a_0, \ldots, a_i] \prod_{j=0}^{i-1} (a_i - a_j),$$

d'où le résultat. D'après les remarques qui suivent le théorème, on peut prendre ξ dans $[\min(a_0, \ldots, a_i), \max(a_0, \ldots, a_i)]$.

Le calcul pratique des différences divisées repose sur la formule suivante :

$$f[a_0, \ldots, a_n] = \frac{f[a_0, \ldots, a_{n-1}] - f[a_1, \ldots, a_n]}{a_0 - a_n}. \tag{11}$$

On la prouve en identifiant les coefficients des monômes de degré n dans la *formule de Neville-Aitken* :

$$(a_0 - a_n) L[a_0, \ldots, a_n; f] = (X - a_n) L[a_0, \ldots, a_{n-1}; f] - (X - a_0) L[a_1, \ldots, a_n; f]. \tag{12}$$

Cette dernière se prouve en remarquant que la différence est un polynôme de degré inférieur ou égal à n et nul en tous les a_i.

La formule (11) permet de calculer récursivement $f[a_0, \ldots, a_n]$ par un tableau de type « triangle de Pascal ». Par exemple, pour $n := 2$, notant **a0** pour a_0, **f0** pour $f[a_0] = f(a_0)$, **f01** pour $f[a_0, a_1]$, etc :

```
f0                          f1                              f2
  f01 := (f0 - f1)/(a0 - a1)     f12 := (f1 - f2)/(a1 - a2)
              f012 := (f01 - f12)/(a0 - a2)
```

De même, pour $n := 5$, le tableau aurait l'allure suivante :

```
          f0        f1        f2        f3        f4        f5
              f01       f12       f23       f34       f45
                  f012      f123      f234      f345
                      f0123     f1234     f2345
                          f01234    f12345
                              f012345
```

Au passage, on a en particulier calculé tous les $f[a_0, \dots, a_i]$. Nous pouvons maintenant calculer $L[A; f]$ par un schéma de type Horner (cela découle de la formule de Newton) :

$$L(A; f) = f[a_0] + (X - a_0)\Big(f[a_0, a_1] + (X - a_1)\big(f[a_0, a_1, a_2] + (X - a_2)(\cdots)\big)\Big).$$

Algorithmiquement, les différences divisées $f[a_i, \dots, a_j]$ (ce sont les seules que nous avons besoin de calculer) peuvent être stockées dans un tableau à deux indices **`F[i,j]`** (seules les cases du tableau telles que $i \leqslant j$ servent donc). En style itératif :

```
pour i de 0 a n faire
   F[i,i] := f(a[i]) ;
pour d de 1 a n faire
   pour i de 1 a n - d faire
      F[i,i+d] := (F[i,i+d-1] - F[i+1,i+d])/(a[i] - a[d]) ;;
```

Puis, en style récursif (pour pratiquer un peu) :

```
Interp(i) := si i = n alors F[0,n]
             sinon F[0,i] + (x - a[i]) * Interp(i+1) ;;
```

C'est, bien entendu, l'appel **`Interp(0)`** qui rend la valeur $L(A; f)(x)$.

3.2 Approximation polynomiale

La qualité de l'approximation obtenue dépend de deux facteurs principaux. La majoration de l'erreur est le premier critère qui vient à l'esprit ; mais il faut également prendre en compte la *stabilité* du procédé. En pratique, la fonction f n'est connue qu'approximativement : les valeurs d'interpolation $f(a_i)$ peuvent provenir de calculs antérieurs tronqués ou de mesures physiques. Il ne faut pas que le calcul de $L[A; f]$ soit trop sensible à de telles variations.

Théorème et définition 15. On munit les espaces vectoriels $\mathcal{C}([a, b], \mathbb{R})$ et $\mathbb{R}_n[X]$ de la norme uniforme sur $[a, b]$, notée, dans les deux cas, $\|-\|_\infty$. L'application linéaire $f \mapsto L[A; f]$ de $\mathcal{C}([a, b], \mathbb{R})$ dans $\mathbb{R}_n[X]$ est continue, et sa norme $\Lambda(A)$ vaut :

$$\Lambda(A) = \left\| \sum_{i=0}^{n} |L_i| \right\|_\infty . \tag{13}$$

On l'appelle *constante de Lebesgue* associée aux points d'interpolation $a_0, \dots, a_n$.

Démonstration. On a, pour tout $x \in [a, b]$:

$$|L[A; f](x)| \leqslant \sum_{i=0}^{n} |f(a_i)|\,|L_i(x)| \leqslant \|f\|_\infty \sum_{i=0}^{n} |L_i(x)|,$$

d'où, en passant à la borne supérieure, l'inégalité $\Lambda(A) \leqslant \left\| \sum_{i=0}^{n} |L_i| \right\|_\infty$. Comme le membre de droite est fini (borne supérieure d'une fonction continue sur un segment), l'application linéaire $f \mapsto L[A; f]$ est bien continue.

Soit $\xi \in [a,b]$ un point qui réalise l'égalité : $\sum_{i=0}^{n} |L_i(\xi)| = \left\| \sum_{i=0}^{n} |L_i| \right\|_\infty$. On peut construire une fonction continue f sur $[a,b]$ telle que $f(a_i) = \operatorname{signe}\big(L_i(\xi)\big)$ pour $i \in [\![0,n]\!]$ et telle que $\|f\|_\infty = 1$: il suffit de prendre f affine par morceaux (son graphe est une ligne brisée joignant les points prescrits d'abscisses a_i et prolongé par des constantes ± 1 à gauche et à droite). On a alors : $\|L[A;f]\| = \|f\|_\infty \left\| \sum_{i=0}^{n} |L_i| \right\|_\infty$, ce qui achève la preuve. ∎

Nous allons maintenant analyser l'erreur et la stabilité pour deux choix des points d'interpolation. Pour une étude générale, indépendante de la fonction f, nous devons nous intéresser à deux nombres : la constante de Lebesgue $\Lambda(A)$ et le facteur $\dfrac{1}{(n+1)!}\|\omega_A\|_\infty$ qui apparaît dans le corollaire 14 de la page 907.

Exercice 14.

On passe du segment $[a,b]$ au segment $[a',b']$ par l'affinité $x' = s(x) = \alpha x + \beta$. Les nœuds d'interpolation a_i ont pour image les nœuds d'interpolation $a'_i = s(a_i)$. Comparer les polynômes L_i, L'_i correspondants, ainsi que les constantes $\Lambda(A)$, $\Lambda(A')$ d'une part, $\|\omega_A\|_\infty$ et $\|\omega_{A'}\|_\infty$ d'autre part, où l'on note $A' := s(A)$.

Solution. On a $L'_i(a'_j) = L'_i(s(a_j)) = \delta_{i,j} = L_i(a_j)$, d'où l'égalité $L_i = L'_i \circ s$. Il en découle que le maximum de $\sum |L_i| = \sum |L'_i \circ s|$ sur $[a,b]$ est égal au maximum de $\sum |L'_i|$ sur $s\big([a,b]\big) = [a',b']$, soit l'égalité $\Lambda(A) = \Lambda(A')$.

Par ailleurs, $\omega_{A'}(\alpha X + \beta) = \prod(\alpha X + \beta - \alpha a_i - \beta) = \alpha^{n+1} \prod(X - a_i) = \alpha^{n+1}\omega_A$, d'où l'on déduit : $\|\omega_{A'}\|_\infty = \alpha^{n+1}\|\omega_A\|_\infty$. Comme $\alpha = \dfrac{b'-a'}{b-a}$, on peut exprimer cela sous la forme : $\dfrac{\|\omega_A\|_\infty}{(b-a)^{n+1}} = \dfrac{\|\omega_{A'}\|_\infty}{(b'-a')^{n+1}}$.

Nous noterons donc $\Omega(A) := \dfrac{\|\omega_A\|_\infty}{(b-a)^{n+1}}$.

Le mode d'emploi de cette constante est le suivant ; on a l'inégalité :

$$\|f(x) - L[A;f]\|_\infty \leqslant \frac{(b-a)^{n+1}}{(n+1)!}\,\Omega(A)\left\|f^{(n+1)}\right\|_\infty. \tag{14}$$

Nous allons donc estimer $\Lambda(A)$ et $\Omega(A)$ dans deux cas intéressants.

3.2.1 Points équidistants

On prend ici $a_i = a + i\dfrac{b-a}{n}$ pour $0 \leqslant i \leqslant n$, ce que nous noterons $a_i = a + ih$, où $h := \dfrac{b-a}{n}$. On notera $f_i := f(a_i)$ les valeurs d'interpolation. L'application affine $t \mapsto a + th$ est bijective de $[0,n]$ sur $[a,b]$ et permet (comme nous venons de le voir) de se ramener à l'interpolation des valeurs f_i aux points $i \in [\![0,n]\!]$. Le calcul des différences divisées se ramène ainsi au calcul des *différences finies*, que nous allons exposer.

À partir de toute suite finie $(x_i)_{0\leqslant i\leqslant N}$, on pose :

$$\forall i \in [\![0, N-1]\!]\ ,\ (\Delta x)_i := x_{i+1} - x_i.$$

Pour simplifier l'écriture, on écrira $(\Delta x)_i = \Delta x_i$, mais cela ne doit pas laisser croire que l'opérateur Δ a été appliqué au terme isolé x_i. On a ainsi défini une suite finie $(\Delta x_i)_{0\leqslant i\leqslant N-1}$.

Ce calcul peut s'itérer : si $y_i := (\Delta x)_i$, alors $(\Delta y)_i = y_{i+1} - y_i = x_{i+2} - 2x_{i+1} + x_i$ pour $0 \leqslant i \leqslant N-2$. On notera $\Delta^2 x_i = \Delta y_i$. Plus généralement, on prouve la formule suivante par récurrence ; pour tout $k \leqslant N$:

$$\forall i \in [\![0, N-k]\!]\ ,\ \Delta^k x_i = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} x_{i+j}.$$

En pratique, les différences finies itérées se calculent par un tableau de type « triangle de Pascal ». On retrouve les différences divisées et le polynôme d'interpolation à l'aide des formules suivantes, où $t := \dfrac{x-a}{h}$:

$$f[a_i, \dots, a_{i+k}] = \frac{1}{k!} \frac{\Delta^k f_i}{h^k},$$

$$(x-a_0)\cdots(x-a_{k-1}) = h^k\, t(t-1)\cdots(t-k+1),$$

$$L[A; f] = \sum_{k=0}^{n} \Delta^k f_0 \binom{t}{k}.$$

Nous allons évaluer l'erreur d'interpolation.

On vérifie d'abord que le maximum de $|t(t-1)\cdots(t-n)|$ est inférieur ou égal à $n!$ sur $[0, n]$. Posons en effet $\varphi(t) := |t(t-1)\cdots(t-n)|$. Alors $\varphi(n-t) = \varphi(t)$, et l'on peut restreindre l'étude à $[0, n/2]$. Puis, pour $1 < t \leqslant n/2$, on a $\dfrac{\varphi(t)}{\varphi(t-1)} = \dfrac{t}{n+1-t} < 1$, et le maximum est donc pris sur $[0, 1]$; et là, la majoration est évidente. On en déduit que $\Omega(A) \leqslant \dfrac{n!}{n^{n+1}}$.

Rappelons que, d'après la formule de Stirling (section 1.4) :

$$n! \sim (n/e)^n \sqrt{2\pi n} \qquad \text{d'où} \qquad \frac{n!}{n^{n+1}} \sim e^{-n} \sqrt{2\pi/n}.$$

Nous allons voir que notre majoration n'est pas très mauvaise.

Lemme 16. On a l'égalité et l'équivalence :

$$\binom{1/2}{n+1} = (-1)^n 2^{-2n-1} \frac{1}{n+1} \binom{2n}{n} \sim \frac{1}{2\sqrt{\pi n^3}}.$$

Démonstration. Les deux formules ont été prouvées à la section 4.2.5. ∎

Le maximum de $|\omega|$ est supérieur ou égal à $\left|(n+1)!\binom{1/2}{n+1}\right| \sim \frac{n!}{2\sqrt{\pi n}}$, ce qui n'est pas si loin du majorant $n!$ retenu plus haut. Pour en revenir à la situation de départ, nous retiendrons la majoration :

$$\|f(x) - L[A;f]\|_\infty \leqslant \frac{1}{(n+1)} \left(\frac{b-a}{n}\right)^{n+1} \left\|f^{(n+1)}\right\|_\infty .$$

Nous allons maintenant étudier la stabilité. Notons L_i les polynômes de Lagrange associés aux a_i et ℓ_i ceux associés aux $i \in [\![0,n]\!]$: on a donc $L_i(x) = \ell_i(t)$, où $t := \frac{x-a}{h}$. On peut démontrer que la constante de Lebesgue est ici de l'ordre de grandeur suivant : $\Lambda(A) \sim \frac{2^{n+1}}{en\ln n}$. Cette quantité croît très rapidement avec n : en augmentant le nombre de points d'interpolation, on améliore peut-être la précision mais au prix d'une grande instabilité. Nous ne prouverons pas l'équivalence ci-dessus, mais seulement une minoration suffisamment probante. Il est clair que $\Lambda(A) \geqslant \sum_{i=0}^{n} |\ell_i(1/2)|$, et nous allons donc estimer le membre de droite. On a :

$$l_i(t) = \prod_{\substack{0\leqslant j\leqslant n\\ j\neq i}} \frac{t-j}{i-j} = \frac{(-1)^{n-i}}{i!(n-i)!} \prod_{\substack{0\leqslant j\leqslant n\\ j\neq i}} (t-j) = (-1)^{n-i}\binom{n}{i}\frac{n+1}{t-i}\binom{t}{n+1}.$$

On déduit du lemme 16 de la page précédente l'égalité

$$l_i(1/2) = (-1)^{i-1}\binom{n}{i}\frac{1}{2i-1}2^{-2n}\binom{2n}{n}$$

puis l'inégalité $|\ell_i(1/2)| \geqslant \frac{1}{2n}2^{-2n}\binom{2n}{n}\binom{n}{i}$, et, en sommant : $\Lambda(A) \geqslant \frac{1}{2n}2^{-n}\binom{2n}{n}$. Comme la formule de Stirling implique que $\binom{2n}{n} \sim \frac{2^{2n}}{\sqrt{\pi n}}$, on confirme la désastreuse croissance de l'instabilité.

3.2.2 Points de Tchebychev

Rappelons (module sur les polynômes de [L1]) que le polynôme de Tchebychev T_{n+1} est caractérisé par la formule $T_{n+1}(\cos\theta) = \cos(n+1)\theta$. Il admet la factorisation :

$$T_{n+1}(X) = 2^n \prod_{i=0}^{n} \left(X - \cos\frac{2i+1}{2n+2}\pi\right).$$

Nous considérerons l'interpolation aux $(n+1)$ points $a_i := \cos\frac{2i+1}{2n+2}\pi$ sur le segment $[-1,1]$ (le cas d'un autre segment s'y ramenant par transformation affine).
On commence par évaluer l'erreur d'interpolation. Avec nos notations habituelles, on a donc $\omega = 2^{-n}T_{n+1}$. Par construction, $\|T_{n+1}\|_\infty = 1$, d'où $\|\omega\|_\infty = 2^{-n}$ et $\Omega(A) = 4^{-n}$. C'est à comparer avec les estimations du cas des points équidistants, qui étaient d'un ordre

compris entre $e^{-n}\sqrt{2\pi/n}$ et $e^{-n}\sqrt{2\pi/n^3}$. Comme $e/4 \simeq 0{,}68 < 1$, l'usage des points de Tchebychev donne une bien meilleure précision.

En ce qui concerne la stabilité, on peut prouver l'estimation suivante : $\Lambda(A) \sim \dfrac{2}{\pi}\ln n$. Nous l'admettrons. Comme dans le cas des points équidistants, l'instabilité croît avec le nombre de points d'interpolations, mais la vitesse de croissance est sans commune mesure : là, de l'ordre de grandeur de 4^n, ici, de l'ordre de grandeur de $\ln n$. L'avantage va sans conteste à l'interpolation aux points de Tchebychev.

3.2.3 Le phénomène de Runge

Même pour des fonctions très régulières, il ne va pas de soi que la précision s'améliore uniformément lorsque l'on augmente le nombre des points d'interpolation. Nous illustrerons ce fait (le *phénomène de Runge*) en interpolant la fonction $f(x) := \dfrac{1}{x^2+\alpha^2}$ sur le segment $[-1,1]$, aux $n := 2m$ points $\pm\dfrac{2i+1}{2m}$, où $0 \leqslant i \leqslant m-1$ (nous ne cherchons pas à indexer ces points de 0 à $n-1$). La constante $\alpha \in \mathbb{R}_+^*$ sera fixée ultérieurement. On a ici :

$$\omega = \prod_{i=0}^{m-1}\left(X - \frac{2i+1}{2m}\right)\left(X + \frac{2i+1}{2m}\right) = \prod_{i=0}^{m-1}\left(X^2 - \left(\frac{2i+1}{2m}\right)^2\right).$$

Notons L le polynôme d'interpolation de f en ces points. Il est de degré au plus $2m-1$. Comme le polynôme $L(-X)$ prend les mêmes valeurs aux mêmes points, il est égal à L, qui est pair (cela a lieu chaque fois que l'on interpole une fonction paire en un ensemble de points symétrique par rapport à l'origine). Ainsi, $\deg L \leqslant 2m-2$. Le polynôme $1-(X^2+\alpha^2)L$ est de degré inférieur ou égal à $2m$ et s'annule en chacun des points d'interpolation. Il est donc de la forme $C\omega$, où C est une constante. Par évaluation en $\alpha\mathrm{i}$, on voit que $C = \dfrac{1}{\omega(\alpha\mathrm{i})}$, d'où la formule : $1-(X^2+\alpha^2)L = \dfrac{\omega}{\omega(\alpha\mathrm{i})}$.

Notre but est de montrer que, si α est bien choisi, $f(1)-L(1) = \dfrac{1}{1+\alpha^2}\dfrac{\omega(1)}{\omega(\alpha\mathrm{i})}$ diverge lorsque m tend vers $+\infty$. Nous allons estimer séparément $\omega(1)$ et $\omega(\alpha\mathrm{i})$. On a d'abord :

$$\begin{aligned}\omega(1) &= \prod_{i=0}^{m-1}\left(1 - \frac{2i+1}{2m}\right)\left(1 + \frac{2i+1}{2m}\right) = (2m)^{-2m}\prod_{j=0}^{2m-1}(2j+1)\\ &= (2m)^{-2m}\frac{(4m)!}{2^{2m}(2m)!} = \frac{(2n)!}{n^n 2^n n!},\end{aligned}$$

d'où l'on tire, à l'aide de la formule de Stirling, l'équivalence : $\omega(1) \sim \left(\dfrac{2}{e}\right)^n\sqrt{2}$. On en déduit l'estimation : $\ln\omega(1) = n(\ln 2 - 1) + \mathrm{o}(1)$, d'où $\ln\omega(1) \sim n(\ln 2 - 1)$. Cette équivalence ne requiert d'ailleurs pas la formule de Stirling, mais seulement l'évaluation plus fa-

cile $\ln n! \sim n \ln n$. D'autre part, on a $\omega(\alpha\mathrm{i}) = \prod_{i=0}^{m-1} \left(-\alpha^2 - \left(\frac{2i+1}{2m}\right)^2\right)$, d'où l'on tire :

$$\ln|\omega(\alpha\mathrm{i})| = \sum_{i=0}^{m-1} \ln\left(\alpha^2 + \left(\frac{2i+1}{2m}\right)^2\right)$$
$$= \sum_{j=1}^{2m} \ln\left(\alpha^2 + \left(\frac{j}{2m}\right)^2\right) - \sum_{i=1}^{m} \ln\left(\alpha^2 + \left(\frac{2j}{2m}\right)^2\right) = S_{2m} - S_m,$$

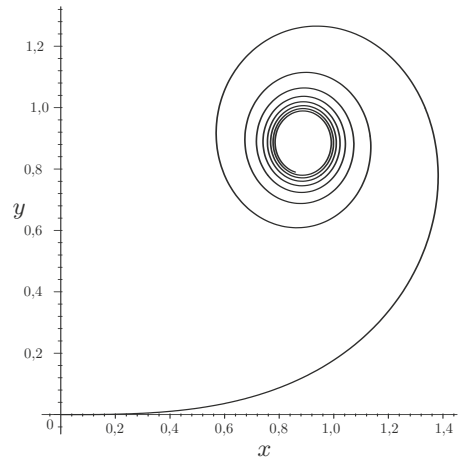
où l'on a posé $S_m := \sum_{i=1}^{m} \ln\left(\alpha^2 + \left(\frac{j}{m}\right)^2\right)$. On reconnaît en $\frac{1}{m}S_m$ une somme de Riemann, de limite (lorsque $m \to +\infty$) l'intégrale : $I_\alpha := \int_0^1 \ln(\alpha^2 + t^2)\,\mathrm{d}t$. On a donc $S_m = I_\alpha m + \mathrm{o}(m)$, d'où $\ln|\omega(\alpha\mathrm{i})| = I_\alpha m + \mathrm{o}(m)$.
Si l'on arrive à déterminer une réel $\alpha > 0$ tel que $2(\ln 2 - 1) > I_\alpha$, on aura donc $\ln\omega(1) - \ln|\omega(\alpha\mathrm{i})| \to +\infty$ lorsque $m \to +\infty$, d'où $|f(1) - L(1)| \to +\infty$. l'existence d'un tel α est laissée en exercice au lecteur (exercice II.8.14 de la page 952).
Pour expliquer ce phénomène, par opposition à ce que donne l'exercice 12 de la page 907, on constate qu'il s'agit dans les deux cas d'approcher des fonctions développables en séries entières. Mais les rayons de convergence de celles de l'exercice étaient infinis, ce que traduisait l'estimation uniforme $\left\|f^{(k)}\right\|_\infty = \mathrm{O}(M^k)$: il s'agissait de fonctions analytiques prolongeables au plan complexe (donc de fonctions *entières*). Au contraire, la fonction $\frac{1}{1+x^2}$ a une barrière en $\pm\mathrm{i}$, son développement en série entière s'en ressent (rayon de convergence 1) et la croissance de ses dérivées le traduit. On peut dire que le comportement chaotique des P_n en 1 traduit une sensibilité à la singularité en $\pm\mathrm{i}$.

4 Calcul approché d'intégrales

Par définition, le calcul de $\int_a^b f(t)\,\mathrm{d}t$ est une opération transcendante : une infinité de calculs sont nécessaires, et un passage à la limite. Dans des cas très particuliers, on sait exprimer une primitive explicite de f en termes de « fonctions élémentaires » et le calcul de l'intégrale se ramène, via des calculs algébriques, aux tables de valeurs de ces fonctions. Ainsi, $\int_2^3 \frac{\mathrm{d}t}{t^2-1} = \frac{1}{2}\ln\frac{3}{2}$, ce que l'on peut évaluer à l'aide d'une table des valeurs de la fonction logarithme népérien. En revanche, aucune expression de ce type n'est connue pour les primitives de e^{-x^2}, et l'*on peut même démontrer qu'il n'en existe pas* (c'est encore un résultat de la théorie de Galois différentielle, évoquée dans le module II.7 à propos de l'équation de Liouville). Cependant, même dans ce cas, certaines intégrales particulières peuvent être exactement évaluées. Ainsi, l'*intégrale de Gauß* : $\int_0^{+\infty} e^{-t^2}\,\mathrm{d}t = \frac{1}{2}\sqrt{\pi}$.
Toutes ces situations sont exceptionnelles. Le cas général est celui où l'on n'a d'autre choix

que d'*évaluer numériquement* l'intégrale définie, autrement dit, d'en donner une *valeur approchée*. Cependant, il faut garder à l'esprit que des méthodes très puissantes permettent le calcul exact lorsqu'il est possible et que beaucoup de ces méthodes sont incorporées dans les systèmes de calcul formel.

Exemple. La « courbe de l'autoroute », ou *clothoïde* (voir page 833) est la solution d'un système différentiel que Maple sait résoudre numériquement et graphiquement. Mais il est difficile, par ce biais, de décider s'il y a un *foyer* (point limite) ou un *cycle limite*. En revanche, Maple sait intégrer *formellement* l'équation en calculant les intégrales : $x_\infty := \displaystyle\int_0^{+\infty} \cos\left(\frac{\sigma^2}{2}\right) \mathrm{d}\sigma = \frac{\sqrt{\pi}}{2}$ et $y_\infty := \displaystyle\int_0^{+\infty} \sin\left(\frac{\sigma^2}{2}\right) \mathrm{d}\sigma = \frac{\sqrt{\pi}}{2}$ (voir les exercices 21 de la page 623 et II.3.30 de la page 656).



Un peu d'Histoire

Pour nous, le calcul d'une aire ou d'une longueur se ramène à une intégration, et c'est, par essence, une opération transcendante. Les mathématiciens Grecs, en particulier Archimède et Eudoxe, savaient calculer certaines aires et longueurs par des méthodes purement géométriques et algébriques. De tels résultats portaient respectivement le nom de *quadratures* et de *rectifications*. Le premier terme est justifié par le problème suivant : comment construire un carré dont l'aire est celle d'une figure donnée. Ainsi, la quadrature d'un cercle de rayon 1 reviendrait à la construction d'un carré de côté $\sqrt{\pi}$. Est-ce possible avec une règle et un compas ? Le mathématicien Wantzel a explicité, en 1837, les conditions pour qu'une grandeur soit ainsi constructible ; nous en avons donné une idée dans le module sur les nombres complexes de [L1] et la formulation générale est précisée dans [MPA1]. Une condition *nécessaire* serait que π soit algébrique ; mais Lindemann a prouvé en 1882 que π était transcendant, confirmant ainsi la conviction des mathématiciens (depuis le XVIII$^{\text{e}}$ siècle) que la quadrature du cercle était impossible.

4.1 Première approche

La première méthode pour calculer numériquement une intégrale est de partir de sa définition. Ainsi, on peut partir d'une somme de Riemann.

Exemple. Pour calculer $\displaystyle\int_0^a t^p \,\mathrm{d}t$, où $p \in \mathbb{N}^*$, on écrit :

$$\int_0^a t^p \,\mathrm{d}t = \lim_{n\to+\infty} \frac{a}{n} \sum_{i=1}^{p} \left(\frac{ia}{n}\right)^p = \lim_{n\to+\infty} \frac{a^{p+1}}{n^{p+1}} s_p(n),$$

où $s_p(n) := \displaystyle\sum_{i=1}^{p} i^p$. Mais on a vu dans le module sur les polynômes de [L1] (exercice II.6.2) qu'il existe un polynôme P_p de terme dominant $\dfrac{1}{p+1}X^{p+1}$ tel que $s_p(n) = P_p(n)$. Il en découle immédiatement que $\displaystyle\int_0^a t^p \,\mathrm{d}t = \frac{a^{p+1}}{p+1}$.

Exercice 15.

Calculer $\displaystyle\int_a^b t^p \,\mathrm{d}t$ à l'aide de la subdivision des aq^i, où $q := \sqrt[n]{\dfrac{b}{a}}$ (méthode due à Fermat).

Solution.

Notant $a_i := aq^i$, on introduit donc les sommes de Riemann $S_n := \displaystyle\sum_{i=0}^{n-1} (a_{i+1} - a_i) f(a_i)$ et l'on aura $\displaystyle\int_a^b t^p \,\mathrm{d}t = \lim_{n\to+\infty} S_n$.

Voici le calcul de S_n :

$$S_n = \sum_{i=0}^{n-1} (aq^{i+1} - aq^i)(aq^i)^p = a^{p+1}(q-1)\frac{1-q^{(p+1)n}}{1-q^{p+1}} = \frac{1-q}{1-q^{p+1}}(b^{p+1} - a^{p+1}),$$

puisque $q^{(p+1)n} = \dfrac{b^{p+1}}{a^{p+1}}$.

Ici, le calcul est très simple car on a une suite géométrique : cela nous rappelle que les subdivisions uniformes ne sont pas les seules. Lorsque $n \to +\infty$, on a $q \to 1$, et $\dfrac{1-q}{1-q^{p+1}} \to \dfrac{1}{p+1}$, d'où enfin : $\displaystyle\int_a^b t^p \,\mathrm{d}t = \frac{b^{p+1} - a^{p+1}}{p+1}$.

Le choix d'un pas variable au lieu d'un pas constant évoque la méthode d'Euler de résolution numérique des équations différentielles : nous y reviendrons. Ici, le pas a été choisi pour simplifier les calculs. En principe c'est plutôt la performance numérique qui dicte le choix.

Exercice 16.
À quelle vitesse converge le calcul ci-dessus ?

Solution. Il s'agit d'évaluer à quelle vitesse a lieu la convergence de $\dfrac{1-q}{1-q^{p+1}}$ vers $\dfrac{1}{p+1}$.

Posons $f(x) := x^{\frac{1}{p+1}}$. De la formule de Taylor à l'ordre 2, on tire l'équivalence :

$$\frac{f(x)-f(1)}{x-1} - f'(1) \sim \frac{x-1}{2} f''(1), \text{ qui donne, ici : } \frac{1-q}{1-q^{p+1}} - \frac{1}{p+1} \sim \frac{p\left(1-\sqrt[n]{c^{p+1}}\right)}{2(p+1)^2},$$

où l'on a posé $c := \dfrac{b}{a}$. Enfin, lorsque $n \to +\infty$, on a $\sqrt[n]{c} - 1 \sim \dfrac{\ln c}{n}$, d'où une erreur de l'ordre de $\dfrac{-p \ln c}{2(p+1)n}$, ce qui est très lent !

4.1.1 La méthode d'Euler

Supposons f continue sur $[a,b]$. La primitive F de f telle que $f(a) = y_0$ est la solution du problème de Cauchy (module II.7) : $\begin{cases} Y' = f(t), \\ Y(a) = y_0. \end{cases}$

Le membre droit de la première égalité ne dépend que de t et pas de Y : c'est le contraire d'un système autonome ! Alors $F(b) - y_0 = \displaystyle\int_a^b f(t)\,\mathrm{d}t$. Ainsi, la méthode d'Euler de résolution approchée des équations différentielles peut-elle servir à calculer des primitives, et donc également des intégrales définies. On est ainsi conduit à introduire une subdivision $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$, de pas $h := \max\limits_{0 \leqslant i \leqslant n-1} h_i$, où $h_i := x_{i+1} - x_i$.

Le calcul est alors le suivant : $\overline{F}_0 := y_0$ puis, pour $0 \leqslant i \leqslant n-1$, $\overline{F}_{i+1} := \overline{F}_i + h_i f(x_i)$. On obtient les valeurs approchées de la primitive : $\overline{F}_i \approx F(x_i)$ et de l'intégrale :

$$\int_a^b f(t)\,\mathrm{d}t \approx \sum_{i=0}^{n-1} h_i f(x_i).$$

Comment évaluer la précision de cette méthode ? Pour une fonction continue quelconque, *il est impossible de majorer l'erreur commise pour une subdivision donnée* : en effet, la donnée des valeurs de f en les x_i ne dit rien des valeurs ailleurs. Tout au plus peut-on espérer prouver que, *lorsque le pas h tend vers* 0, l'erreur commise tend vers 0. Au contraire, lorsque f est au moins de classe $\mathcal{C}^1$, on peut quantifier ces estimations. Par ailleurs, il faut distinguer l'erreur commise « par principe » dans la méthode d'Euler en approchant $F(x_{i+1})$ par $F(x_i) + h_i f(x_i)$ (formule de Taylor au premier ordre), erreur qui s'apparente à une erreur de troncature, de l'erreur d'arrondi dans les données (par exemple y_0) et les calculs intermédiaires.

L'*erreur de consistance* e_i à la $i^{ème}$ étape est définie par la formule suivante :

$$e_i := \big(F(x_{i+1}) - \overline{F}_{i+1}\big) - \big(F(x_i) - \overline{F}_i\big) = F(x_{i+1}) - F(x_i) - h_i f(x_i).$$

Les numériciens disent que c'est « l'erreur sur $\overline{F}_{i+1}$ en supposant $\overline{F}_i$ exact ». Par exemple, si f est de classe $\mathcal{C}^1$, donc F de classe $\mathcal{C}^2$, la formule de Taylor à l'ordre 2 pour $F(x_{i+1})$ donne : $e_i = \dfrac{1}{2}h_i^2 f'(x_i) + \mathrm{o}(h_i^2)$.

Proposition et définition 17. Lorsque le pas h tend vers 0, l'erreur de consistance totale $\displaystyle\sum_{i=0}^{n-1} |e_i|$ tend vers 0. On dit que *la méthode d'Euler est consistante*.

Démonstration.
Soit $\varepsilon > 0$. La fonction f étant continue sur le segment $[a,b]$, elle y est uniformément continue ; il existe donc $\delta > 0$ tel que $|x' - x| < \delta \Rightarrow |f(x') - f(x)| < \dfrac{\varepsilon}{b-a}$.
On suppose le pas $h < \delta$. Du théorème des accroissements finis, on déduit l'existence de $c_i \in]x_i, x_{i+1}[$ tel que $e_i = F(x_{i+1}) - F(x_i) - h_i f(x_i) = h_i\big(f(c_i) - f(x_i)\big)$, d'où $|e_i| \leqslant h_i \dfrac{\varepsilon}{b-a}$. En sommant, on trouve : $\displaystyle\sum_{i=0}^{n-1} |e_i| \leqslant \sum_{i=0}^{n-1} h_i \frac{\varepsilon}{b-a} = \varepsilon$. ∎

En pratique, on ne connaît qu'une valeur approchée $\tilde{F}_0 = \tilde{y}_0 = y_0 + \varepsilon_0$ de la condition initiale y_0 (l'erreur ε_0 étant de signe inconnu). De plus, chaque calcul intermédiaire donne lieu à une erreur d'arrondi, et l'on obtient une valeur approchée $\tilde{F}_{i+1} = \tilde{F}_i + h_i f(x_i) + \varepsilon_{i+1}$ (l'erreur ε_i étant de signe inconnu). L'estimation suivante est alors triviale :

Proposition et définition 18.
L'erreur finale admet la majoration : $\left|\tilde{F}_n - F_n\right| \leqslant S \displaystyle\sum_{i=0}^{n} |\varepsilon_i|$, avec $S = 1$.
On dit que *la méthode d'Euler est stable* et que la *constante de stabilité* S vaut 1.

La dernière propriété est un corollaire des deux précédentes :

Proposition et définition 19. On suppose la condition initiale approchée à ε_0 près, mais les calculs exacts. Lorsque $\max(|\varepsilon_0|, h)$ tend vers 0, l'erreur finale $\left|\tilde{F}_n - F_n\right|$ tend vers 0. On dit que *la méthode d'Euler est convergente*.

Supposons maintenant que f est de classe $\mathcal{C}^1$. L'estimation de l'erreur de consistance $e_i = \dfrac{1}{2}h_i^2 f'(x_i) + \mathrm{o}(h_i^2)$ suggère que, dans les régions où $|f'|$ est grand (et où f varie vite), il faut diminuer les pas intermédiaires h_i pour conserver une précision raisonnable des calculs. Ainsi, si l'on vise une erreur de l'ordre de $\varepsilon > 0$, on tentera d'obtenir une erreur de consistance $|e_i|$ majorée par $\delta := \dfrac{\varepsilon}{b-a}$ en prenant $\left|\dfrac{1}{2}h_i^2 f'(x_i)\right| \leqslant \delta$. En pratique, on ne connaît probablement pas la valeur de $f'(x_i)$, mais on peut l'approcher par $\dfrac{f(x_{i+1}) - f(x_i)}{h_i}$.

On pose donc $e_i^* := \frac{1}{2}\big(f(x_{i+1}) - f(x_i)\big)$ (quantité associée au pas $h_i = x_{i+1} - x_i$) et l'on choisit le prochain pas selon les règles suivantes :

1. Si $\dfrac{\delta}{3} \leqslant \dfrac{|e_i^*|}{h_i} \leqslant \delta$, alors $h_{i+1} := h_i$.
2. Si $\dfrac{|e_i^*|}{h_i} < \dfrac{\delta}{3}$, alors $h_{i+1} := 1{,}25\, h_i$.
3. Si $\dfrac{|e_i^*|}{h_i} > \delta$, alors $h_{i+1} := 0{,}8\, h_i$.

En réalité, le choix du pas est encore affecté par deux facteurs : il sera majoré par un maximum h_{max} dicté par l'idée que l'on se fait *a priori* de la finesse souhaitable de la subdivision ; et il sera minoré par un minimum h_{min} dépendant de la précision autorisée par la machine et du coût tolérable des calculs (plus le pas est petit, plus il y a de calculs). Dans la deuxième règle, on borne de toutes façons h_{i+1} à h_{max}. Dans la troisième règle, si l'on doit passer en dessous de h_{min}, on peut être conduit à interrompre les calculs. Par exemple, à l'approche d'une singularité, les calculs « divergent ».

4.2 Méthodes de quadratures

Dans la méthode d'Euler, on obtient une formule : $\displaystyle\int_a^b f(t)\,\mathrm{d}t \approx \sum h_i f(x_i)$, que l'on peut encore écrire $(b-a)\sum w_i f(x_i)$, où $w_i := \dfrac{h_i}{b-a}$, ce qui est une formule de « moyenne » :

$$\frac{1}{b-a}\int_a^b f(t)\,\mathrm{d}t \approx \sum w_i f(x_i), \text{ avec } \sum w_i = 1 \quad \text{et} \quad \forall i\ ,\ w_i \geqslant 0.$$

Notons que le signe positif des *poids* w_i n'est pas indifférent : il est intervenu dans la preuve de la stabilité. Supposons en effet que l'on effectue un calcul comme le précédent, mais avec des erreurs d'arrondis : $\tilde{f}_i - f(x_i) = \varepsilon_i$. Les signes de ces erreurs sont, bien entendu, inconnus ; mais on peut les supposer majorées : $\forall i\ ,\ |\varepsilon_i| \leqslant \varepsilon$. L'erreur totale est alors majorée :

$$\Big|\sum w_i\big(\tilde{f}_i - f(x_i)\big)\Big| \leqslant \varepsilon \sum |w_i|\,.$$

Si tous les w_i sont positifs, le majorant est $\varepsilon \sum w_i = \varepsilon$, ce qui est optimal : c'est l'erreur que l'on ferait en intégrant une fonction constante dont la valeur est connue à ε près. Mais, si certains des w_i sont strictement négatifs, $\sum |w_i| > 1$. On évitera donc les poids négatifs.

4.2.1 Quadratures élémentaires et quadratures composées

On appellera *(méthode de) quadrature élémentaire* ou *simple* (l'origine du terme « quadrature » est expliquée page 915) une formule de la forme :

$$\int_a^b f(t)\,\mathrm{d}t \approx (b-a)\sum_{i=0}^{n} w_i f(x_i), \tag{15}$$

avec des points x_i tels que $a \leqslant x_0 < \cdots < x_n \leqslant b$ et des poids w_i tels que $\sum_{i=0}^{n} w_i = 1$ et, *de préférence*, $\forall i \in [\![1, n]\!]\ ,\ w_i \geqslant 0$. Notons que, par définition, cette formule est exacte lorsque f est constante (aux erreurs d'arrondi près - ce dont nous ne discuterons plus).

Définition 2. On dit que la méthode de quadrature élémentaire donnée par la formule (15) est *d'ordre supérieur ou égal à* N si la formule est exacte pour toute fonction f polynomiale de degré inférieur ou égal à N ; et qu'elle est *exactement d'ordre* N si, de plus, elle est inexacte pour au moins une fonction polynomiale f de degré $N+1$.

Comme par hypothèse $\sum_{i=0}^{n} w_i = 1$, tout méthode est au moins d'ordre 0.

Proposition 20. Soient des points x_i tels que $a \leqslant x_0 < \cdots < x_n \leqslant b$. Soient L_i les polynômes d'interpolation de Lagrange correspondants. Pour que la méthode (15) soit d'ordre supérieur ou égal à n, il faut, et il suffit, que les poids w_i soient donnés par la formule :

$$w_i := \frac{1}{b-a} \int_a^b L_i(t)\,\mathrm{d}t.$$

Démonstration. Si la méthode est d'ordre supérieur ou égal à n, elle donne une égalité pour chaque L_i, ce qui implique immédiatement la formule ci-dessus.

Réciproquement, définissons les w_i par cette formule. De l'égalité $\sum_{i=0}^{n} L_i = 1$ (obtenue en interpolant la fonction constante 1 aux points x_i), on déduit en intégrant que $\sum_{i=0}^{n} w_i = 1$.

On a donc bien une méthode de quadrature. Par construction, cette méthode est exacte pour tous les polynômes L_i, donc, par linéarité, pour tous les polynômes de $\mathbb{R}_n[X]$, dont les L_i forment une base. ∎

En revanche, il n'est pas assuré que les w_i soient positifs. Cependant, *nous ne considérerons que des méthodes de quadrature élémentaire dont les poids sont calculés de cette manière*.

Corollaire 21. L'intégrale approchée se calcule en intégrant le polynôme d'interpolation ; en notant $A := \{x_0, \ldots, x_n\}$:

$$\int_a^b f(t)\,\mathrm{d}t \approx (b-a) \sum_{i=0}^{n} w_i f(x_i) = \int_a^b L(A; f)(t)\,\mathrm{d}t. \tag{16}$$

Démonstration. Il suffit d'intégrer l'égalité $L(A; f) = \sum_{i=0}^{n} f(x_i) L_i$. ∎

Exemples. Si $n := 0$, il y a un seul point x_0 et un seul poids $w_0 := 1$.

- Pour $x_0 := a$, on a la formule $\int_a^b f(t)\,\mathrm{d}t \approx (b-a) f(a)$ (*méthode des rectangles à gauche*).

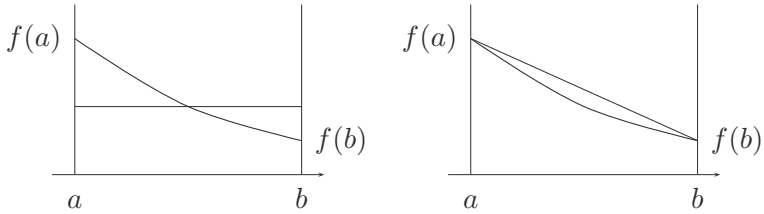
- Pour $x_0 := b$, on a la formule $\displaystyle\int_a^b f(t)\,\mathrm{d}t \approx (b-a)f(b)$ (*méthode des rectangles à droite*).

En prenant $f(x) := x$, on a $\displaystyle\int_a^b t\,\mathrm{d}t = \frac{b^2-a^2}{2}$ différent de $\neq a(b-a)$ et de $b(b-a)$. On voit qu'aucune des deux formules n'est exacte : ce sont des méthodes d'ordre 0 exactement.

- Pour $x_0 := \dfrac{a+b}{2}$, on a la formule $\displaystyle\int_a^b f(t)\,\mathrm{d}t \approx (b-a)f\left(\frac{a+b}{2}\right)$ (*méthode du point milieu*).

  Il y a égalité pour $f(x) := x$ car $\dfrac{b^2-a^2}{2} = (b-a)\dfrac{a+b}{2}$, mais pas pour $f(x) = x^2$, car $\displaystyle\int_a^b t^2\,\mathrm{d}t = \frac{b^3-a^3}{3} \neq (b-a)\left(\frac{a+b}{2}\right)^2$. C'est donc une méthode d'ordre 1.

Pour $n := 1$, on doit choisir deux points x_0 et x_1. Si l'on prend $x_0 := a$ et $x_1 := b$, on obtient le polynôme d'interpolation linéaire $L(A;f) = \dfrac{(x-a)f(b)-(x-b)f(a)}{b-a}$ et la formule $\displaystyle\int_a^b f(t)\,\mathrm{d}t \approx (b-a)\frac{f(a)+f(b)}{2}$, autrement dit, $w_0 = w_1 = \dfrac{1}{2}$ (*méthode des trapèzes*). La formule est celle de l'aire du trapèze (voir la figure).



Point milieu Trapèzes

Comme $\dfrac{b^3-a^3}{3} \neq (b-a)\dfrac{a^2+b^2}{2}$, cette méthode est exactement d'ordre 1.

Exercice 17.

Comment obtient-on un ordre supérieur à n ? Examiner le cas $n := 1$ sur $[0,1]$.

Solution. Il suffit d'obtenir une formule exacte pour *un* polynôme de degré $n+1$ pour qu'elle soit exacte pour tous (linéarité). En prenant $\omega := \displaystyle\prod_{i=0}^{n}(X - x_i)$, on voit que la condition est : $\displaystyle\int_a^b \omega(t)\,\mathrm{d}t = 0$. C'est une condition algébrique sur les x_i.

Par exemple, si $n := 1$, $a := 0$ et $b := 1$, la condition s'écrit :

$$\int_0^1 (t-x_0)(t-x_1)\,\mathrm{d}t = 0 \qquad \text{c'est-à-dire} \qquad \frac{1}{3} - \frac{1}{2}(x_0+x_1) + x_0x_1 = 0.$$

Lorsque l'on sait majorer l'erreur d'approximation par un polynôme d'interpolation, on peut majorer l'erreur d'une formule de quadrature.

Exemples. Posons $c := \dfrac{a+b}{2}$ et supposons que f est de classe $\mathcal{C}^2$. De la formule de Taylor, on déduit l'inégalité

$$\left|f(x) - f(c) - (x-c)f'(c)\right| \leqslant \frac{(x-c)^2}{2}\left\|f''\right\|_\infty$$

(il s'agit, comme toujours, de la norme uniforme sur le segment $[a,b]$). En intégrant, on obtient une majoration de l'erreur dans la méthode du point milieu :

$$\left|\int_a^b f(t)\,\mathrm{d}t - (b-a)f\left(\frac{a+b}{2}\right)\right| \leqslant \frac{(b-a)^3}{24}\left\|f''\right\|_\infty.$$

Pour la méthode des trapèzes, on reprend l'estimation de l'erreur dans l'interpolation linéaire. Il résulte du lemme 7 de la page 898 que :

$$|f(x) - L(a,b;x)| \leqslant \frac{1}{2}\left\|f''\right\|_\infty |(x-a)(x-b)|.$$

En intégrant cette inégalité, on obtient une majoration de l'erreur dans la méthode des trapèzes :

$$\left|\int_a^b f(t)\,\mathrm{d}t - (b-a)\frac{f(a)+f(b)}{2}\right| \leqslant \frac{(b-a)^3}{12}\left\|f''\right\|_\infty.$$

Exercice 18.
Déduire du théorème 13 de la page 906 une majoration de l'erreur de quadrature en général.

Solution. De la formule (7) de la page 906, on déduit en intégrant :

$$\begin{aligned}\left|\int_a^b f(t)\,\mathrm{d}t - (b-a)\sum_{i=0}^n w_i f(x_i)\right| &\leqslant \frac{1}{(n+1)!}\int_a^b |\omega_A(t)|\,\mathrm{d}t\left\|f^{(n+1)}\right\|_\infty \\ &\leqslant \frac{b-a}{(n+1)!}\left\|\omega_A\right\|_\infty\left\|f^{(n+1)}\right\|_\infty\end{aligned}$$

De même que dans l'interpolation augmenter le nombre de points ne garantit pas la convergence (phénomène de Runge), de même augmenter le nombre des points x_i dans une méthode de quadrature élémentaire ne garantit pas que l'erreur de quadrature tendra vers 0. On est donc conduit à introduire des *méthodes de quadrature composées*, définies de la manière suivante : on subdivise le segment $[a,b]$ en des segments $[a_i, a_{i+1}]$, où $a = a_0 < \cdots < a_p = b$, et l'on applique la relation de Chasles $\displaystyle\int_a^b = \sum_{i=0}^{p-1}\int_{a_i}^{a_{i+1}}$, puis on utilise, sur chaque segment $[a_i, a_{i+1}]$, une méthode de quadrature élémentaire, avec des points $x_{i,j} \in [a_i, a_{i+1}]$ $(0 \leqslant j \leqslant n_i)$ et des poids $w_{i,j}$:

$$\int_a^b f(t)\,\mathrm{d}t \approx \sum_{i=0}^{p-1}(a_{i+1}-a_i)\sum_{j=0}^{n_i} w_{i,j} f(x_{i,j}).$$

Exemples. Si l'on prend tous les $n_i := 0$, et $x_i := x_{i,0} \in [a_i, a_{i+1}]$, la formule obtenue n'est autre qu'une somme de Riemann :

$$\int_a^b f(t)\,\mathrm{d}t \approx \sum_{i=0}^{p-1}(a_{i+1}-a_i)f(x_i).$$

La méthode (composée) des rectangles à gauche (resp. des rectangles à droite, resp. du point milieu) consiste à utiliser l'approximation de $\displaystyle\int_a^b f(t)\,\mathrm{d}t$ par :

$$\sum_{i=0}^{p-1}(a_{i+1}-a_i)f(a_i),\ \text{resp.}\ \sum_{i=0}^{p-1}(a_{i+1}-a_i)f(a_i),\ \text{resp.}\ \sum_{i=0}^{p-1}(a_{i+1}-a_i)f\left(\frac{a_i+a_{i+1}}{2}\right).$$

Si l'on prend maintenant tous les $n_i := 1$ et $x_{i,0} := a_i$, $x_{i,1} := a_{i+1}$, on obtient la méthode (composée) des trapèzes :

$$\begin{aligned}\int_a^b f(t)\,\mathrm{d}t &\approx \sum_{i=0}^{p-1}(a_{i+1}-a_i)\frac{f(a_i)+f(a_{i+1})}{2}\\ &= \frac{a_1-a_0}{2}f(a_0)+\sum_{i=1}^{p-1}\frac{a_{i+1}-a_{i-1}}{2}f(a_i)+\frac{a_p-a_{p-1}}{2}f(a_p).\end{aligned}$$

4.2.2 Méthodes de Newton-Cotes

On appelle *méthode de Newton-Cotes de rang* n, ce que l'on abrégera en NC_n, une méthode de quadrature composée dans laquelle, sur chaque segment $[a_i, a_{i+1}]$, on a $n_i = n$ et les $(n+1)$ points $x_{i,j}$ forment une subdivision régulière :

$$\forall i \in [\![0, p-1]\!]\ ,\ \forall j \in [\![0, n]\!]\ ,\ x_{i,j} = a_i + j\frac{a_{i+1}-a_i}{n}\cdot$$

La méthode NC_n est, par construction, d'ordre supérieur ou égal à n ; nous verrons que cette valeur peut être améliorée (proposition 22 de la page suivante). Nous pouvons reprendre les calculs de la section 3.2.1 de la page 910. Notons l_i les polynômes de Lagrange associés aux points $i \in [\![0, n]\!] \subset [0, n]$:

$$l_j(s) := \prod_{\substack{0\leqslant k\leqslant n\\ k\neq j}}\frac{s-k}{j-k}\cdot$$

Pour chaque $i \in [\![0, p-1]\!]$, les polynômes $L_{i,j}$ d'interpolation sur le segment $[a_i, a_{i+1}]$ sont donnés par la formule :

$$\forall j \in [\![0, n]\!]\ ,\ L_{i,j}(x) = l_j(s),\ \text{où}\ s := n\frac{x-a_i}{a_{i+1}-a_i}\cdot$$

La formule de quadrature élémentaire associée est donc la suivante :

$$\int_{a_i}^{a_{i+1}} f(t)\,\mathrm{d}t \approx (a_{i+1}-a_i)\sum_{j=0}^{n} w_{i,j}f(x_{i,j}),$$

$$\text{où}\quad w_{i,j} := \frac{1}{a_{i+1}-a_i}\int_{a_i}^{a_{i+1}} L_{i,j}(t)\,\mathrm{d}t = \frac{1}{n}\int_0^n l_j(s)\ \mathrm{d}s.$$

Exemples.
Pour $n := 1$, il s'agit d'une interpolation linéaire aux points $x_{i,0} := a_i$ et $x_{i,1} := a_{i+1}$, donc de la méthode des trapèzes composée.

Pour $n := 2$, on a les points $x_{i,0} := a_i$, $x_{i,1} := \dfrac{a_i + a_{i+1}}{2}$ et $x_{i,2} := a_{i+1}$. Les l_j sont : $l_0(s) = \dfrac{(s-1)(s-2)}{2}$, $l_1(s) = -s(s-2)$ et $l_2(s) = \dfrac{(s-1)(s-2)}{2}$, d'où les poids :

$$w_{i,0} = \frac{1}{2}\int_0^2 l_0(s)\,\mathrm{d}s = \frac{1}{6},\ w_{i,1} = \frac{1}{2}\int_0^2 l_1(s)\,\mathrm{d}s = \frac{2}{3} \quad \text{et} \quad w_{i,2} = \frac{1}{2}\int_0^2 l_2(s)\,\mathrm{d}s = \frac{1}{6}\cdot$$

La formule de quadrature associée (*méthode de Simpson*) est :

$$\int_a^b f(t)\,\mathrm{d}t \approx \sum_{i=0}^{p-1}(a_{i+1} - a_i)\left(\frac{1}{6}f(a_i) + \frac{2}{3}f\left(\frac{a_i + a_{i+1}}{2}\right) + \frac{1}{6}f(a_{i+1})\right).$$

Proposition 22. (i) Les poids de la méthode NC_n vérifient la propriété de symétrie :

$$\forall i \in [\![0, p-1]\!]\ ,\ \forall j \in [\![0, n]\!]\ ,\ w_{i,j} = w_{i,n-j}.$$

(ii) Si n est pair, la méthode NC_n est d'ordre supérieur ou égal à $n+1$.

Démonstration. (i) Remarquons d'abord (avec les notations qui précèdent) que le polynôme $l_j(n-s)$ est de degré inférieur ou égal à n, qu'il vaut 1 en $n-j$ et 0 aux autres entiers de $[0, n]$. On a donc $l_j(n-s) = l_{n-j}(s)$. En intégrant cette égalité sur $[0, n]$, on trouve $w_{i,j} = w_{i,n-j}$.
(ii) Il suffit de vérifier l'exactitude de la formule de quadrature élémentaire pour un polynôme de degré $(n+1)$. Prenons $P := \left(X - \dfrac{a_i + a_j}{2}\right)^{n+1}$. Puisque n est pair, on a $P(a_i + a_j - X) = -P(X)$. En intégrant sur le segment $[a_i, a_{i+1}]$, on trouve donc 0. Mais la somme $\sum_{j=0}^{n} w_{i,j}P(x_{i,j})$ est nulle, car le changement d'indice $j \leftarrow n-j$ donne :

$$\sum_{j=0}^{n} w_{i,j}P(x_{i,j}) = \sum_{j=0}^{n} w_{i,n-j}P(x_{i,n-j}) = \sum_{j=0}^{n} w_{i,j}P(a_i + a_j - x_{i,j}) = -\sum_{j=0}^{n} w_{i,j}P(x_{i,j}).$$

On peut démontrer que les méthodes de Newton-Cotes de rangs $2m$ et $2m+1$ sont exactement d'ordre $2m+1$. Pour cette raison, on utilise en général des méthodes de Newton-Cotes d'ordre pair (à l'exception de la méthode des trapèzes).
Nous ne démontrerons la convergence des méthodes de quadratures que pour les méthodes composées à partir de méthodes élémentaires semblables (en particulier, NC_n).
Autrement dit, nous supposons que $n_0, \ldots, n_{p-1}$ sont égaux à un même entier n et que, pour chaque $j \in [\![0, n]\!]$, $w_{0,j}, \ldots, w_{p-1,j}$ sont égaux à un même poids w_j.

On a alors $\sum_{j=0}^{n} w_j = 1$. Ainsi, pour toute subdivision du segment $[a, b]$ en des segments $[a_i, a_{i+1}]$, où $a = a_0 < \cdots < a_p = b$, il ne reste à choisir que les $(n+1)$ points d'interpolation $x_{i,j}$ sur chaque $[a_i, a_{i+1}]$.

Théorème 23. Lorsque le pas $\max_{0\leqslant i\leqslant p-1}(a_{i+1}-a_i)$ tend vers 0, la formule de quadrature $\sum_{i=0}^{p-1}(a_{i+1}-a_i)\sum_{j=0}^{n_i} w_{i,j}f(x_{i,j})$ tend vers $\int_a^b f(t)\,\mathrm{d}t$.

Démonstration. On permute les indices de sommation :

$$\sum_{i=0}^{p-1}(a_{i+1}-a_i)\sum_{j=0}^{n_i} w_{i,j}f(x_{i,j}) = \sum_{i=0}^{p-1}(a_{i+1}-a_i)\sum_{j=0}^{n} w_j f(x_{i,j})$$
$$= \sum_{j=0}^{n} w_j \sum_{i=0}^{p-1}(a_{i+1}-a_i)f(x_{i,j}).$$

Mais chaque somme $\sum_{i=0}^{p-1}(a_{i+1}-a_i)f(x_{i,j})$ est une somme de Riemann, qui tend vers l'intégrale de f lorsque le pas tend vers 0.

La somme totale tend donc vers $\sum_{j=0}^{n} w_j \int_a^b f(t)\,\mathrm{d}t = \int_a^b f(t)\,\mathrm{d}t$, puisque $\sum_{j=0}^{n} w_j = 1$. ∎

Ce résultat s'applique aux méthodes de Newton-Cotes. En revanche, il n'est pas vrai qu'une méthode de Newton-Cotes élémentaire de rang n converge lorsque n tend vers $+\infty$. Et l'on constate de plus que, pour n assez grand, des poids négatifs apparaissent (exercice II.8.18 de la page 953).

4.2.3 Estimation de l'erreur dans une méthode de quadrature

Une méthode systématique d'évaluation de l'erreur repose sur une formule intégrale que nous allons détailler. Nous noterons, pour tout $y \in \mathbb{R}$ et tout $N \in \mathbb{N}^*$:

$$y_+ := \max(0, y) \quad \text{et} \quad y_+^N := (y_+)^N.$$

On prendra garde que, lorsque N est pair, on n'a pas nécessairement $y_+^N := (y^N)_+$.

Fixons $t \in \mathbb{R}$. Pour tout $N \in \mathbb{N}^*$, la fonction $x \mapsto (x-t)_+^N$ est de classe $\mathcal{C}^{N-1}$ et, pour tout $k \in [\![0, N-1]\!]$, sa dérivée $k^{ème}$ est la fonction $x \mapsto \binom{N}{k}(x-t)_+^{N-k}$.

Soit f une fonction de classe $\mathcal{C}^{N+1}$ sur un segment $[a,b]$. La formule de Taylor avec reste intégral s'écrit : $f(x) = P(x) + R(x)$, où $P := \sum_{k=0}^{N} \frac{f^{(k)}(a)}{k!}(x-a)^k \in \mathbb{R}_N[X]$ et où le reste $R(x)$ a la forme :

$$R(x) := \int_a^x \frac{1}{N!}(x-t)^N f^{(N+1)}(t)\,\mathrm{d}t = \int_a^b \frac{1}{N!}(x-t)_+^N f^{(N+1)}(t)\,\mathrm{d}t.$$

Cette dernière écriture est une conséquence immédiate de la définition, puisque $(x-t)_+^N$ est égal à $(x-t)^N$ pour $t \in [a,x]$ et à 0 pour $t \in [x,b]$.

Considérons maintenant une méthode de quadrature élémentaire que nous écrirons sous la forme :

$$I(f) := \int_a^b f(t)\,\mathrm{d}t \approx J(f) := \sum_{i=0}^{n} \lambda_i f(x_i).$$

Nous noterons $E(f) := I(f) - J(f)$ l'erreur de quadrature par cette méthode, pour la fonction f.

Définition 3. Supposons que la méthode de quadrature ci-dessus est d'ordre N. Le *noyau de Peano* associé à cette méthode est la fonction qui, à $t \in [a,b]$, associe :

$$K_N(t) := E\big(x \mapsto (x-t)_+^N\big). \tag{17}$$

Par définition des notations $(x-t)_+^N$ et $E(f)$, on a donc :

$$K_N(t) := \int_a^b (x-t)_+^N\,\mathrm{d}t - \sum_{i=0}^{n} \lambda_i (x_i - t)_+^N. \tag{18}$$

Théorème 24. Pour toute fonction f de classe $\mathcal{C}^{N+1}$ sur $[a,b]$, on a la formule :

$$E(f) = \int_a^b \frac{1}{N!} K_N(t) f^{(N+1)}(t)\,\mathrm{d}t. \tag{19}$$

Démonstration.

Avec les notations $f = P+R$ de la formule de Taylor, on a $E(f) = E(P)+E(R) = E(R)$, puisque $\deg P \leqslant N$ et que la méthode est d'ordre N. Ainsi :

$$E(f) = E(R) = \int_a^b \int_a^b \frac{(x-t)_+^N}{N!} f^{(N+1)}(t)\,\mathrm{d}t\,\mathrm{d}x - \sum_{i=0}^{n} \lambda_i \int_a^b \frac{(x_i-t)_+^N}{N!} f^{(N+1)}(t)\,\mathrm{d}t.$$

Du théorème de Fubini (théorème 133 de la page 648) on déduit que le premier terme du membre droit vaut :

$$\int_a^b \left(\int_a^b (x-t)_+^N\,\mathrm{d}x \right) \frac{f^{(N+1)}(t)}{N!}\,\mathrm{d}t,$$

et la linéarité de l'intégrale), prouve que le membre gauche est égal à :

$$\int_a^b \sum_{i=0}^{n} \lambda_i (x_i - t)_+^N \frac{f^{(N+1)}(t)}{N!}\,\mathrm{d}t.$$

Finalement, $E(f) = \displaystyle\int_a^b \left(\int_a^b (x-t)_+^N\,\mathrm{d}x - \sum_{i=0}^{n} \lambda_i (x_i-t)_+^N \right) \frac{f^{(N+1)}(t)}{N!}\,\mathrm{d}t$, ce qui est la formule indiquée. ∎

Corollaire 25.
Si le noyau de Peano K_N est de signe constant, il existe $c \in]a,b[$ tel que :

$$E(f) = f^{(N+1)}(c)\int_a^b \frac{1}{N!}K_N(t)\,\mathrm{d}t \quad \text{et} \quad |E(f)| \leqslant \left\|f^{(N+1)}\right\|_\infty \int_a^b \frac{1}{N!}\,|K_N(t)|\,\mathrm{d}t.$$

Cela découle de la formule de la moyenne. On peut montrer que la condition du corollaire est vérifiée par la méthode NC_n.

Erreur de quadrature dans la méthode du point milieu. La méthode est d'ordre $N = 1$. Le noyau de Peano est donné par la formule :

$$\begin{aligned} K_1(t) &= \int_a^b (x-t)_+\,\mathrm{d}x - (b-a)\left(\frac{a+b}{2} - t\right)_+ \\ &= \frac{(b-t)^2}{2} - (b-a)\left(\frac{a+b}{2} - t\right)_+ \\ &= \begin{cases} \dfrac{(t-a)^2}{2} \text{ si } a \leqslant t \leqslant \dfrac{a+b}{2}, \\ \dfrac{(b-t)^2}{2} \text{ si } \dfrac{a+b}{2} \leqslant t \leqslant b, \end{cases} \end{aligned}$$

car $\displaystyle\int_a^b (x-t)_+\,\mathrm{d}x = \frac{(b-t)^2}{2}$. Le noyau de Peano est donc de signe constant positif sur $[a,b]$, d'où la formule d'erreur pour une fonction f de classe $\mathcal{C}^2$:

$$\begin{aligned} E(f) &= f''(c)\int_a^b K_1(t)\,\mathrm{d}t \\ &= f''(c)\int_a^{\frac{a+b}{2}} \frac{(t-a)^2}{2}\,\mathrm{d}t + f''(c)\int_{\frac{a+b}{2}}^b \frac{(b-t)^2}{2}\,\mathrm{d}t \\ &= \frac{(b-a)^3}{24}f''(c). \end{aligned}$$

Par conséquent, $|E(f)| \leqslant \dfrac{(b-a)^3}{24}\left\|f''\right\|_\infty$. Considérons maintenant une méthode du point milieu composée sur $[a,b]$, dans laquelle tous les segments $[a_i, a_{i+1}]$ ont même longueur $\dfrac{b-a}{p}$. Les p erreurs de quadratures élémentaires s'additionnent :

$$E(f) = \sum_{i=0}^{p-1} \frac{(a_{i+1}-a_i)^3}{24p^3}f''(c_i),$$

où $c_i \in]a_i, a_{i+1}[$ pour $0 \leqslant i \leqslant p-1$. Le théorème des valeurs intermédiaires appliqué à la fonction continue f'' implique que $\dfrac{1}{p}\displaystyle\sum_{i=0}^{p-1} f''(c_i) = f''(c)$ pour un $c \in [a,b]$. L'erreur a donc la forme :

$$E(f) = \frac{(b-a)^3}{24p^2}f''(c).$$

Exercice 19.

Calculer l'erreur de quadrature dans la méthode des trapèzes.

Solution. La méthode est d'ordre $N = 1$. Le noyau de Peano est donné par la formule :

$$\begin{aligned} K_1(t) &= \int_a^b (x-t)_+ \,\mathrm{d}x - (b-a)\left(\frac{a+b}{2} - t\right)_+ \\ &= \left[\frac{(x-t)_+^2}{2}\right]_{x=a}^{x=b} - \frac{b-a}{2}\big((a-t)_+ + (b-t)_+\big) \\ &= \frac{(b-t)^2}{2} - \frac{b-a}{2}(b-t) \\ &= \frac{(b-t)(a-t)}{2}, \end{aligned}$$

qui est de signe constant négatif sur $[a,b]$, d'où la formule pour une fonction f de classe $\mathcal{C}^2$:

$$E(f) = f''(c)\int_a^b \frac{(b-t)(a-t)}{2}\,\mathrm{d}t = -\frac{(b-a)^3}{12}f''(c).$$

Par conséquent, $|E(f)| \leqslant \dfrac{(b-a)^3}{12}\,\|f''\|_\infty$.

Erreur de quadrature dans la méthode de Simpson. La méthode est d'ordre $N = 3$. On a :

$$\begin{aligned} K_3(t) &= \int_a^b (x-t)_+^3 \,\mathrm{d}x - (b-a)\left(\frac{1}{6}(a-t)_+^3 + \frac{2}{3}(\frac{a+b}{2} - t)_+^3 + \frac{1}{6}(b-t)_+^3\right) \\ &= \frac{(b-t)^4}{4} - (b-a)\left(\frac{2}{3}(\frac{a+b}{2} - t)_+^3 + \frac{1}{6}(b-t)^3\right) \\ &= \begin{cases} \dfrac{(t-a)^3(3t-2b-a)}{12} & \text{si } a \leqslant t \leqslant \dfrac{a+b}{2}, \\ \dfrac{(b-t)^3(b+2a-3t)}{12} & \text{si } \dfrac{a+b}{2} \leqslant t \leqslant b, \end{cases} \end{aligned}$$

qui est de signe constant négatif sur $[a,b]$, en vertu des inégalités :

$$\frac{2a+b}{3} \leqslant \frac{a+b}{2} \leqslant \frac{a+2b}{3}.$$

On a donc la formule d'erreur suivante pour une fonction f de classe $\mathcal{C}^2$:

$$E(f) = \frac{f^{(4)}(c)}{6}\int_a^b K_3(t)\,\mathrm{d}t = -\frac{(b-a)^5}{2880}f^{(4)}(c)$$

(vérification laissée au lecteur). Par conséquent, $|E(f)| \leqslant \dfrac{(b-a)^5}{2880}\left\|f^{(4)}\right\|_\infty$. Considérons maintenant une méthode de Simpson composée sur $[a,b]$, dans laquelle tous les segments $[a_i, a_{i+1}]$ ont même longueur $\dfrac{b-a}{p}$. Les p erreurs de quadratures élémentaires s'additionnent, et l'on trouve de la même manière que dans le cas du point milieu :

$$E(f) = -\frac{(b-a)^5}{2880p^2}f^{(4)}(c).$$

4.3 Méthode de Gauß

On fixe maintenant une « fonction poids » sur $[a,b]$, c'est-à-dire une fonction w continue positive non nulle sur $]a,b[$. On désire trouver une formule de quadrature élémentaire :

$$\int_a^b w(t)f(t)\,\mathrm{d}t \approx \sum_{i=0}^{n} \lambda_i f(x_i),$$

d'ordre le plus élevé possible (pour n donné) dans le sens suivant : cette formule est dite d'ordre supérieur ou égal à N si elle est exacte pour toute $f \in \mathbb{R}_N[X]$.

Exercice 20.

Quel ordre peut-on raisonnablement espérer avec $(n+1)$ points ?

Solution. Pour que la méthode ci-dessus soit d'ordre N, il faut, et il suffit, que les $(N+1)$ équations suivantes soient satisfaites :

$$\forall k \in [\![0,N]\!]\ ,\ \sum_{i=0}^{n} \lambda_i x_i^k = \frac{b^{k+1}-a^{k+1}}{b-a}\cdot$$

Il y a $2(n+1)$ inconnues, les λ_i et les x_i. On peut espérer que, pour $N+1 = 2(n+1)$, c'est-à-dire $N = 2n+1$, le système *non linéaire* ci-dessus ait une solution, et que celle-ci soit utilisable, *i.e.* que les x_i soient distincts et dans l'intervalle $[a,b]$.

Afin de pouvoir appliquer la théorie des polynômes orthogonaux de [L2-L3], nous ferons l'hypothèse suivante : tous les moments $\displaystyle\int_a^b t^k w(t)\,\mathrm{d}t$ sont finis.

On peut donc définir un produit scalaire $\langle P,Q\rangle := \displaystyle\int_a^b w(t)P(t)Q(t)\,\mathrm{d}t$ sur $\mathbb{R}[X]$. Il existe alors une unique suite $(P_n)_{n\in\mathbb{N}}$ de polynômes deux à deux orthogonaux tels que chaque P_n est unitaire de degré n.

Théorème 26. Il existe un unique choix des points $x_0,\dots,x_n$ et des poids $\lambda_0,\dots,\lambda_n$ tel que la formule ci-dessus soit d'ordre supérieur ou égal à $(2n+1)$; et elle est alors d'ordre exactement $(2n+1)$.

Démonstration. Supposons d'abord les x_i et λ_i trouvés tels que la formule soit d'ordre supérieur ou égal à $(2n+1)$. On notera $\omega := \displaystyle\prod_{i=0}^{n}(X-x_i)$. Alors, pour tout $P \in \mathbb{R}_n[X]$, $\deg \omega P \leqslant 2n+1$, d'où $\displaystyle\int_a^b w(t)\omega(t)P(t)\,\mathrm{d}t = 0$. Le polynôme ω est unitaire de degré $(n+1)$ et orthogonal à $\mathbb{R}_n[X]$, c'est donc le polynôme P_{n+1} et les x_i sont les $(n+1)$ racines de P_{n+1}. Notant temporairement L_i les polynômes d'interpolation associés, on doit avoir, pour chaque j :

$$\int_a^b w(t)L_j(t)\,\mathrm{d}t = \sum_{i=0}^{n} \lambda_i L_j(x_i) = \lambda_j,$$

ce qui détermine également les coefficients λ_i.

Réciproquement (encore d'après le module sur les polynômes orthogonaux de [L2-L3]), P_{n+1} a $(n+1)$ racines réelles distinctes dans $]a,b[$; prenons pour x_i ses racines, notons L_i les polynômes d'interpolation associés et posons $\lambda_i := \displaystyle\int_a^b w(t)L_j(t)\,\mathrm{d}t$. La formule de quadrature $\displaystyle\int_a^b w(t)f(t)\,\mathrm{d}t \approx \sum_{i=0}^{n} \lambda_i f(x_i)$ est donc exacte pour les L_i, elle est donc d'ordre supérieur ou égal à n. Soit $P \in \mathbb{R}_{2n+1}[X]$. Dans la division euclidienne $P = QP_{n+1} + R$, on a donc $\deg Q, \deg R \leqslant n$. On écrit :

$$\begin{aligned}\int_a^b w(t)P(t)\,\mathrm{d}t &= \int_a^b w(t)Q(t)P_{n+1}(t)\,\mathrm{d}t \int_a^b w(t)R(t)\,\mathrm{d}t \\ &= \langle Q, P_{n+1}\rangle + \sum_{i=0}^{n} \lambda_i R(x_i) = \sum_{i=0}^{n} \lambda_i P(x_i).\end{aligned}$$

En effet, $\langle Q, P_{n+1}\rangle = 0$ parce que P_{n+1} est orthogonal à $\mathbb{R}_n[X]$; la formule de quadrature est exacte pour R parce que son degré est inférieur ou égal à n ; et $R(x_i) = P(x_i)$ parce que P_{n+1} divise $P - R$ et s'annule en les x_i. La méthode est donc bien d'ordre supérieur ou égal à $(2n+1)$.

Enfin, on a $\displaystyle\int_a^b w(t)\omega^2(t)\,\mathrm{d}t > 0$, alors que $\displaystyle\sum_{i=0}^{n} \lambda_i \omega^2(x_i) = 0$: comme $\deg \omega^2 = 2n+2$, la méthode est d'ordre exactement $(2n+1)$. ∎

Exemples. Prenons $a := -1$, $b := 1$ et $w := 1$. Les polynômes orthogonaux correspondants sont les polynômes de Legendre. On a $P_0 = 1$, $P_1 = X$ et $P_2 = X^2 - \dfrac{1}{3}$. Pour $n := 0$, le point d'interpolation est $x_0 = 0$, le polynôme d'interpolation est $L_0 = 1$, le poids est $\lambda_0 = 2$ et l'on trouve la formule de quadrature : $\displaystyle\int_{-1}^{1} f(t)\,\mathrm{d}t \approx 2f(0)$; c'est la formule du point milieu. Pour $n := 1$, $x_0 = -\sqrt{\dfrac{1}{3}}$, $x_1 = \sqrt{\dfrac{1}{3}}$, $\lambda_0 = \lambda_1 = 1$ (sans calcul : ils sont égaux par symétrie et leur somme est la longueur de l'intervalle, soit 2). On trouve la formule de quadrature :

$$\int_{-1}^{1} f(t)\,\mathrm{d}t \approx f\left(-\sqrt{\frac{1}{3}}\right) + f\left(\sqrt{\frac{1}{3}}\right).$$

Prenons $a := -1$, $b := 1$ et $w := \dfrac{1}{\sqrt{1-x^2}}$. Les polynômes orthogonaux correspondants sont les polynômes de Tchebychev. On a $P_0 = 1$, $P_1 = X$ et $P_2 = X^2 - \dfrac{1}{2}$. Pour $n := 0$, on trouve comme ci-dessus la formule de quadrature : $\displaystyle\int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}}\,\mathrm{d}t \approx 2f(0)$; c'est encore la formule du point milieu. Pour $n := 1$, $x_0 = -\sqrt{\dfrac{1}{2}}$, $x_1 = \sqrt{\dfrac{1}{2}}$, $\lambda_0 = \lambda_1 = 1$

(sans calcul : ils sont égaux par symétrie et leur somme est la longueur de l'intervalle, soit 2). On trouve la formule de quadrature :

$$\int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}}\,\mathrm{d}t \approx f\left(-\sqrt{\frac{1}{2}}\right) + f\left(\sqrt{\frac{1}{2}}\right).$$

Exercice 21.

Reprendre le cas des polynômes de Legendre pour $n := 2$.

Solution. On a $P_3 = X^3 - \dfrac{3}{5}X$, donc $x_0 = -a$, $x_1 = 0$ et $x_2 = a$, où l'on a noté $a := \sqrt{\dfrac{3}{5}}$. Les polynômes d'interpolation sont $L_0 = \dfrac{X(X-a)}{2a^2}$, $L_1 = \dfrac{a^2 - X^2}{a^2}$ et $L_2 = \dfrac{X(X+a)}{2a^2}$. Les poids sont $\lambda_0 = \lambda_2 = \dfrac{5}{9}$ et $\lambda_1 = \dfrac{8}{9}$. La formule de quadrature est donc :

$$\int_{-1}^{1} f(t)\,\mathrm{d}t \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right).$$

---

5 Analyse matricielle

L'analyse matricielle est un vaste domaine à l'intérieur de l'analyse numérique, qui comprend notamment la résolution de systèmes linéaires (objet de cette section) et l'optimisation et la recherche d'éléments propres (qui sont abordés dans [MPA1]).

Les sciences naturelles et les sciences de l'ingénieur peuvent nécessiter la résolution de *très gros* systèmes linéaires : plusieurs milliers de variables et d'équations. En particulier, la résolution approchée par *discrétisation* des systèmes différentiels ou aux dérivées partielles (dont un exemple est traité dans l'exercice II.8.0 de la page 949), ainsi que l'extension à plusieurs variables de la méthode de Newton de résolution approchée des équations (section 2) posent de tels problèmes (ces domaines sont abordés dans [MPA1]).

Les systèmes qui apparaissent dans les problèmes concrets sont *souvent* dotés de propriétés particulières : matrices symétriques, voire définies positives, matrices bandes ou même tridiagonales. C'est pourquoi la présence de telles conditions dans certains théorèmes ou algorithmes ne doit pas être considérée comme artificielle.

On a déjà vu, dans le module de calcul matriciel de [L1], une méthode très efficace de résolution : la méthode du pivot de Gauß. Nous approfondirons ici les aspects *numériques* du problème : question de *stabilité* (influence des erreurs d'arrondi liées à la représentation machine), possibilité de résolution *approchée*, usage de l'algèbre bilinéaire. Nous nous en tiendrons aux matrices *réelles*, mais la plupart des algorithmes s'étendent au cas des matrices complexes.

5.1 Méthodes directes

On veut résoudre le système :

$$AX = B, \text{ où } A \in GL_n(\mathbb{R}) \quad \text{et} \quad B \in M_{n,1}(\mathbb{R}). \tag{20}$$

Rappelons que l'inversion de A à l'aide de la comatrice ou l'usage des formules de Cramer impliquent un coût (nombres d'opérations) de l'ordre de grandeur de $n!$, alors que le coût de la méthode de Gauß est $\mathrm{O}(n^3)$.

Exercice 22.

On dit (« loi de Moore ») que la puissance du matériel informatique double tous les dix-huit mois . On suppose qu'à un moment donné, la résolution d'un système linéaire de taille $n := 1000$ est possible en un temps « raisonnable ». Combien de temps faut-il pour que la résolution avec $n := 1001$ soit possible en un temps raisonnable, si l'on suppose que c'est la méthode de Cramer qui est utilisée ? Et si c'est la méthode de Gauß ?

Solution. Passer de 1000 à 1001 par la méthode de Cramer *en gardant un temps de calcul équivalent* requiert une puissance 1001 fois plus grande. Admettons que la puissance (ici comprise comme la vitesse) soit régulièrement croissante : elle est donc multipliée par $2^{2/3}$ chaque année. On résout $(2^{2/3})^k = 1001$, c'est-à-dire $k = \dfrac{3\ln 1001}{2\ln 2} \approx 14{,}95$. Il faudra quinze ans. Avec la méthode de Gauß, il faut résoudre $(2^{2/3})^k = (1{,}001)^3$, c'est-à-dire $k = \dfrac{9\ln 1{,}001}{2\ln 2} \approx 0{,}0065$. Il faudra moins de deux jours et demi.

Toutes les méthodes dites *directes* de résolution ont pour but de ramener le système (20) à un système triangulaire :

$$\begin{aligned} a'_{1,1}x_1 + \cdots + a'_{1,n}x_n &= b'_1, \\ \ddots \cdots &= \cdots \\ a'_{n,n}x_n &= b'_n. \end{aligned}$$

Ce système est alors résolu très simplement par l'algorithme suivant :

```
pour i de n a 1 faire
   (x[i] := b'[i] ;
    pour j de i+1 a n faire x[i] := x[i] - a'[i,j]*x[j];
    x[i] := x[i]/a'[i,i]) ;;
```

Le coût de résolution d'un système triangulaire est en $\mathrm{O}(n^2)$. La méthode de Gauß pour s'y ramener a un coût en $\mathrm{O}(n^3)$: le nombre d'opérations arithmétiques est équivalent à $2n^3/3$, et le coût des recherches de pivots et des échanges de lignes est en $\mathrm{O}(n^2)$.

5.1.1 Stabilité et conditionnement

Les données A et B du système (20) ne sont en général pas connues exactement : ce peuvent être des résultats de mesures physiques ou de calculs approchés antérieurs, et, de toutes façons, il a fallu les convertir en représentation machine, donc arrondir. Le résultat dépend-il beaucoup de cette erreur de départ ? Disons que l'on résout en fait un système de matrice $A + \delta A$ et de terme constant $B + \delta B$, où les termes d'erreur δA et δB sont supposés « petits ». Nous analyserons séparément les deux causes d'erreur (sur A et sur B) en admettant que, lorsqu'elles ont lieu simultanément, leurs effets se cumulent. Commençons par l'erreur sur B. Notant X la solution de (20) et $X + \delta X$ la solution du système approché, on a donc :

$$A(X + \delta X) = B + \delta B \Longrightarrow A\delta X = \delta B \Longrightarrow \delta X = A^{-1}\delta B.$$

Pour aller plus loin, on choisit une norme $\|-\|$ sur $\mathbb{R}^n$ et l'on note également $\|-\|$ la norme matricielle subordonnée sur $M_n(\mathbb{R})$.

$$\|\delta X\| \leqslant \left\|A^{-1}\right\| \|\delta B\| \Longrightarrow \frac{\|\delta X\|}{\|X\|} \leqslant \left\|A^{-1}\right\| \frac{\|\delta B\|}{\|X\|} \leqslant \left\|A^{-1}\right\| \|A\| \frac{\|\delta B\|}{\|A\| \, \|X\|}$$
$$\leqslant \left\|A^{-1}\right\| \|A\| \frac{\|\delta B\|}{\|B\|},$$

en vertu de l'inégalité $\|B\| \leqslant \|A\| \, \|X\|$.

Définition 4. Le nombre $\mathrm{cond}(A) := \left\|A^{-1}\right\| \|A\|$ est appelé *conditionnement* de A ; il dépend donc du choix des normes. On notera en particulier cond_∞ et cond_2 les conditionnements calculés relativement aux normes subordonnées aux normes uniforme et euclidienne.

On a donc la majoration de l'erreur relative sur le résultat en fonction de l'erreur relative sur la donnée :

$$\frac{\|\delta X\|}{\|X\|} \leqslant \mathrm{cond}(A) \frac{\|\delta B\|}{\|B\|}.$$

De plus, l'égalité peut avoir lieu (voir l'exemple ci-dessous, dont l'argument se généralise). Si $\mathrm{cond}(A)$ est grand, la matrice est dite *mal conditionnée*. Pour une telle matrice, l'erreur relative $\dfrac{\|\delta X\|}{\|X\|}$ sur le résultat peut être grande même si l'erreur relative $\dfrac{\|\delta B\|}{\|B\|}$ sur la donnée du second membre ne l'est pas. Considérons maintenant l'effet d'une erreur sur A. Notant $X + \delta X$ la solution du système approché, on a donc :

$$(A + \delta A)(X + \delta X) = B \Longrightarrow A\delta X = -\delta A(X + \delta X) \Longrightarrow \delta X = -A^{-1}\delta A(X + \delta X).$$

On en déduit comme précédemment la majoration :

$$\frac{\|\delta X\|}{\|X + \delta X\|} \leqslant \mathrm{cond}(A) \frac{\|\delta A\|}{\|A\|}.$$

L'égalité peut d'ailleurs avoir lieu (raisonnement laissé au lecteur). Ici, l'erreur relative est calculée relativement à la solution modifiée, mais la morale est la même : pour une

matrice mal conditionnée, il peut y avoir amplification. Ces majorations sont complétées par l'exercice II.8.23 de la page 953.

Exemple.
La *matrice de Hilbert* d'ordre n est la matrice H_n dont le coefficient d'indices (i,j) vaut $\dfrac{1}{i+j-1}$. Par exemple, $H_2 = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}$, et $H_2^{-1} = \begin{pmatrix} 4 & -6 \\ -6 & 12 \end{pmatrix}$.
Ces matrices sont très mal conditionnées. Prenons la norme uniforme $\|X\|_\infty$ sur $\mathbb{R}^n$. La norme subordonnée est donnée par la formule :

$$\|(a_{i,j})_{1\leqslant i,j\leqslant n}\| = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{i,j}| .$$

Ainsi, $\|H_2\| = 3/2$ et $\left\|H_2^{-1}\right\| = 18$, d'où un conditionnement $\mathrm{cond}_\infty(H_2) = 27$.
Pour réaliser le cas le pire dans la majoration ci-dessus, on choisit d'abord :
- X tel que $\|H_2X\|_\infty = \|H_2\|\,\|X\|_\infty$; par exemple :

$$X := \begin{pmatrix} 6 \\ 6 \end{pmatrix} \qquad \text{d'où} \qquad B = H_2X = \begin{pmatrix} 9 \\ 15 \end{pmatrix} ;$$

- puis δB tel que $\left\|H_2^{-1}\delta B\right\|_\infty = \left\|H_2^{-1}\right\|\,\|\delta B\|_\infty$; par exemple :

$$\delta B := \begin{pmatrix} -1 \\ 1 \end{pmatrix} \qquad \text{d'où} \qquad \delta X = H_2^{-1}\delta B = \begin{pmatrix} -10 \\ 18 \end{pmatrix} .$$

On obtient les conditionnements des matrices de Hilbert (pour la norme uniforme) grâce à l'instruction Maple **`cond(hilbert(n))`**.
Pour n variant de 1 à 7, cela donne successivement : 1, 27, 748, 28375, 943656, $29070279 = 0{,}29070279.10^8$ et $\dfrac{1970389773}{2} = 0{,}9851948865.10^9$.

Exercice 23.

Expliciter les calculs qui découlent de l'exemple.

Solution.

On résout d'abord le système $H_2X = B := \begin{pmatrix} 9 \\ 5 \end{pmatrix}$; on trouve $X = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$ (comme prévu).
On résout ensuite le système légèrement modifié $H_2X' = B'$, où $B' := B + \delta B = \begin{pmatrix} 8 \\ 6 \end{pmatrix}$, *i.e.* un vecteur d'erreur $\delta B = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$. On trouve $X' = \begin{pmatrix} -4 \\ 24 \end{pmatrix}$, qui est très différent de X ! Pour quantifier cette amplification de l'erreur, on calcule $\|B\|_\infty = 9$ et $\|\delta B\|_\infty = 1$, d'où une erreur relative sur la donnée B égale à $\dfrac{\|\delta B\|_\infty}{\|B\|_\infty} = \dfrac{1}{9}$. Puis on calcule $\|X\|_\infty = 6$, l'erreur sur le résultat $\delta X = \begin{pmatrix} -6 \\ 18 \end{pmatrix}$, sa norme $\|\delta X\|_\infty = 18$,

d'où une erreur relative sur le résultat X égale à $\dfrac{\|\delta X\|_\infty}{\|X\|_\infty} = 3$. L'erreur relative a donc été multipliée par 27 (le conditionnement).

Exercice 24.

Modifier le système précédent en prenant pour matrice approchée $H_2' := \begin{pmatrix} 1 & 1/2 \\ 1/2 & 0,3 \end{pmatrix}$.

Solution.

On trouve pour solution $X' = \begin{pmatrix} 4 \\ 10 \end{pmatrix}$. La matrice d'erreurs est $\delta H_2 = \begin{pmatrix} 0 & 0 \\ 0 & -1/30 \end{pmatrix}$, dont la norme est $\|\delta H_2\| = 1/30$: l'erreur relative sur la matrice est donc $\dfrac{\|\delta H_2\|}{\|H_2\|} = 1/45$. L'erreur sur le résultat est $\delta X := X' - X = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$, de norme $\|\delta X\|_\infty = 4$, et l'erreur relative est donc (calculée relativement à la solution de base) $\dfrac{\|\delta X\|_\infty}{\|X\|_\infty} = 2/3$, ou bien (calculée relativement à la solution modifiée) $\dfrac{\|\delta X\|_\infty}{\|X'\|_\infty} = 0,4$. Selon le premier calcul, l'erreur relative a été multipliée par 30, selon le second calcul, elle a été multipliée par 18 (ces deux facteurs d'amplification sont donc de l'ordre de grandeur du conditionnement).

Remarque. Puisque le calcul approché comporte le risque d'accumulation des erreurs d'arrondi, pourquoi (lorsque cela est possible) ne pas calculer exactement, avec des rationnels ? Le problème est que le procédé d'élimination de Gauß peut entraîner une « explosion » (croissance rapide) des numérateurs et dénominateur. Appliquons, par exemple, la procédure Maple **`gausselim`** (qui effectue l'algorithme du pivot) à la matrice $hilb(n)$ en prenant $n = 10, 20, \ldots, 100$. Rappelons que le plus grand entier qui apparaît dans ces matrices est 199. Le coefficient d'indices (n, n) de la matrice triangulaire obtenue est chaque fois de la forme $\pm 1/d$, où le dénominateur d vaut respectivement :

```
> 44914183600
> 25506187972572561213960000
> 28039854179866828629616005775348893600 0
> 20608160372148925033536516752903298030551706251776000
> 76002791635166556774070084664439971770328756104155443723222620000
> 23929250337685557167056728685533706028331577499534324080258482965024172155904000
> 35193627432456346600806183185850139109631057645743383840374677144497677938466412915846579200 0
```

```
> 32458094751717938832714268678884226741407473066651274754
  41275963653805975263096148734960726956328795033600000
> 69142316217523023837935238818391697713057685940211014377
  94742976116343766265300097660852279286292910322409503
  195542000000
> 13761155029966070093165244415949005805790939374575168114
  41697303672278007465521109081410991940974015187190319
  90583585871833617462600000
```

(ce dernier a donc 132 chiffres). Il est clair que les calculs exacts deviennent alors impraticables. On contourne cette difficulté en ayant recours à l'*arithmétique modulaire*, c'est-à-dire que l'on remplace les entiers par leurs images dans un corps fini $\mathbb{Z}/p\mathbb{Z}$; mais cela dépasse le cadre de cet ouvrage.

Lorsque l'on utilise la norme euclidienne $\|-\|_2$ sur $\mathbb{R}^n$ et la norme subordonnée $\|-\|$ sur $M_n(\mathbb{R})$, on peut calculer le conditionnement de la manière suivante. Soit $A \in GL_n(\mathbb{R})$. La matrice $B := {}^tAA$ est symétrique définie positive. Soient $0 < \mu_1 \leqslant \cdots \leqslant \mu_n$ ses valeurs propres et $(X_1, \ldots, X_n)$ une base orthonormée de vecteurs propres. Pour tout $X := \sum_{i=1}^{n} x_i X_i \in \mathbb{R}^n$, on a $BX = \sum_{i=1}^{n} \mu_i x_i X_i$, d'où :

$$\|AX\|_2^2 = {}^tXBX = \sum_{i=1}^{n} \mu_i x_i^2,$$

d'où les inégalités :

$$\sqrt{\mu_1}\,\|X\|_2 \leqslant \|AX\|_2 \leqslant \sqrt{\mu_n}\,\|X\|_2\,.$$

L'égalité peut être réalisée à gauche (en prenant $X := X_1$) et à droite (en prenant $X := X_n$). On en déduit les valeurs (pour la norme $\|-\|$ subordonnée à la norme euclidienne $\|-\|_2$) :

$$\|A\| = \sqrt{\mu_n} \quad \text{et} \quad \left\|A^{-1}\right\| = \sqrt{\mu_1^{-1}}.$$

Le conditionnement, calculé relativement à ces normes, est donc :

$$\mathrm{cond}_2(A) = \sqrt{\frac{\mu_n}{\mu_1}}.$$

Notons que $\mathrm{cond}_2(A) \geqslant 1$, avec égalité si, et seulement si, A est proportionnelle à une matrice orthogonale.

Proposition 27. Pour une matrice inversible symétrique A, le conditionnement calculé relativement à la norme subordonnée à la norme euclidienne est donné par la formule :

$$\mathrm{cond}_2(A) = \rho(A)\rho(A^{-1}). \tag{21}$$

Démonstration. Si A est symétrique, $B = A^2$ a pour valeurs propres les carrés des valeurs propres de A. Dans ce cas, $\sqrt{\mu_n} = \rho(A)$ (rayon spectral de A) et $\sqrt{\mu_1^{-1}} = \rho(A^{-1})$. ∎

Exemple. Considérons, pour tout $\alpha \geqslant 0$, la matrice symétrique :

$$A_\alpha := \begin{pmatrix} 2+\alpha & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2+\alpha & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2+\alpha & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2+\alpha & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2+\alpha & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2+\alpha \end{pmatrix} \in M_n(\mathbb{R}).$$

Il découle de l'exercice II.8.24 de la page 953 que les valeurs propres de A sont $\lambda_1 < \cdots < \lambda_n$, où $\lambda_k := 2 + \alpha - 2\cos\dfrac{k\pi}{n+1} = \alpha + 4\sin^2\dfrac{k\pi}{2(n+1)}$ pour $1 \leqslant k \leqslant n$. Elles sont toutes strictement positives, donc A_α est symétrique définie positive. La plus petite valeur propre est $\lambda_1 = \alpha + 4\sin^2\dfrac{\pi}{2(n+1)}$; la plus grande est $\lambda_n = \alpha + 4\cos^2\dfrac{\pi}{2(n+1)}$.

Le rayon spectral de A est λ_n, celui de A^{-1} est λ_1^{-1} et le conditionnement de A est

$$\mathrm{cond}_2(A) = \frac{\alpha + 4\cos^2\dfrac{\pi}{2(n+1)}}{\alpha + 4\sin^2\dfrac{\pi}{2(n+1)}}.$$

5.1.2 Décomposition LU et stratégies de pivot

Lorsque, dans la méthode de Gauß, on peut choisir à chaque étape pour pivot le premier coefficient disponible (donc le coefficient diagonal), on obtient la *décomposition LU*. Celle-ci a été détaillée dans la section 3.2.1 du module I.3, où a également été énoncée une condition *suffisante* pour que cela se produise : la non nullité des déterminants mineurs principaux $D_0(A), \dots, D_n(A)$. On obtient ainsi une écriture $A = LU$, où la matrice L est triangulaire inférieure unipotente et où la matrice U est triangulaire supérieure. Cette réduction comporte un nombre d'opérations arithmétiques équivalent à $2n^3/3$ et aucune permutation de lignes. La résolution du système (20) se ramène alors à celle des deux systèmes triangulaires : $LY = B$, $UX = Y$, lesquelles ont des coûts en $\mathrm{O}(n^2)$. Dans le cas de *matrices bandes*, c'est-à-dire dont les coefficients sont concentrés autour de la diagonale, le coût de la réduction peut être notablement réduit (exercice II.8.25 de la page 953).

En principe, la décomposition LU est presque sûrement possible : autrement dit, pour toute mesure « raisonnable » sur $GL_n(\mathbb{R})$, l'ensemble des matrices ne vérifiant pas les conditions suffisantes indiquées est de mesure nulle. En pratique, on peut très bien rencontrer un coefficient diagonal nul en position de candidat pivot. Mais surtout, la présence d'un *petit* pivot peut s'avérer aussi gênante numériquement que celle d'un pivot nul, à cause de son effet sur les erreurs d'arrondi.

Exemple. Considérons le système :

$$\begin{cases} 10^{-4}x - y = -1, \\ x + y = 2, \end{cases}$$

dont la solution exacte est

$$x = \frac{1}{1+10^{-4}} = 0{,}99989998\cdots\,,\ y = \frac{1+2\,.\,10^{-4}}{1+10^{-4}} = 1{,}00010001\cdots$$

(développements décimaux périodiques). Les écritures avec trois chiffres significatifs de ces nombres sont $x \approx 1,\ y \approx 1$. Si les calculs eux-mêmes sont faits avec trois chiffres significatifs, en prenant pour pivot le premier coefficient 10^{-4}, on trouve le système équivalent de lignes L_1 et $L_2 - \dfrac{1}{10^{-4}}L_1$:

$$\begin{cases} 10^{-4}x - y = -1, \\ (1+10^4)y = 2+10^4, \end{cases}$$

qui s'arrondit en :

$$\begin{cases} 0{,}0001x - y = -1, \\ 10000y = 20000, \end{cases}$$

d'où la solution $y \approx 1,\ x \approx 0$, ce qui est difficilement admissible.
Si l'on utilise au contraire pour pivot le *deuxième* coefficient de la première colonne, on trouve le système équivalent de lignes $L_1 - 10^{-4}L_2$ et L_2 :

$$\begin{cases} (-1-10^{-4})y = -1, \\ x + y = 2, \end{cases}$$

qui s'arrondit en :

$$\begin{cases} -1{,}00y = -1, \\ x + y = 2, \end{cases}$$

d'où la solution correcte $y \approx 1,\ x \approx 1$.

Point Méthode

On ne choisit donc pas n'importe quel pivot non nul dans la colonne en cours d'examen, mais le plus grand d'entre eux : c'est la stratégie du *pivot partiel*. Son coût n'est pas essentiellement plus grand que celui de la décomposition LU, mais cette méthode est beaucoup plus stable. Le résultat est une décomposition LU non pas de la matrice de départ A mais de la matrice PA, où P est une matrice de permutation.

Dans la stratégie du *pivot total*, on cherche le plus grand pivot disponible parmi tous ceux des colonnes non encore traitées. On opère donc également des permutations sur les colonnes, et le résultat est une décomposition de la matrice PAP', où P et P' sont des matrices de permutation. La stabilité en est un peu améliorée, mais le coût est également un peu augmenté, la programmation est plus lourde, et il faut garder trace du lien entre les variables et les équations. Cette méthode n'est donc pas très utilisée.

5.1.3 Décomposition de Cholesky et décomposition QR

Les méthodes évoquées ici relèvent de l'algèbre bilinéaire. Lorsque la matrice A est symétrique définie positive, la décomposition LU est toujours possible. On lui préfère une variante, la décomposition de Cholesky (proposition 50 de la page 135) : on écrit $A = {}^tTT$, où T est triangulaire supérieure à coefficients diagonaux strictement positifs ; cette décomposition est unique. On en a donné une programmation récursive à la page 378. Quel que soit le style (récursif ou itératif), le nombre d'opérations arithmétiques est équivalent à $n^3/3$, soit deux fois moins que LU ou la stratégie du pivot partiel. Il faut cependant noter que n extractions de racines carrées sont requises, et que celles-ci sont plus imprécises ou plus coûteuses que les autres opérations arithmétiques. Une fois obtenue la décomposition de Cholesky, la résolution du système (20) se ramène à celle des deux systèmes triangulaires : ${}^tTY = B$, $TX = Y$, de coûts $\mathrm{O}(n^2)$.

Exercice 25.

Donner la décomposition de Cholesky de H_2.

Solution.

On écrit $\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix} = \begin{pmatrix} x & 0 \\ y & z \end{pmatrix} \begin{pmatrix} x & y \\ 0 & z \end{pmatrix}$, soit $x^2 = 1$, $xy = 1/2$ et $y^2 + z^2 = 1/3$. Avec les conditions $x, z > 0$, cela donne $T = \begin{pmatrix} 1 & 0 \\ 1/2 & 1/\sqrt{12} \end{pmatrix}$. La réussite du calcul prouve que H_2 est définie positive (voir aussi l'exercice II.8.22 de la page 953).

Sans aucune hypothèse sur $A \in GL_n(\mathbb{R})$, on a démontré, par la méthode de Householder (théorème 47 de la page 173), l'existence d'une décomposition $A = QR$, dans laquelle la matrice Q est orthogonale et la matrice R est triangulaire supérieure à coefficients diagonaux strictement positifs ; cette décomposition est unique. Cette décomposition est essentiellement équivalente à l'orthonormalisation de la base des colonnes de A, mais l'application du procédé d'orthogonalisation de Gram-Schmidt est considérée comme un processus instable. C'est donc la méthode de Householder qui est recommandée dans ce cas.

Exemple. Nous allons vérifier, grâce à Maple, la plus grande stabilité du procédé de Householder par rapport au procédé de Gram-Schmidt. Voici les instructions Maple :

```
> restart: with(linalg):
> n := 5: H := hilbert(n): Hnum := map(evalf,H):
  QRdecomp(H,Q='Q'): Qex := map(evalf,Q):
  QRdecomp(Hnum,Q='Qnum'):
  Q[n,n];Qex[n,n];Qnum[n,n];
```

La matrice H vaut $hilbert(5)$ (calculé exactement, c'est-à-dire en rationnels) et **Hnum** est son évaluation flottante, donc arrondie (par défaut, avec dix chiffres significatifs). La procédure Maple **QRdecomp** utilise le procédé de Householder pour la décomposition $H = QR$. Nous obtenons dans la variable **Q** le facteur Q calculé exactement. Son dernier coefficient $\big($d'indices $(5,5)\big)$, appelé en Maple **Q[n,n]**, vaut ici : $\dfrac{70\sqrt{33001}}{33001}$.

L'évaluation en écriture flottante de **Q** est calculée dans **Qex**, dont le dernier coefficient, écrit **Qex[n,n]**, vaut $0{,}3853314794$ (dans le style anglo-saxon de Maple, cela s'écrit en fait **0.3853314794**). Enfin, si l'on applique la procédure **QRdecomp** à la matrice *numérique* **Hnum**, on obtient le facteur **Qnum** calculé numériquement. Son dernier coefficient vaut, selon Maple, $0{,}3853314641886493 86$: cela concorde plutôt bien. Notons d'ailleurs que Maple donne ici plus de chiffres qu'il n'est raisonnable, et que cela sera encore le cas ci-dessous.

Pour appliquer le procédé de Gram-Schmidt à la liste des colonnes de H, on a recours aux instructions suivantes :

```
> BON := map(normalize,GramSchmidt([col(H,1..n)])):
  BONex := map(evalf,BON):
  BONnum := map(normalize,GramSchmidt([col(Hnum,1..n)])):
  BON[n][n];BONex[n][n];BONnum[n][n];
> evalb(simplify(Q[n,n]) = simplify(BON[n][n]));
```

On trouve dans **BON** la base orthonormale calculée exactement par le procédé de Gram-Schmidt à partir de **H** et dans **BONex** son évaluation flottante. Le dernier coefficient (appelé **BON[n][n]**) du dernier vecteur de la base vaut $\dfrac{\sqrt{13098096900}}{297009}$, dont l'évaluation flottante (appelée **BONex[n][n]**) vaut $0{,}3853314796$. Enfin, le résultat du même calcul effectué numériquement (et appelé (**BONnum[n][n]**) vaut $0{,}3651452037$. La précision est déjà moins parfaite (deuxième chiffre après la virgule erroné). Pour être certain que le coefficient exact donné par Householder ou par Gram-Schmidt est bien le même. c'est-à-dire que $\dfrac{70\sqrt{33001}}{33001} = \dfrac{\sqrt{13098096900}}{297009}$, on utilise la dernière instruction, qui nous rend bien **true**.

Reprenons maintenant les mêmes instructions, mais avec $n := 6$. Le dernier coefficient, calculé exactement par le procédé de Householder, est $\dfrac{252\sqrt{819005}}{819005}$, qui s'évalue en $0{,}2784564385$. Le procédé de Householder appliqué numériquement (calcul flottant) donne $0{,}2784575695471827 59$, ce qui concorde. Le procédé de Gram-Schmidt appliqué exactement donne $\dfrac{\sqrt{6293221315920}}{9009055}$, dont l'évaluation vaut $0{,}2784564384$, ce qui va très bien ; d'ailleurs, Maple confirme l'égalité exacte des expressions avec racines carrées. En revanche, le calcul effectué numériquement donne $0{,}4676011712$, ce qui est très incorrect : erreur relative de 68 % !

Enfin, pour $n := 7$, le coefficient exact converti en écriture flottante vaut $0{,}1994567621$, celui calculé numériquement par Householder $0{,}1994632340682597 48$ et celui calculé numériquement par Gram-Schmidt $-0{,}5185833546$: c'est catastrophique !

Le nombre d'opérations de l'algorithme de Householder est équivalent à $2n^3$ (dont $2n$ extractions de racines carrées). Une fois obtenue la décomposition QR, la résolution du système (20) se ramène à celle d'un seul système triangulaires : $RX = {}^tQB$ (puisque $Q^{-1} = {}^tQ$). Mais l'avantage principal de cette méthode est que l'on peut démontrer qu'elle ne détériore pas sensiblement le conditionnement de A.

Exercice 26.
Donner la décomposition QR de H_2.

Solution. Soient $U := \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}$ et $V := \begin{pmatrix} 1/2 \\ 1/3 \end{pmatrix}$ les colonnes de H_2.

Alors $\|U\|_2 = \dfrac{\sqrt{5}}{2}$, et l'on pose $U' := \dfrac{1}{\|U\|_2} U = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}$.

Ensuite, $W := V - < V, U' > U' = V - \dfrac{4}{3\sqrt{5}} U' = \begin{pmatrix} -1/30 \\ 1/15 \end{pmatrix}$, d'où $\|W\|_2 = \dfrac{\sqrt{5}}{30}$, d'où $W' := \dfrac{1}{\|W\|_2} W = \begin{pmatrix} \frac{-1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}$. La matrice Q de colonnes U', W' est orthogonale et l'on a les relations : $U = \dfrac{\sqrt{5}}{2} U'$ et $V = \dfrac{4}{3\sqrt{5}} U' + \dfrac{\sqrt{5}}{30} W'$. finalement :

$$H_2 = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{5}}{2} & \frac{4}{3\sqrt{5}} \\ 0 & \frac{\sqrt{5}}{30} \end{pmatrix}.$$

5.2 Méthodes itératives

Lorsque les matrices sont grosses, les méthodes directes peuvent être trop coûteuses ou trop instables et l'on préfère parfois des méthodes itératives. Notons que l'on a déjà vu (section 1.1, page 883) un algorithme de division itérative alors qu'il existe un algorithme « direct ». Nous n'aborderons pas les méthodes réellement en usage (méthodes de descente, de gradient), qui s'apparentent à l'optimisation et seront étudiées dans [MPA1] ; mais celles qui suivent sont efficaces et instructives.

Toutes les méthodes qui vont suivre, pour résoudre le système $AX = B$ relèvent de l'idée suivante : on écrit $A = M - N$, où M est inversible. On a alors l'équivalence : $AX = B \Leftrightarrow X = PX + C$, où $P := M^{-1}N$ et $C := M^{-1}B$. On considère le système $X = PX + C$ comme un problème de point fixe, que l'on résout par itération (méthode de Picard) : on pose $X_0 := 0$ et $X_{k+1} := PX_k + C$. Remarquons que c'est une extension de la formule $\dfrac{1}{1-x} = 1 + x + x^2 + \cdots$, car la relation $X = (I_n - P)^{-1}C = C + PC + P^2C + \cdots$ fait précisément apparaître les solutions approchées $X_k = C + PC + P^2C + \cdots + P^{k-1}C$. Il faut, bien entendu, imposer des conditions à P (et donc à la décomposition $A = M - N$) pour que ce processus itératif converge :

Théorème 28. Soit $P \in GL_n(\mathbb{R})$. Il y a équivalence entre les propriétés suivantes :
(i) La suite définie par $X_0 := 0$ et $X_{k+1} := PX_k + C$ converge quel que soit le vecteur $C \in M_{n,1}(\mathbb{R})$.
(ii) On a $\lim\limits_{k\to+\infty} P^k = 0$.
(iii) Le rayon spectral de P vérifie $\rho(P) < 1$.
(iv) Il existe une norme matricielle subordonnée telle que $\|P\| < 1$.
Lorsque ces propriétés sont vérifiées, la matrice $I_n - P$ est inversible et la limite de la suite $(X_k)_{k\in\mathbb{N}}$ du (i) est $(I_n - P)^{-1}C$.

Démonstration. Cela découle immédiatement de l'exercice II.1.32 de la page 472. ∎

On peut évaluer la vitesse de convergence de l'algorithme comme suit. On fixe une norme matricielle subordonnée (le plus souvent, à la norme uniforme sur $\mathbb{R}^n$). Si $X := (I_n - P)^{-1}C$, le reste $X - X_k$ est égal à P^kX, dont la norme est majorée par $\|P\|^k \left\|(I_n - P)^{-1}\right\| \|C\|$. Cette majoration est optimale (autrement dit, l'égalité peut être réalisée). Du fait que le rayon spectral $\rho(P)$ est la borne inférieure des $\|P\|$ (pour toutes les normes matricielles subordonnées), on considère même que l'erreur après k itérations est de l'ordre de $\rho(P)^k$. Il s'agit donc d'une convergence *linéaire*, au sens de la section 1.3, page 889. Le nombre de chiffres exacts (en base b) après k itérations est de l'ordre de $k\big(-\log_b \rho(P)\big)$. On dira donc que le processus converge d'autant plus vite que $\rho(P)$ est plus petit (et ne converge pas si $\rho(P) \geqslant 1$).
Remarquons enfin que l'on ne calcule pas réellement la matrice P et le vecteur C ; on calcule plutôt X_{k+1} en résolvant le système $MX_{k+1} = NX_k + C$, qui est équivalent à l'égalité $X_{k+1} = PX_k + C$. Il faut donc que la matrice M soit telle que le système soit facile à résoudre : par exemple diagonale, ou triangulaire. On dit alors que M est *facilement inversible*, même si l'on n'a pas l'intention de l'inverser.
On note traditionnellement D, E et F les matrices respectivement diagonale, triangulaire inférieure stricte et triangulaire supérieure stricte telles que $A = D - E - F$. Ainsi, pour $i, j \in [\![1, n]\!]$:

$$(D)_{i,j} = \begin{cases} (A)_{i,j} & \text{si } i = j, \\ 0 & \text{sinon,} \end{cases} \quad (E)_{i,j} = \begin{cases} -(A)_{i,j} & \text{si } i > j, \\ 0 & \text{sinon,} \end{cases} \quad (F)_{i,j} = \begin{cases} -(A)_{i,j} & \text{si } i < j, \\ 0 & \text{sinon.} \end{cases}$$

Dans tout ce qui suit, *on suppose la matrice diagonale D inversible* : autrement dit, les coefficients diagonaux de A sont tous non nuls.

5.2.1 Les méthodes de Jacobi et de Gauß-Seidel

Dans la *méthode de Jacobi*, on prend $M := D$, qui est bien facilement inversible ; on prend donc $N := E + F$. L'itération à l'étape k prend la forme $DX_{k+1} = (E + F)X_k + B$. Si

l'on note $(x_1^{(k)}, \ldots, x_n^{(k)}) := X_k$, on a : $x_1^{(0)} := \cdots := x_n^{(0)} := 0$ et, pour tout $k \geqslant 0$:

$$
\begin{aligned}
x_1^{(k+1)} &:= -a_{1,1}^{-1}\left(a_{1,2}x_2^{(k)} + \cdots + a_{1,n}x_n^{(k)} - b_1\right),\\
&\cdots\cdots\\
x_i^{(k+1)} &:= -a_{i,i}^{-1}\left(a_{i,1}x_1^{(k)} + \cdots + \widehat{a_{i,i}x_i^{(k)}} + \cdots + a_{i,n}x_n^{(k)} - b_i\right),\\
&\cdots\cdots\\
x_n^{(k+1)} &:= -a_{n,n}^{-1}\left(a_{1,1}x_1^{(k)} + \cdots + a_{1,n-1}x_{n-1}^{(k)} - b_n\right),
\end{aligned}
$$

où le chapeau marque un terme omis.
Le coût (en nombre d'opérations arithmétiques) de chaque itération est donc $\mathrm{O}(n^2)$. En termes plus algorithmiques, on a besoin d'entretenir en mémoire *deux* vecteurs $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_n)$ et de réaliser successivement deux séries d'affectations. Au départ, $x_1 := \cdots := x_n := y_1 := \cdots := y_n := 0$. Puis, à chaque itération :

$$
\begin{aligned}
y_1 &:= -a_{1,1}^{-1}\left(a_{1,2}x_2 + \cdots + a_{1,n}x_n - b_1\right),\\
&\cdots\cdots\\
y_i &:= -a_{i,i}^{-1}\left(a_{i,1}x_1 + \cdots + \widehat{a_{i,i}x_i} + \cdots + a_{i,n}x_n - b_i\right),\\
&\cdots\cdots\\
y_n &:= -a_{n,n}^{-1}\left(a_{1,1}x_1 + \cdots + a_{1,n-1}x_{n-1} - b_n\right),\\
x_1 &:= y_1,\\
&\cdots\cdots\\
x_i &:= y_i,\\
&\cdots\cdots\\
x_n &:= y_n.
\end{aligned}
$$

En effet, toutes les anciennes valeurs sont nécessaires jusqu'au bout. En pseudo-code, cela donne :

```
pour i de 1 a n faire (x[i] := 0 ; y[i] := 0) ;
répète
  (pour i de 1 a n faire
     (y[i] := - b[i] ;
      pour j de 1 a i-1 faire y[i] := y[i] + a[i,j]*x[i] ;
      pour j de i+1 a n faire y[i] := y[i] + a[i,j]*x[i] ;
      y[i] := y[i]/a[i,i]);
   pour i de 1 a n faire x[i] := y[i])
jusque (* précision obtenue souhaitée *)
```

Dans la *méthode de Gauß-Seidel*, le calcul de chaque composante de X_{k+1} s'effectue *en*

tenant compte des composantes déjà calculées. Ainsi, la formule :

$$x_i^{(k+1)} := -a_{i,i}^{-1}\left(a_{i,1}x_1^{(k)} + \cdots + \widehat{a_{i,i}x_i^{(k)}} + \cdots + a_{i,n}x_n^{(k)} - b_i\right)$$
$$= -a_{i,i}^{-1}\left(a_{i,1}x_1^{(k)} + \cdots + a_{i,i-1}x_{i-1}^{(k)} + a_{i,i+1}x_{i+1}^{(k)} + \cdots + a_{i,n}x_n^{(k)} - b_i\right)$$

est remplacée par la formule :

$$x_i^{(k+1)} := -a_{i,i}^{-1}\left(a_{i,1}x_1^{(k+1)} + \cdots + a_{i,i-1}x_{i-1}^{(k+1)} + a_{i,i+1}x_{i+1}^{(k)} + \cdots + a_{i,n}x_n^{(k)} - b_i\right),$$

qui a bien un sens puisque $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ sont connus au moment du calcul de $x_i^{(k+1)}$. Algorithmiquement, cela signifie que l'on n'a pas besoin de garder les vieilles valeurs, un seul vecteur est nécessaire et une seule passe d'affectation a lieu :

$$x_1 := -a_{1,1}^{-1}\left(a_{1,2}x_2 + \cdots + a_{1,n}x_n - b_1\right),$$
$$\cdots\cdots$$
$$x_i := -a_{i,i}^{-1}\left(a_{i,1}x_1 + \cdots + \widehat{a_{i,i}x_i} + \cdots + a_{i,n}x_n - b_i\right),$$
$$\cdots\cdots$$
$$x_n := -a_{n,n}^{-1}\left(a_{1,1}x_1 + \cdots + a_{1,n-1}x_{n-1} - b_n\right).$$

En pseudo-code, cela donne :

```
pour i de 1 a n faire (x[i] := 0) ;
répète
  (pour i de 1 a n faire
     (x[i] := - b[i] ;
      pour j de 1 a i-1 faire x[i] := x[i] + a[i,j]*x[i] ;
      pour j de i+1 a n faire x[i] := x[i] + a[i,j]*x[i] ;
      x[i] := x[i]/a[i,i]))
jusque (* précision obtenue souhaitée *)
```

L'interprétation matricielle est la suivante : l'itération résout le système :

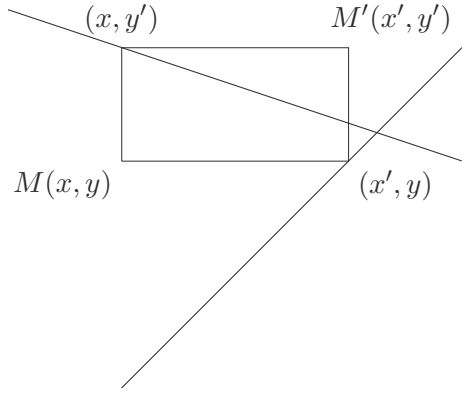
$$DX_{k+1} = EX_{k+1} + FX_k + B \quad \text{qui équivaut à} \quad (D-E)X_{k+1} = FX_k + B.$$

On a donc posé $M := D - E$ et $N := F$. Le coût en temps de chaque itération est le même que dans la méthode de Jacobi, mais le coût en espace est meilleur, puisque l'on ne stocke qu'un vecteur (pour de très gros systèmes, cela n'est pas négligeable).

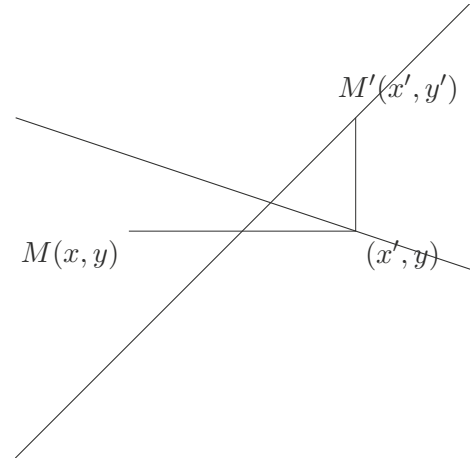
La convergence de la méthode de Gauß-Seidel, qui est liée au rayon spectral de la matrice $P := M^{-1}N = D^{-1}(E+F)$, est souvent meilleure que celle de la méthode de Jacobi, qui est liée au rayon spectral de $P := M^{-1}N = (D-E)^{-1}F$. Nous étudierons plus loin (section 5.2.3) la convergence de ces deux méthodes dans une classe importante de cas particuliers.

On peut en donner une interprétation géométrique simple de ces deux méthodes lorsque $n := 2$. Notons alors x, y les coordonnées du vecteur X. Notons de plus $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ et $B = \begin{pmatrix} e \\ f \end{pmatrix}$.

La résolution du système $AX = B$ revient donc à la recherche du point d'intersection des deux droites D d'équation $ax + by = e$ et D' d'équation $cx + dy = f$.
Dans la méthode de Jacobi, chaque itération transforme le vecteur X de coordonnées (x, y) en un vecteur X' de coordonnées (x', y') comme suit : le point $M(x, y)$ est projeté horizontalement sur un point $(x', y) \in D$ et verticalement sur un point $(x, y') \in D'$. L'horizontale passant par (x, y') et la verticale passant par (x', y) se rencontrent en $M'(x', y')$, nouvelle position, peut-être plus proche de l'intersection. Les calculs de x' et de y' sont indépendants l'un de l'autre.
Dans la méthode de Gauß-Seidel, le point $M(x, y)$ est projeté horizontalement sur un point $(x', y) \in D$; *puis* le point (x', y) ainsi obtenu est projeté verticalement sur le point $M'(x', y') \in D'$. La nouvelle position est nécessairement sur D'. Le deuxième mouvement part de l'arrivée du premier, le deuxième calcul utilise l'arrivée du premier.



Jacobi



Gauss-Seidel

5.2.2 Les méthodes de relaxation

La transformation de la $i^{ème}$ coordonnée dans le processus de Gauß-Seidel peut être écrite (en style algorithmique, c'est-à-dire par mise à jour du vecteur X) :

$$x_i := x_i - r_i, \text{ où } r_i := a_{i,i}^{-1} \left(a_{i,1}x_1 + \cdots + a_{i,n}x_n - b_i\right).$$

On peut considérer r_i comme une *correction* : si l'on avait $AX = B$, ce terme serait nul. Il est a comparer au terme $r(x) := \dfrac{f(x)}{f'(x)}$ du procédé de Newton $x := x - r(x)$ (section 2.2.3). Comme toute correction de trajectoire, on peut imaginer qu'elle va dans la bonne direction, mais trop ou pas assez vite. Les *méthodes de relaxation* consistent à remplacer le terme correctif r_i par un multiple ωr_i, où $\omega \in \mathbb{R}_+^*$ est bien choisi. La même idée pourrait d'ailleurs être appliquée au processus de Jacobi, mais, en pratique, on juge que cela n'apporte rien.
Soit donc $\omega > 0$. La méthode de relaxation dérivée du processus de Gauß-Seidel est donnée

par les formules :

$$x_i^{(k+1)} := x_i^{(k)} - \omega a_{i,i}^{-1}\left(a_{i,1}x_1^{(k+1)} + \cdots + a_{i,i-1}x_{i-1}^{(k+1)} + a_{i,i}x_i^{(k)} + \cdots + a_{i,n}x_n^{(k)} - b_i\right)$$

(l'initialisation est la même que pour Jacobi et Gauß-Seidel). Pour $\omega := 1$, on retrouve donc le processus de Gauß-Seidel. Nous laissons au lecteur le soin de l'écrire en pseudo-code. Le coût en temps et en espace est essentiellement le même que pour le processus de Gauß-Seidel. Le lecteur vérifiera que la version matricielle est la suivante :

$$\left(\frac{1}{\omega}D - E\right)X_{k+1} = \left(\frac{1-\omega}{\omega}D + F\right)X_k + B.$$

On a donc pris $M := \dfrac{1}{\omega}D - E$ et $N := \dfrac{1-\omega}{\omega}D + F$. On dit que, par comparaison avec le processus de Gauß-Seidel, « une partie de D est passée dans N ».

Théorème 29.
Pour que cette méthode de relaxation converge, il est *nécessaire* que $\omega \in]0,2[$.

Démonstration. On a ici :

$$P := M^{-1}N = \left(\frac{1}{\omega}D - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right).$$

Les deux facteurs sont triangulaires et l'on calcule facilement le déterminant :

$$\det P = \frac{\det\left(\dfrac{1-\omega}{\omega}D + F\right)}{\det\left(\dfrac{1}{\omega}D - E\right)} = \frac{\det\left(\dfrac{1-\omega}{\omega}\right)^n \det D}{\det\left(\dfrac{1}{\omega}\right)^n \det D} = (1-\omega)^n.$$

De l'inégalité $|\det P| \leqslant \rho(P)^n$, on déduit alors $|1-\omega| \leqslant \rho(P)$. On en déduit : $\rho(P) < 1 \Rightarrow 0 < \omega < 2$. ∎

On prendra donc *toujours* $\omega \in]0,2[$. Il est difficile de donner, en général, une condition *suffisante* de convergence. Signalons sans démonstration le fait suivant :

PROPRIÉTÉ. *Si la matrice A est symétrique définie positive, la condition $\omega \in]0,2[$ est suffisante pour assurer la convergence de la méthode de relaxation. En particulier, le processus de Gauß-Seidel converge.*

5.2.3 Le cas des matrices tridiagonales

Des matrices tridiagonales apparaissent fréquemment lorsque l'on discrétise des équations différentielles ou aux dérivées partielles ; un exemple est détaillé dans l'exercice II.8.0 de la page 949. Pour de telles matrices, des calculs explicites peuvent être menés à bien.
Nous supposerons donc A *tridiagonale* : $a_{i,j} = 0$ dès que $|j-i| \geqslant 2$. Nous noterons, pour simplifier : $a_i := a_{i,i}$ (pour $1 \leqslant i \leqslant n$) ; $b_i := a_{i,i+1}$ (pour $1 \leqslant i \leqslant n-1$) ; et $c_i := a_{i,i-1}$ (pour $2 \leqslant i \leqslant n$). La matrice E a donc pour seuls coefficients non nuls les $(E)_{i,i-1} = -c_i$ (pour $2 \leqslant i \leqslant n$) et la matrice F a pour seuls coefficients non nuls les $(F)_{i,i+1} = -b_i$

(pour $1 \leqslant i \leqslant n-1$). Nous noterons de plus $J := D^{-1}(E+F)$ (matrice d'itération du processus de Jacobi) et $\mathcal{L}_\omega := \left(\frac{1}{\omega}D - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right)$ (matrice d'itération du processus de relaxation). En particulier, $\mathcal{L}_1$ est la matrice d'itération du processus de Gauß-Seidel.

Lemme 30. (i) Soit $\mu \in \mathbb{C}^*$. Alors μ est une valeur propre de $\mathcal{L}_\omega$ si, et seulement si, $\dfrac{(\mu+\omega-1)^2}{\omega^2\mu}$ est le carré d'une valeur propre de J (ou, ce qui revient au même, une valeur propre de J^2).
(ii) En particulier, l'application $\lambda \mapsto \lambda^2$ établit une bijection entre les paires de valeurs propres non nulles de J et les valeurs propres non nulles de $\mathcal{L}_1$.

Démonstration. Soit $\lambda \in \mathbb{C}^*$. Alors λ^2 est une valeur propre de $\mathcal{L}_\omega$ si, et seulement si, la matrice $\left(\frac{1-\omega}{\omega}D + F\right) - \lambda^2\left(\frac{1}{\omega}D - E\right)$ est non inversible, ce qui équivaut à : $\dfrac{\lambda^2+\omega-1}{\omega}D - \lambda^2 E - F$ non inversible.
Comme λ est non nul, la matrice $\mathrm{Diag}(1, \lambda, \ldots, \lambda^{n-1})$ est inversible et elle conjugue $\dfrac{\lambda^2+\omega-1}{\omega}D - \lambda^2 E - F$ à $\dfrac{\lambda^2+\omega-1}{\omega}D - \lambda E - \lambda F$; en effet, cette conjugaison laisse D invariant, conjugue E à $\lambda^{-1}E$ et F à λF. Ce calcul facile (laissé au lecteur) repose sur le fait que A est tridiagonale. La matrice $\dfrac{\lambda^2+\omega-1}{\omega}D - \lambda E - \lambda F$ est non inversible si, et seulement si, $\dfrac{\lambda^2+\omega-1}{\omega\lambda}$ est une valeur propre de $D^{-1}(E+F) = J$. On en conclut déjà que le spectre de J est symétrique (l'opposé d'une valeur propre de J est une valeur propre de J). Puis, en élevant au carré, on trouve la condition énoncée. L'application au cas où $\omega := 1$ est alors immédiate. ∎

Théorème 31. Les processus de Jacobi et de Gauß-Seidel convergent ou divergent tous les deux ; s'ils convergent, le processus de Gauß-Seidel est deux fois plus rapide. Plus précisément :

$$\rho(\mathcal{L}_1) = \rho(J)^2. \tag{22}$$

Démonstration. L'égalité $\rho(\mathcal{L}_1) = \rho(J)^2$ est une conséquence du lemme. Les autres affirmations découlent alors de la section 5.2. ∎

Pour étudier la convergence du processus de relaxation, nous ferons une hypothèse :

HYPOTHÈSE. *On suppose dorénavant que toutes les valeurs propres de J sont réelles.*

C'est en particulier vrai dans le cas (fréquent) où la matrice A est symétrique, car J l'est alors aussi. Lorsque x appartient au spectre de J^2, l'égalité $\dfrac{(\mu+\omega-1)^2}{\omega^2\mu} = x$ équivaut

à $T_x(\mu) = 0$, où :

$$T_x(\mu) := \mu^2 + \big(2(\omega - 1) - \omega^2 x\big)\mu + (\omega - 1)^2$$

est un trinôme de discriminant $\Delta := x\omega^2\big(x\omega^2 - 4(\omega - 1)\big)$.
Supposons d'abord que $\rho(J) \geqslant 1$. Ni le processus de Jacobi ni le processus de Gauß-Seidel ne convergent donc. Pour $x := \rho(J)^2$ (qui est une valeur propre de J^2), $T_x(1) = \omega^2(1-x)$ est négatif ou nul, ce qui montre que T_μ a des racines réelles et que 1 est situé entre ces racines. L'une de ces racines est donc supérieure ou égale à 1, donc $\rho(\mathcal{L}_\omega) \geqslant 1$: le processus de relaxation ne converge donc pour aucune valeur de ω.
Supposons maintenant que $\rho(J) < 1$. Le processus de Jacobi et le processus de Gauß-Seidel convergent donc. On peut alors démontrer le fait suivant, que nous admettrons :

PROPRIÉTÉ. *Pour tout $\omega \in]0, 2[$, la méthode de relaxation converge. Le rayon spectral $\rho(\mathcal{L}_\omega)$ est minimum pour la valeur suivante du paramètre de relaxation :*

$$\omega = \omega_0 := \frac{2}{1 + \sqrt{1 - \rho(J)^2}}. \tag{23}$$

Le rayon spectral optimum vaut :

$$\rho(\mathcal{L}_{\omega_0}) = \omega_0 - 1 = \frac{1 - \sqrt{1 - \rho(J)^2}}{1 + \sqrt{1 - \rho(J)^2}}. \tag{24}$$

Exemple. Nous reprenons les notations de l'exemple de la page 937.
On a donc $D = (2 + \alpha)I_n$ et $J = I_n - \dfrac{1}{2+\alpha}A_\alpha$. Les valeurs propres de J sont donc les $1 - \dfrac{\lambda_k}{2+\alpha} = \dfrac{2}{2+\alpha}\cos\dfrac{k\pi}{n+1}$, pour $1 \leqslant k \leqslant n$. Le rayon spectral est donc $\rho(J) = \dfrac{2}{2+\alpha}\cos\dfrac{\pi}{n+1}$, et le processus de Jacobi converge, ainsi que le processus de Gauß-Seidel et la méthode de relaxation pour tout $\omega \in]0, 2[$. Un cas important est détaillé dans l'exercice II.8.0 de la page ci-contre.

Exercice type corrigé

II.8.0 On va étudier l'équation $-y'' + cy = f$, où $c > 0$ et où f est de classe $\mathcal{C}^2$ sur $[0,1]$. On impose les *conditions aux limites* $u(0) = u(1) = 0$. Ce n'est donc pas un problème de Cauchy, mais il découle de l'exercice 17 de la page 836 qu'il admet une solution unique u. Il est évident que cette solution est de classe $\mathcal{C}^4$ sur $[0,1]$, et l'on notera : $M := \left\|u^{(4)}\right\|_\infty$ (norme uniforme sur $[0,1]$).

(i) Pour toute fonction φ de classe $\mathcal{C}^4$ sur $[0,1]$ et tout $x \in]0,1[$, démontrer la formule :

$$\forall h > 0 \text{ (assez petit)} \ , \ \exists \xi_x \in]x-h, x+h[\ : \ \varphi(x+h)+\varphi(x-h)-2\varphi(x) = h^2\varphi''(x)+\frac{h^4}{12}\varphi^{(4)}(\xi_x).$$

(ii) Pour un entier $n \geqslant 2$, on pose $h := \dfrac{1}{n}$, $x_i := ih$ et $f_i := f(x_i)$ $(0 \leqslant i \leqslant n)$. On suppose trouvés des réels $u_0, \dots, u_n$ tels que $u_0 = u_n = 0$ et, pour tout $i \in [\![1, n-1]\!]$:

$$-\frac{u_{i-1} + u_{i+1} - 2u_i}{h^2} + cu_i = f_i.$$

On note $e_i := u(x_i) - u_i$ (« erreur d'approximation »). Vérifier que $e_0 = e_n = 0$ et démontrer, pour tout $i \in [\![1, n-1]\!]$:

$$-\frac{e_{i-1} + e_{i+1} - 2e_i}{h^2} + ce_i = \frac{h^2}{12}v_i \ , \ \text{ où } |v_i| \leqslant M.$$

(iii) Quelle est la matrice $A \in M_{n-1}(\mathbb{R})$ telle que $AU = B$, où $U := {}^t(u_1, \dots, u_{n-1})$ et $B := {}^t(h^2 f_1, \dots, h^2 f_{n-1})$? Calculer les valeurs propres de A et vérifier que c'est une matrice symétrique définie positive. Qu'en déduire au sujet de l'existence des u_i ?
(Utiliser les exemples des pages 937 et 948 en prenant garde aux tailles des matrices.)

(iv) À l'aide du système linéaire satisfait par le vecteur d'erreur $E := (e_1, \dots, e_{n-1})$, majorer la norme euclidienne de celui-ci puis sa norme uniforme. Qu'en déduire au sujet de la méthode de discrétisation ?

(v) Estimer et comparer les performances des méthodes de Jacobi, de Gauß-Seidel et de relaxation optimale appliquées au système $AU = B$.

Solution. (i) La formule de Taylor donne ici :

$$\varphi(x+h) = \varphi(x) + h\varphi'(x) + \frac{h^2}{2}\varphi''(x) + \frac{h^3}{6}\varphi'''(x) + \frac{h^4}{24}\varphi^{(4)}(\xi'), \text{ où } \xi' \in]x, x+h[\quad \text{et}$$

$$\varphi(x-h) = \varphi(x) - h\varphi'(x) + \frac{h^2}{2}\varphi''(x) - \frac{h^3}{6}\varphi'''(x) + \frac{h^4}{24}\varphi^{(4)}(\xi''), \text{ où } \xi' \in]x-h, x[.$$

On additionne ces deux formules, on retranche $2\varphi(x)$ et l'on applique le théorème des valeurs intermédiaires à $\varphi^{(4)}$ sur $[\xi', \xi'']$, ce qui donne la formule voulue.

(ii) Les égalités $e_0 = e_n = 0$ sont évidentes. Soit $i \in [\![1, n-1]\!]$:

$$\begin{aligned} -\frac{e_{i-1} + e_{i+1} - 2e_i}{h^2} + ce_i &= +\frac{u_{i-1} + u_{i+1} - 2u_i}{h^2} - cu_i - \frac{u(x_{i-1}) + u(x_{i+1}) - 2u(x_i)}{h^2} + cu(x_i) \\ &= -f_i - \left(u''(x_i) + \frac{h^2}{12}u^{(4)}(\xi_i)\right) + cu(x_i), \end{aligned}$$

pour un $\xi_i \in]x_{i-1}, x_{i+1}[$, d'après la question (i) et les définitions $e_j := u(x_j) - u_j$. Par hypothèse, $f_i = f(x_i) = -u''(x_i) + cu(x_i)$, et il ne reste que le terme $-\dfrac{h^2}{12}u^{(4)}(\xi_i)$, auquel on applique la majoration $\left|u^{(4)}(\xi_i)\right| \leqslant M$.

(iii) Selon les notations des exemples évoqués, la matrice $A := A_{ch^2}$ *mais d'ordre* $n-1$ est telle que $AU = B$. La matrice A a pour valeurs propres les $\lambda_k := ch^2 + 4\sin^2\dfrac{k\pi}{2n}$ pour $1 \leqslant k \leqslant n-1$ (insistons sur le fait qu'il faut remplacer n par $n-1$ dans toutes les formules). Elle est symétrique définie positive (exemple de la page 937), et, en particulier, inversible. Les conditions de la question (ii) définissent donc bien un unique vecteur $(u_1, \ldots, u_{n-1})$.

(iv) On a $AE = V$, où $V := {}^t\left(\dfrac{h^4}{12}v_1, \ldots, \dfrac{h^4}{12}v_{n-1}\right)$, d'où $\|E\|_2 \leqslant \left\|A^{-1}\right\| \|V\|_2$.

Pour la norme subordonnée à la norme euclidienne, on a, d'après l'exemple de la page 937, $\left\|A^{-1}\right\| = \rho(A^{-1}) = \lambda_1^{-1}$. Par ailleurs, $\|V\|_2 \leqslant M\dfrac{h^4}{12}\sqrt{n}$. Comme $\|E\|_\infty \leqslant \|E\|_2$, on obtient (en remplaçant h par $1/n$) la majoration :

$$\|E\|_\infty \leqslant \frac{M}{12C_n}\, n^{-3/2}, \text{ où } C_n := c + 4n^2 \sin^2\frac{\pi}{2n}\cdot$$

Comme $\lim\limits_{n\to+\infty} C_n = c + \pi^2 > 0$, l'erreur est en $\mathrm{O}(n^{-3/2})$. Lorsque l'on raffine la subdivision, la méthode de discrétisation converge.

(v) On sait déjà que les trois méthodes convergent (exemple de la page 948). Le rayon spectral de J est $\rho(J) = \dfrac{2}{2+ch^2}\cos\dfrac{\pi}{n}\cdot$ Comme $h = \dfrac{1}{n}$ et $\cos\dfrac{\pi}{n} = 1 - \dfrac{\pi^2}{2n^2} + \mathrm{o}\left(\dfrac{1}{n^2}\right)$, on trouve :

$$\rho(J) = 1 - \frac{c+\pi^2}{2n^2} + \mathrm{o}\left(\frac{1}{n^2}\right).$$

Il faut donc (par exemple) $-\dfrac{1}{\ln\rho(J)} \equiv \dfrac{2n^2}{c+\pi^2}$ itérations pour diviser l'erreur par e. Pour la méthode de Gauß-Seidel, comme $\rho(\mathcal{L}_1) = \rho(J)^2$, il faut $-\dfrac{1}{\ln\rho(\mathcal{L}_1)} \equiv \dfrac{n^2}{c+\pi^2}$ soit deux fois moins d'itérations. Pour la relaxation optimale, on a (en vertus de calculs de développements limités laissés au lecteur) :

$$\rho(\mathcal{L}_\omega) = \frac{1-\sqrt{1-\rho(J)^2}}{1+\sqrt{1-\rho(J)^2}} = 1 - 2\frac{\sqrt{c+\pi^2}}{n} + \mathrm{o}\left(\frac{1}{n}\right).$$

Le nombre d'itérations nécessaires pour diviser l'erreur par e est donc $-\dfrac{1}{\ln\rho(\mathcal{L}_\omega)} \sim \dfrac{n}{\sqrt{c+\pi^2}}$: on a gagné un ordre de grandeur !

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.8.1 Soit $y \in [0,1[$. Justifier la possibilité d'écrire $y = \sum_{j \geqslant 1} d_j b^{-j}$, avec $0 \leqslant d_j \leqslant b-1$. Montrer que cette écriture est unique, sauf dans le cas où y est de la forme $\dfrac{a}{b^p}$, qui admet alors deux écritures exactement.

II.8.2 $*$ Proposer et étudier un algorithme de calcul de la racine carrée entière dans l'esprit de l'algorithme itératif d'inversion. On adaptera aux entiers l'algorithme de Babylone (exemple de la page 904).

II.8.3 On suppose que u_n tend vers x avec une erreur $u_n - x \sim \dfrac{c}{n}$, resp. $u_n - x \sim \dfrac{c}{\sqrt{n}}$, resp. $u_n - x \sim \dfrac{c}{\ln n}$. Combien d'étapes sont nécessaires pour avoir k chiffres significatifs ?

II.8.4 Sur la couverture du célèbre ouvrage *Computational Methods of Linear Algebra*, de V.N. Faddeeva (éditions Dover), on trouve les chiffres suivants :

```
30000000
32000000
32300000
32370000
32374000
32374600
32374650
32374658
```

Qu'est-ce qui est ainsi illustré ?

II.8.5 Montrer que, s'il existe $\mu \in]0,1[$ tel que $|x_{n+1} - r| \sim \mu |x_n - r|$, resp. $M \in \mathbb{R}_+$ tel que $|x_{n+1} - r| \leqslant M |x_n - r|^2$, alors la convergence est linéaire, resp. quadratique.

II.8.6 Avec les notations de la section 1.4, démontrer l'égalité : $I_{2p+1} = \dfrac{4^p}{(2p+1)\binom{2p}{p}}$.

En déduire la *formule de Wallis* :

$$\frac{\pi}{2} = \lim_{n \to +\infty} \frac{2}{1}\frac{2}{3} \cdots \frac{2n}{2n-1}\frac{2n}{2n+1}.$$

Exercices sur la section 2

II.8.7 Écrire l'algorithme de recherche dichotomique d'un zéro de f.

II.8.8 Appliquer la méthode de la sécante à $f(x) := x^3 - 6x + 3$ sur $\left[0, \sqrt{2}\right]$.

II.8.9 On cherche à calculer une racine $p^{ème}$ de $\alpha > 0$ comme point fixe. On pose donc, pour un certain réel t (à déterminer), $\varphi(x) := (1-t)x + \dfrac{\alpha t}{x^{p-1}}$.
Comment choisir t pour avoir un point fixe superattractif ?

II.8.10 Localiser les racines réelles de $x^5 - x - 1 = 0$ dans des segments de longueur 1. Évaluer celle du milieu à 10^{-2} près par les méthodes de dichotomie, de la sécante et de Newton. Peut-on envisager une méthode de point fixe ?

II.8.11 $*$ Si $\varphi(x) = x + \lambda x^k + o(x^k)$, $\lambda \neq 0$, $k \in \mathbb{N} \setminus \{0, 1\}$, étudier la stabilité du point fixe 0.

Exercices sur la section 3

II.8.12

Interpoler x^4 sur $[-1, 1]$ aux points $-1, 0, 1$, puis aux points de Tchebychev $-\dfrac{\sqrt{3}}{2}, 0, \dfrac{\sqrt{3}}{2}$.
Dans les deux cas, comparer l'erreur effective à l'erreur théorique.

II.8.13 $*$ Combien d'opérations sont nécessaires au calcul direct des L_i et de $L[A; f]$? Combien sont nécessaires par la méthode des différences divisées ?

II.8.14 $*$ (i) Démontrer que, pour $\alpha > 0$ assez petit, on a l'inégalité :

$$2(\ln 2 - 1) > \int_0^1 \ln(\alpha^2 + t^2)\, \mathrm{d}t.$$

(ii) Obtenir la même conclusion par application du théorème 120 de la page 631.

II.8.15 $**$ On se donne des points distincts $a_1, \ldots, a_n \in K$ (corps commutatif) et des multiplicités $\mu_1, \ldots, \mu_n \in \mathbb{N}^*$. Montrer que l'application qui, à un polynôme $P \in K[X]$, associe la famille $\left(P^{(j)}(a_i)\right)_{\substack{1 \leqslant i \leqslant n \\ 0 \leqslant j < \mu_i}}$, a pour noyau l'idéal engendré par $\prod (X - a_i)^{\mu_i}$. En déduire qu'elle induit un isomorphisme de $K_{N-1}[X]$ sur K^N, où $N = \sum_{i=1}^{n} \mu_i$ (*interpolation de Lagrange-Hermite*).
Lorsque $K := \mathbb{R}$, étendre à cette situation la formule (7) du théorème 7 de la page 906 (il faut d'abord étendre le lemme 12).

II.8.16 $**$ Que deviennent les différences divisées $f[a_0, \ldots, a_n]$ lorsque les points d'interpolation *confluent* ? Par exemple, lorsque a_0 et a_1 tendent vers une valeur commune ?

Exercices sur la section 4

II.8.17 Quel est l'ordre exact de la méthode de Simpson ?

II.8.18 Calculer les poids dans la méthode de Newton-Cotes d'ordre 8.

II.8.19 (i) Calculer $I := \displaystyle\int_0^1 \frac{\mathrm{d}x}{1+x^2}$ par la méthode du point milieu, puis par la méthode des trapèzes. Évaluer l'erreur commise et la comparer à celle estimée dans le texte.
(ii) Calculer I par la méthode de Simpson, puis par la méthode de Gauß avec trois points. Évaluer l'erreur commise.

II.8.20 $*$ Déterminer une formule de quadrature élémentaire à deux points sur $[a,b]$ d'ordre le plus grand possible.

II.8.21 $**$ Expliciter la formule de quadrature de Gauß associée aux polynômes de Tchebychev pour tout entier n.

Exercices sur la section 5

II.8.22 $*$ Démontrer que la matrice de Hilbert H_n est symétrique définie positive.

II.8.23 Donner, en fonction de $\mathrm{cond}(A)$, une majoration de l'erreur relative $\dfrac{\|\delta X\|}{\|X\|}$ lorsque la matrice A est approchée par $A+\delta A$ et le vecteur B par $B+\delta B$.
On supposera l'erreur relative $\dfrac{\|\delta A\|}{\|A\|}$ « suffisamment petite ».

II.8.24 Démontrer que le déterminant de la matrice

$$\begin{pmatrix} 2X & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2X & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2X & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2X & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2X & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2X \end{pmatrix}$$

(d'ordre n) est égal au polynôme de Tchebychev U_n (module sur les polynômes de [L1]). En déduire les valeurs propres de la matrice A_α de l'exemple de la page 937.

II.8.25 (i) Estimer le coût de la décomposition LU d'une matrice tridiagonale.
(ii) Donner la décomposition LU de $A := \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 2 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 2 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 2 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 2 \end{pmatrix}$.

II.8.26 À quelle vitesse converge le processus $X_{k+1} := PX_k + C$ lorsque $\rho(P) = 0$?

II.8.27 Dans la formule (24) de la page 948, vérifier que le rayon spectral optimal est inférieur ou égal à celui de $\mathcal{L}_1$. À quelle condition le processus de Gauß-Seidel est-il le processus optimal de relaxation ?

II.8.28 Dans l'exemple de la page 948, vérifier que le spectre de J est symétrique (*i.e.* invariant par $\lambda \mapsto -\lambda$).

Bibliographie

[Berger] **M. Berger** *Géométrie*, Éditions Nathan, 1990

[Ciarlet] **P. G. Ciarlet** *Introduction à l'analyse matricielle et à l'optimisation* éditions Masson, 1990

[Demailly] **Jean-Pierre Demailly** *Analyse numérique et équations différentielles*, éditions des Presses Universitaires de Grenoble, 1996

[DW2] **C. Deschamps, A. Warusfel.** *Mathématiques* $2^{ème}$ *année*, Dunod Paris, 2001.

[D1] **J. Dieudonné.** *Éléments d'analyse, Tome I, fondements de l'analyse moderne*, Gauthier-Villars Paris, 1969.

[D2] **J. Dieudonné.** *Calcul infinitésimal*, Collection Méthodes, Hermann Paris, 1968.

[Efimov] **N. Éfimov** *Géométrie supérieure*, Éditions de Moscou, 1981.

[G2] **R. Godement** *Analyse mathématique II*, Springer-Verlag, 1998.

[Knuth] **D. E. Knuth** *Art of Computer Programming, volume II*, éditions Addison Wesley), 1981

[Landau] **L. Landau et E. Lifchitz** *Théorie des champs*, Éditions de Moscou, 1970.

[Muller] **J.-M. Muller**, *Arithmétique des ordinateurs*, éditions Masson, 1989

[R3] **E. Ramis, C. Deschamps, J. Odoux** *Cours de Mathématique spéciales, tome* 3 *: topologie et éléments d'analyse*, Masson Paris, 1976.

[L1] **J.P. Ramis, A. Warusfel.** *Mathématiques, Tout-en-un pour la licence, niveau L1*, Dunod Paris, 2005.

[L2-L3] **J.P. Ramis, A. Warusfel.** *Mathématiques, Tout-en-un pour la licence, niveau L2-L3,*, Dunod Paris, à paraître en 2014.

[MPA1] **J.P. Ramis, A. Warusfel, F. Moulin** *Cours de Mathématiques pures et appliquées, volume* 1 *algèbre et géométrie*, De Boeck Bruxelles, 2010.

[MPA] **J.P. Ramis, A. Warusfel, F. Moulin** livre en préparation dans la série *Mathématiques pures et appliquées*, De Boeck Bruxelles, à paraître.

Indications

I.1 : Compléments d'algèbre

I.1.1 Si $\dfrac{a}{b} = \dfrac{a'}{b'}$, vérifier que :

$$v_p(a) - v_p(b) = v_p(a') - v_p(b').$$

En revanche, $\dfrac{a}{b} = \dfrac{a'}{b'}$ n'entraîne pas :

$$\frac{a^2+1}{b^2+1} = \frac{a'^2+1}{b'^2+1}.$$

I.1.2 En ce qui concerne $\mathcal{R}$, seule la transitivité n'est pas triviale. Pour vérifier que $E/\mathcal{R}$ est un anneau, utiliser (après l'avoir justifié) le fait que trois éléments quelconques de $E/\mathcal{R}$ ont nécessairement des représentants définis sur un même intervalle J.

I.1.3 Non : trouver $a \not\equiv_9 b$ tels que $3a \equiv_9 3b$.

I.1.4 (i) Surjectivité : la classe de (a,b) admet pour antécédent $a-b$. Injectivité : elle découle du texte (les éléments de M étant simplifiables). (ii)
L'application $M \times M \to A \times (A \setminus \{0\})$ (inclusion naturelle) passe au quotient et induit une bijection du groupe des fractions de M avec K^* et cette bijction est un morphisme. Attention ! Il faut noter multiplicativement la loi du groupe commutatif M.

I.1.5 Vérifier que $(a,b) \mapsto f(a)f(b)^{-1}$ passe au quotient.

I.1.6 Ce sont les diviseurs des éléments de S. Pour $S = A^*$, on trouve $S^{-1}A = A$. Si $s \in S$ était nilpotent, on aurait $0 = s^n \in S$. Quand $0 \in S$, on a $0 = 1$ dans $S^{-1}A$ et ce dernier est trivial.

I.1.7 (i) Montrer que $(a,s) \mapsto f(s)^{-1}f(a)$ passe au quotient en un morphisme g. (ii) et (iv) découlent de (i). Dans (iii), la surjectivité vient en écrivant $a/t = at'/s$, où $s = tt'$. Attention : il y a un risque d'ambiguïté de notation entre $a/s \in S^{-1}A$ et $a/s \in T^{-1}A$ (tant qu'on n'a pas justifié l'identification).

I.1.8 Si $a^{-1}b = h \in H$, alors $b = ah$, donc $bH = ahH = aH$. Si $aH = bH$, alors $b \in bH = aH$, donc $b = ah$, $h \in H$, donc $a^{-1}b = h \in H$.

I.1.9 Le fait que $\iota_a \circ \iota_b = \iota_{ab}$ résulte d'un calcul facile. Le noyau est le centre de G. La dernière affirmation est tautologique.

I.1.10 Considérer le noyau de $G \to G' \to G'/H'$. La deuxième assertion est fausse en général, vraie si f est surjectif.

I.1.11 Le noyau est formé des $a \in \mathbb{Z}$ tels que $n = mp$ divise ma.

I.1.12 La stabilité de HG' vient de l'égalité $(h_1 g_1')(h_2 g_2') = (h_1 h_2')(g_1' g_2')$, où $h_2' := g_1' h_2 {g_1'}^{-1} \in H$ par hypothèse. Il suffit en fait de supposer que H est inclus dans le normalisateur de G', c'est-à-dire le sous-groupe de G formé des g tels que $gG'g^{-1} = G'$.

I.1.13 La classe $\overline{a} \in \mathbb{C}/\mathbb{Z}$ de $a \in \mathbb{C}$ est de torsion si, et seulement s'il existe $n \in \mathbb{N}^*$ tel que $n\overline{a} = 0$, *i.e.* si $a \in \mathbb{Q}$. Pour un tel élément et un morphisme f $\mathbb{C}/\mathbb{Z}$ dans $\mathbb{C}$, on a $nf(\overline{a}) = 0$, donc $\overline{a} = 0$. Par l'isomorphisme usuel entre $\mathbb{C}^*$ et $\mathbb{C}/\mathbb{Z}$, on en déduit que tout morphisme de groupes de $\mathbb{C}^*$ dans $\mathbb{C}$ est trivial sur le groupe des racines de l'unité.

I.1.14 Puisque $n = mp$, on a $n\mathbb{Z} \subset m\mathbb{Z}$ d'où le passage au quotient. L'élément $a \pmod m$ admet $a \pmod n$ pour antécédent.

I.1.15 L'anneau des entiers de Gauß.

I.1.16 La preuve donnée pour les idéaux des anneaux commutatifs s'étend sans problème.

I.1.17 Si $M \in I$ est non nulle, calculer les $E_{i,j} M E_{k,l}$.

I.1.18 L'application $E \to \dfrac{E+F}{F}$ est surjective car la classe de $e+f$ admet pour antécédent e.

I.1.19 On sait (module sur les polynômes de [L1]) que $<P>$ et $K_{n-1}[X]$ sont supplémentaires.

I.1.20 Soient $\operatorname{Ker}\lambda_1, \ldots, \operatorname{Ker}\lambda_k$ ces hyperplans. Appliquer la proposition à $x \mapsto \big(\lambda_1(x), \ldots, \lambda_k(x)\big)$.

I.1.21 (i) L'étape clé est le fait que $\sum \alpha_i d_i = \in F$ entraîne que tous les α_i sont nuls (en notant $(d_i) := \mathcal{D}$). (ii) En écrivant la matrice par blocs correspondant aux deux bases juxtaposées, la condition est que l'un des blocs extra-diagonaux soit nul. Les deux matrices demandées sont les deux blocs diagonaux. (Le dernier bloc ne sert pas ici.)

I.1.22 Les éléments de $I \cap J$ sont les polynomes sans terme constant et à coefficients pairs, d'où $I \cap J = 2XA$. Les idéaux $(I+J)^2$ et $I^2 + IJ + J^2$ sont tous deux égaux à $4A + 2XA + X^2 A$.

I.1.23 On a $I^n = <a^n, a^{n-1}b, \ldots, ab^{n-1}, b^n>$.

I.1.24 (i) Si I est un tel idéal, et si $(i_1, i_2) \in I$, alors $(i_1, 0) = (i_1, i_2)(1, 0)$ et $(0, i_2) = (i_1, i_2)(0, 1)$ sont éléments de I. (ii) Le produit $B_1 \times B_2$ de deux anneaux est un corps si, et seulement si, l'un de ces anneaux est un corps et l'autre est trivial. (ii) Le produit $B_1 \times B_2$ de deux anneaux est intègre si, et seulement si, l'un de ces anneaux est intègre et l'autre est trivial.

I.1.25 Si $P = \prod P_i^{r_i}$ est la décomposition de P en produit de facteurs premiers, $K[X]/ < P > \simeq \prod K[X]/ < P_i^{r_i} >$ admet des nilpotents si, et seulement si, l'un des r_i est $\geqslant 2$. Dans le cas contraire, on a bien un produit de corps.

I.1.26 (i) Ce corps a pour éléments 0, 1, $x := \mathrm{cl}(x)$ et $1+x$ (avec $0, 1 \in \mathbb{F}_2$) : il faut écrire la table d'addition et la table de multiplication. (ii) Prendre $\mathbb{F}_2[X]/<P>$ où $\deg P = 3$ et P irréductible (par exemple $X^3 + X + 1$). De même, prendre $\mathbb{F}_3[X]/<Q>$ où $\deg Q = 2$ et Q irréductible.

I.1.27 (ii) S'il y avait une fonction $f \in I$ qui ne s'annule pas sur $\left[a, \dfrac{a+b}{2}\right]$ et une fonction $g \in I$ qui ne s'annule pas sur $\left[\dfrac{a+b}{2}, b\right]$ la fonction $f^2 + g^2 \in I$ serait inversible.

I.1.28 Appliquer le théorème de Krull à l'idéal $<a>$.

I.1.29 Vérifier successivement que $i(I)$ est inclus dans $S^{-1}I$, que ce dernier est un idéal et que tout idéal contenant $i(I)$ contient $S^{-1}I$.

I.1.30 Si $(a/s)(b/t) \in S^{-1}\mathcal{P}$, il existe $u \in S$ tel que $uab \in \mathcal{P}$, d'où $a \in \mathcal{P}$ ou $b \in \mathcal{P}$ puisque $u \notin \mathcal{P}$.

I.1.31 Appliquer l'exercice I.1.30 à $S := \{x^n \mid n \in \mathbb{N}\}$ et invoquer le théorème de Krull.

I.1.32 La racine de I est l'image réciproque du nilradical de A/I.

I.1.33 Il faut prouver que $I + JK = A \Rightarrow I + J = I + K = A$, ce qui découle des inclusions $I + JK \subset I + J$, $I + JK \subset I + K$.

I.1.34 (i) découle de l'équivalence

$$(a_1, a_2)(b_1, b_2) = (1, 1) \Leftrightarrow a_1 b_1 = 1 \text{ et } a_2 b_2 = 1$$

(attention, ce n'est pas le même 1 !).

I.1.35 Décompositions sur $\mathbb{Z}$:

$$X^4 - 1 = (X-1)(X+1)(X^2+1),$$

$X^4 + 1$ est irréductible,

$$X^4 + X^2 + 1 = (X^2 + X + 1)(X^2 - X + 1).$$

I.1.36 Notant $I := \mathrm{Ker}\, f$, le noyau est $IA[X]$ (polynômes à coefficients dans I).

I.1.37 Soit $b := -a$. Alors, si $(1 - bX)P = 1$, on a $P(0) = 1$ et $(1 - b^n X^n)P = 1 + bX + \cdots + b^{n-1}X^{n-1}$. Si $n > \deg P$, le coefficient du terme de degré n du membre gauche est $-b^n$ qui est donc nul. L'existence de $\mathfrak{M}$ est conséquence du théorème de Krull.

I.1.38 Si $x = \dfrac{a}{b}$ (fraction irréductible), le noyau est $<bX - a>$.

I.1.39 Si V est un K-espace vectoriel et un anneau (avec la même addition), c'est une K-algèbre si, et seulement si, l'application $(x, y) \mapsto x \star y$ est K-bilinéaire.

I.1.40 Le noyau $\mathrm{Ker}\, \varphi$ est un idéal propre du corps K, donc il est trivial. La condition de centralité entraîne $x \star (\lambda.y) = \lambda.(x \star y)$, l'égalité $(\lambda.x) \star y = \lambda.(x \star y)$ venant de l'associativité de A.

I.1.41 L'algèbre $K[X]/<X^2-1>$ est isomorphe à $K\times K$ par le morphisme $\overline{P}\mapsto(P(-1),P(1))$.
L'algèbre $K[X]/<X^2>$ admet la description concrète $K\oplus K\varepsilon$ avec $\varepsilon^2=0$.
L'algèbre $K[X]/<X^2+1>$ est un corps si -1 n'est pas un carré dans K. Si $-1=a^2$, elle est isomorphe à $K\times K$ par le morphisme $\overline{P}\mapsto(P(-a),P(a))$.

I.1.42 Si $x\neq 0$, considérer l'application linéaire injective $\lambda_x : y\mapsto xy$.

I.1.43 Si K est de caractéristique 0, non ; s'il est de caractéristique $p>0$, on a $D^p=0$.

I.1.44 On trouve selon le cas une algèbre isomorphe à : $\mathbb{C}$; $\mathbb{C}[X]/<X^2>$; $\mathbb{C}\times C$.

I.1.45 (iii) Soit $x:=a+b\delta$ la classe de $a+bX$ dans $K[X]/<X^2-d>$. La matrice de λ_x dans la base $(1,\delta)$ est $\begin{pmatrix}a & db\\ b & a\end{pmatrix}$, d'où $N(x)=a^2-db^2$.
Cela s'applique à tous les exemples, sauf celui de $\mathbb{H}$ (qui se traite aussi en écrivant la matrice).

I.1.46 (i) S'inspirer du cas des nombres algébriques, *cf.* [L1].

I.1.47 Pour l'inversibilité, plagier la preuve de la proposition 51 de la page 32. Pour l'intégrité, considérer les ordres.

I.1.48 On trouve $\displaystyle\sum_{n\geqslant 0}\frac{j^{n+1}-j^{2n+2}}{j-j^2}X^n$.

I.1.49 Pour prouver $G_k/G_{k+1}\simeq K$, considérer le morphisme $a\mapsto(1+aX^k)\pmod{G_{k+1}}$.

I.1.50 Invoquer l'exercice 21 de la page 37.

I.1.51 Sous l'hypothèse indiquée (et en admettant le résultat de Hadamard), l'annulation de l'un des $H_{m+j}^{(k)}$ pour $1\leqslant j\leqslant r$ entraîne celle de ses deux voisins.

I.1.52 Utiliser l'exercice I.1.51 de la page 55.

I.1.53 Raisonner par continuité.

I.1.54 Par identification des coefficients, on trouve $\displaystyle\sum_{n\geqslant 0} n!\,X^{n+1}$.

I.1.55 Si $K[[X]]$ est compact, le fermé discret K est fini. Réciproquement, on applique le critère de Bolzano-Weierstraß. Soit (f_k) une suite de terme général $f_k:=\sum a_n^{(k)}X^n$. Il y a une infinité d'indices k tels que les $a_0^{(k)}$ aient une même valeur $a_0\in K$ (puisque K est fini). Parmi ces indices, il y en a une infinité tels que les $a_1^{(k)}$ aient une même valeur $a_1\in K$, etc. Alors $a_0+a_1X+\cdots$ est un point d'accumulation.

I.1.56 S'inspirer de l'exercice 18 de la page 34. Cette série sera étudiée dans la section « Partitio numerorum » du chapitre de combinatoire de [MPA].

I.1.57 Par calcul direct, on trouve :

$$P_{i,k}(a_1,\dots,a_k)=\sum_{\substack{j_1+\cdots+j_i=k\\ j_1,\dots,j_i\in\mathbb{N}^*}} a_{j_1}\cdots a_{j_i}.$$

I.1.58 On *compose* f à droite par la série formelle $0 \in XK[[X]]$.

I.1.59 Dériver $\sin^2 + \cos^2$.

I.1.60 Dériver $\sqrt{1-4X} = 1 - 2\sum c_n X^{n+1}$.

I.1.61 Remplacer v_n par $1 + w_n$ et appliquer massivement l'inégalité triangulaire.

I.1.62 On trouve $\binom{p-1/2}{n} \sim \dfrac{C\,(-1)^n}{n^{p+1/2}}$, où $C := \left(\dfrac{-1}{4}\right)^p \dfrac{(2p)!}{p!}\dfrac{1}{\sqrt{\pi}}$.

I.2 : Actions de groupes

I.2.3 Penser au théorème de la base incomplète.

I.2.4 On peut, par exemple, se ramener à l'exercice précédent à l'aide de translations.

I.2.6 Prendre l'isobarycentre (centre de gravité) de l'orbite d'un point de Y sous G.

I.2.8 Si a n'est pas une racine de l'unité, $A = \mathbb{C}$ ne contient que les constantes. Si a est une racine primitive de l'unité d'ordre n, c'est-à-dire $a := \exp(2ki\pi/n)$ où $n \in \mathbb{N}^*$ et $k \in [\![0, n-1]\!]$ sont premiers entre eux, $A = \mathbb{C}[X^n]$ (polynômes en X^n).

I.2.9 Pour 3, c'est essentiellement le théorème de la base incomplète.

I.2.10 Pour 2, considérer une orbite Y de X sous K.

I.2.11 L'orbite d'une paire de droites $\{D, D'\}$ est caractérisée par la distance de D à D' si D, D' sont parallèles, et par l'angle de D et D' sinon, cet angle appartenant à $]0, \pi/2]$.

I.2.13 Soit g un élément du centre. Montrer d'abord que g est une translation, en exprimant le fait que g commute avec chaque translation (*cf.* la formule (6) de la page 64). Si par exemple $g = t_{\vec{u}}$, montrer que $\vec{u}$ est fixé par tout automorphisme (linéaire) de E, direction de X.

I.2.17 Soient H un sous-groupe d'indice n de G et K l'intersection des conjugués de H dans G. Majorer l'indice de K dans G en fonction de n. Montrer ensuite que, si $d \in \mathbb{N}^*$ et si Γ est un groupe fini d'ordre d, il n'y a qu'un nombre fini de morphismes de G dans Γ (utiliser une partie génératrice finie de G).

I.2.20 Le sous-groupe de $\mathfrak{S}_9$ engendré par les 3-cycles $(1\,2\,3), (4\,5\,6), (7\,8\,9)$ et par l'élément $(1\,4\,7)(2\,5\,8)(3\,6\,9)$ convient.

I.2.22 Pour 2, l'orbite d'une paire $\{A, A'\}$ d'éléments de E est caractérisée par le cardinal de $A \cap A'$. En 3, l'image de φ est contenue dans le groupe alterné $\mathfrak{A}(E)$ si, et seulement si, le coefficient binomial $\binom{N-2}{k-1}$ est impair.

I.2.23 Pour chaque entier k tel que $1 \leqslant k \leqslant n/2$, les produits de k transpositions à supports deux à deux disjoints forment une de ces classes de conjugaison, et on les obtient toutes ainsi. Le cardinal de cette classe de conjugaison correspondant à k est $n(n-1)\cdots(n-2k+1)/(2^k k!)$.

I.2.26 Construire la suite $(s_1, \dots, s_r)$ par récurrence.

I.2.32 Le nombre de tels coloriages est $(q^6 + 3q^4 + 12q^3 + 8q^2)/24$.

I.2.33 L'ordre de G est 48, réponse contenue dans l'énoncé de 3 !

I.2.35 Nombre de bases : $\prod_{k=0}^{n-1} (q^n - q^k)$.

Nombre de drapeaux : $\prod_{k=1}^{n} (1 + q + \cdots + q^k)$.

I.2.36 L'ordre est $q^n \prod_{k=0}^{n-1} (q^n - q^k)$.

I.3 : Algèbre bilinéaire

I.3.1 Nous supposerons $n \geqslant 3$.
(i) Les matrices $E_{ij} + E_{nn} - E_{in} - E_{nj}$, avec $1 \leqslant i \leqslant n-1, 1 \leqslant j \leqslant n-1$ $(i \neq j)$, forment une base de $\mathcal{M}_0$. La dimension de $\mathcal{M}_0$ est donc égale à $(n-1)^2$.
(ii) Soit J la matrice de $M_n(\mathbb{R})$ dont tous les coefficients sont égaux à 1, pour toute matrice A de $\mathcal{M}_1$, $A - lJ$, si $l = l_i(A)$, appartient à $\mathcal{M}_0$. On en déduit que $\mathcal{M}_1$ est de dimension égale à $(n-1)^2 + 1$.
(iii) Les formes linéaires Tr et d restreintes à $\mathcal{M}_0$, sont indépendantes, donc $\mathcal{M}_2$ est de dimension égale à $(n-1)^2 - 2$.
(iv) En raisonnant comme en (ii), on obtient que la dimension de $\mathcal{M}_3$ est égale à $n(n-2)$.

I.3.3 a) Suggestion : Utiliser les matrices élémentaires.

b) Appliquer les résultats du théorème 11 de la page 102.

I.3.5 On montre que les formes (f_k) forment un système libre.

Supposons $\sum_{k=0}^{n} \lambda_k f_k = 0$, pour tout P dans $\mathbb{R}_n[T]$, on a :

$$\int_0^1 \sum_{k=0}^{n} \lambda_k t^k P(t) \, \mathrm{d}t = 0.$$

Si $P(t) = \sum_{k=0}^{n} \lambda_k t^k$,

alors $\int_0^1 P(t)^2 \, \mathrm{d}t = 0$ et l'on en déduit que P est nul (P^2 est une fonction continue et positive sur $[0,1]$).

Pour tout $k \in [\![0, n]\!]$ et pour tout $P = \sum_{j=0}^{n} a_j t^j$,

$f_k(P) = \int_0^1 \sum_{j=0}^{n} a_j t^{k+j} \, \mathrm{d}t$, donc

$f_k(P) = \sum_{j=0}^{n} \frac{a_j}{k+j+1}$.

La matrice de passage de la base duale de $\mathcal{B}$ à la base (f_k) est donc :

$$M = \left(\frac{1}{j+k+1} \right)_{0 \leqslant j,k \leqslant n}.$$

I.3.9 On obtient les résultats suivants.

- Φ_1 est de rang 2 et de signature $(1,1)$ (on peut « réécrire » les sommations).
- Φ_2 est définie positive (appliquer une décomposition de Gauß).
- Φ_3 est positive de signature $(n-1, 0)$ (on pourra rechercher son cône isotrope).
- Φ_4 est de rang 2 et de signature $(2,0)$, si $\alpha \neq 0$.
- Φ_5 est de rang 1 ou 2 (discuter suivant les valeurs des θ_i).
- Le rang et la signature de Φ_6 dépendent de $s = \sum_{i=1}^{n} \alpha_i^2$:
  - si $s < 1$, alors Φ est définie positive,

- si $s = 1$, la signature de Φ est $(n-1, 0)$,
- si $s > 1$, la signature de Φ est (n-1,1).

I.3.10 Grâce à une décomposition de Gauß, on montre :

$$\Phi(x) = \sum_{k=1}^{n} (kx_k + x_{k+1} + \cdots + x_n)^2.$$

I.3.12 Pour tout $x \in E$, $f(x) = \sum_{i=1}^{n} a_i x_i$, avec $(a_1, \ldots, a_n) \in K^n$, donc

$$\Phi(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j x_i x_j$$

et la matrice $M = (a_i a_j)_{1 \leqslant i,j \leqslant n}$ est de rang 1 $(f \neq 0)$.
Réciproquement, si M est une matrice de $S_n(K)$ de rang 1, il existe $(a_1, a_2, \ldots, a_n) \in K^n \setminus \{0\}$ et $(b_1, b_2, \ldots, b_n) \in K^n \setminus \{0\}$ tels que :

$$m_{i,j} = a_i b_j = a_j b_i.$$

Soit $i_0 \in [\![1, n]\!]$ tel que $a_{i_0} \neq 0$, pour tout $j \in [\![1, n]\!]$, $b_j = \frac{b_{i_0}}{a_{i_0}} a_j$, donc $m_{ij} = \lambda a_i a_j$ et :

$$\Phi(x) = \lambda \left(\sum_{i=1}^{n} a_i x_i \right)^2.$$

La forme quadratique Φ est donc proportionnelle au carré d'une forme linéaire non nulle.

I.3.13 On a

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_i x_j}{a_i + a_j} = \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^{+\infty} x_i x_j e^{-(a_i + a_j)t} \, \mathrm{d}t = \int_0^{+\infty} \left(\sum_{i=1}^{n} x_i e^{-a_i t} \right)^2 \mathrm{d}t.$$

On remarque ensuite que les réels $(a_1, a_2, \ldots, a_n)$ étant deux à deux distincts les fonctions $(e^{-a_i t})_{1 \leqslant i \leqslant n}$ sont indépendantes.

I.3.15 Suggestion.

On utilise un raisonnement par récurrence sur n et l'on se ramène au cas d'une matrice de la forme :

$$A = \begin{pmatrix} 1 & 0 & \ldots & 0 & a_1 \\ 0 & 1 & \ldots & 0 & a_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & a_n \\ a_1 & a_2 & \ldots & a_n & a_{n+1} \end{pmatrix}$$

On a $\det A = a_{n+1} - \sum_{i=1}^{n} a_i^2$.

Pour tout x dans R^{n+1}, on a :

$${}^t X A X = \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} a_i x_i x_{n+1} + a_{n+1} x_{n+1}^2.$$

On en déduit :

$${}^t X A X = \sum_{i=1}^{n} (x_i + a_i x_{n+1})^2 + (\det A) x_{n+1}^2.$$

D'après l'hypothèse on sait que le déterminant de A est strictement positif, d'où le résultat .

I.4 : Espaces préhilbertiens

I.4.2 Soit $(e_1, e_2 \ldots e_n)$, la base canonique de $\mathbb{R}^n$. On applique l'orthogonalisation de Gram-Schmidt à la base $(e_1 - e_n, e_2 - e_n, \ldots, e_{n-1} - e_n)$ de l'hyperplan. On obtient ainsi une base orthonormée de l'hyperplan :

$\sqrt{\frac{k}{k+1}} \big(e_k - \frac{1}{k}(e_1 + \cdots + e_{k-1} + e_n) \big)$,
$k \in [\![1, n-1]\!]$.

I.4.7 Une réponse est :

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

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

puis

$$P = \begin{pmatrix} \frac{5}{\sqrt{30}} & 0 & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{30}} & \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} \\ \frac{-1}{\sqrt{30}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} \end{pmatrix}$$

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 7 \end{pmatrix}.$$

I.4.8 Le minimum recherché est égal au carré de la distance de l'application $t \mapsto \ln(t)$ à $\mathbb{R}_n[T]$, l'espace E des applications continues sur $]0,1]$, à valeurs dans $\mathbb{R}$, de carré intégrable étant muni du produit scalaire : $(f|g) = \displaystyle\int_0^1 f(t)g(t)\,\mathrm{d}t$.

On suit une démarche analogue à celle qui a été utilisée dans l'exercice type. On introduit ici la fraction rationnelle $F(t) := \displaystyle\sum_{k=0}^{n} \frac{a_k}{k+1+t} + \frac{1}{(t+1)^2}$, et l'on montre que :

$$\inf f = -F'(0) = \frac{1}{(n+1)^2}.$$

I.4.9 b) Pour tout g dans V et h dans W, on a $(gh + g'h') = gh'' + g'h' = (gh')'$, d'où :

$$(g|h) = g(1)h'(1) - g(0)h'(0) = 0.$$

Soit h l'application :

$$t \mapsto \frac{\beta\sinh(t) - \alpha\sinh(t-1)}{\sinh(1)}.$$

Elle est dans W et $h(0) = \alpha$, $h(1) = \beta$.
Pour tout f de E, si $\alpha = f(0)$ et $\beta = f(1)$, l'application $f - h$ est dans V ; V et W sont donc supplémentaires orthogonaux.
c) On en déduit :

$$\inf_{f\in F} = \int_0^1 (f^2(t) + f'^2(t))\,\mathrm{d}t = \|h\|^2 = \frac{(\alpha^2+\beta^2)\cosh(1) - 2\alpha\beta}{\sinh(1)}.$$

I.4.11 a) Pour tout $P \in E$,

$$u(P)(x) = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} \int_0^1 t^k P(t)\,\mathrm{d}t,$$

donc $u(P) = 0$ si et seulement si pour tout $k \in [\![0,n]\!]$, $\displaystyle\int_0^1 t^k P(t)\,\mathrm{d}t = 0$, un tel polynôme P orthogonal à tout l'espace est nul.
En appliquant la formule de Fubini (calcul d'intégrales doubles) on montre que u est auto-adjoint.
b) Les relations proposées s'établissent facilement en calculant les coordonnées dans la base $(P_0, P_1 \ldots, P_n)$ du polynôme $(x+y)^n$, y étant donné dans $\mathbb{R}$.
c) Par définition :

$$\mathrm{Tr}(u) = \sum_{k=0}^{n} (u(P_k)|P_k) = \int_0^1 \sum_{k=0}^{n} \lambda_k \big(P_k(t)\big)^2\,\mathrm{d}t;$$

d'après b), $\displaystyle\sum_{k=0}^{n} \lambda_k (P_k(t))^2 = (2t)^n$, donc :

$$\mathrm{Tr}\,u = \frac{2^n}{(n+1)}.$$

On procède de même pour calculer $\mathrm{Tr}(u^2)$ et l'on obtient $\mathrm{Tr}(u^2) = \dfrac{2^{2n+1}-1}{(n+1)(2n+1)}$.

I.4.14 a) Considérer le produit scalaire sur $\mathbb{R}^n$: $(x,y) \mapsto {}^tXAY$, et appliquer la formule de Cauchy-Schwarz.

b) Noter que $\dfrac{\det A_i}{\det A}$ est le terme d'ordre i de la diagonale de A^{-1}.

I.4.17 (i) On vérifie que u^* est l'application $x \mapsto (\,x \mid b\,)\,a$.

(ii) On note que $f(x) = 0$ si et seulement si $(x|a) = 0$ et $(x|b) = 0$. Si $n \geqslant 3$, 0 est valeur propre de f de multiplicité $(n-2)$, l'espace propre étant $a^\perp \cap b^\perp$. La matrice associée à la restriction de f au plan $\mathrm{Vect}(a,b)$, la base choisie étant (a,b), est :

$$\begin{pmatrix} (a|b) & \|b\|^2 \\ \|a\|^2 & (a|b) \end{pmatrix}$$

La droite dirigée par $\|b\|a + \|a\|b$ est la droite propre associée à la valeur propre $(a|b) + \|b\|\|a\|$.
La droite dirigée par $\|b\|a - \|a\|b$ est la droite propre associée à la valeur propre $(a|b) - \|b\|\|a\|$.

I.4.19 Soit $n \geqslant 3$. Soient $(C_1, C_2, \ldots, C_n)$ les vecteurs colonnes de M.
Soit $V = (1, 1, \ldots, 1)$ dans $\mathbb{R}^n$.
Pour tout $k \in [\![1, n]\!]$, $C_k = C_1 + (k-1)V$, M est donc de rang 2, M étant diagonalisable, on sait que zéro est valeur propre de M de multiplicité $(n-2)$, et l'espace propre associé est défini par les deux équations : $x_1 + x_2 + \cdots + x_n = 0$ et $x_1 + 2x_2 + \cdots + kx_k + \cdots + nx_n = 0$.
On recherche alors les valeurs propres non nulles de M en remarquant que les vecteurs propres associés sont dans le plan engendré par C_1 et V. Ce sont les deux racines du trinôme :

$$\lambda^2 - n^2\lambda - \frac{1}{12}n^2(n^2-1).$$

I.4.20 On vérifie que :

$$\Psi(x) = (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^2 + x_4^2.$$

En appliquant la méthode décrite dans la remarque pratique du corollaire 59, on détermine la base $\mathcal{B}$ de $\mathbb{R}^4$ formée des vecteurs :
$v_1 = (1,0,0,0)$, $v_2 = (1,1,0,0)$, $v_3 = (1,1,1,0)$, $v_4 = (1,1,1,1)$, lesquels sont à la fois Ψ orthonormés et Φ orthogonaux, la matrice associée à Φ, dans cette base est $\mathrm{Diag}\,(1,2,-1,0)$.

I.4.22 Une réponse est :

$$P = \begin{pmatrix} \dfrac{2}{3} & \dfrac{4+3i}{3\sqrt{10}} & \dfrac{4-3i}{3\sqrt{10}} \\ \dfrac{1}{3} & \dfrac{2-6i}{3\sqrt{10}} & \dfrac{2+6i}{3\sqrt{10}} \\ -\dfrac{2}{3} & \dfrac{5}{3\sqrt{10}} & \dfrac{5}{3\sqrt{10}} \end{pmatrix}$$

et $D = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 3i & 0 \\ 0 & 0 & -3i \end{pmatrix}$.

I.4.25 Suggestion :
On montrera que $B = (I - A)(I + A)^{-1}$.

I.5 : Réduction des matrices

I.5.2 En 2, le noyau est formé des matrices diagonales, et l'image est formée des matrices de trace nulle.

I.5.5 Si $d \in \mathbb{N}^*$ et si $b_0, \ldots, b_{d-1} \in L$ sont des inconnues, l'égalité :

$$A^d + b_{d-1}A^{d-1} + \cdots + b_1 A + b_0 I_n = 0$$

se traduit par un système linéaire de n^2 équations à d inconnues, à coefficients et second membre *dans* K.

I.5.6 Il y a huit solutions.

I.5.8 On peut par exemple utiliser le critère de diagonalisabilité portant sur le polynôme minimal.

I.5.9 Pour la diagonalisabilité, comparer les polynômes annulateurs de A et de u.

I.5.10 La nécessité est vraie sur n'importe quel corps, on peut utiliser l'exercice I.5.7. La suffisance est fausse sur $\mathbb{R}$. Il faut donc, en dernière analyse, utiliser le théorème de d'Alembert-Gauß.

I.5.11 Considérer la somme des dimensions des sous-espaces propres, pour u ou pour u^p.

I.5.15 Si $p \neq q$, A est diagonalisable si, et seulement si, $t \neq 1 - n^2/4$. Si $p = q$, A est diagonalisable si, et seulement si, $t^2 - 2(n-1)t + n^2 + 2n - 3 \neq 0$.

I.5.17 On trouve $A^{1789} = \begin{pmatrix} 1 & 0 & 0 \\ 895 & 0 & 1 \\ 894 & 1 & 0 \end{pmatrix}$.

I.5.20 Se ramener d'abord au cas où I est fini. Raisonner ensuite par récurrence sur $n + \operatorname{card} I$. Tout revient en fait à trouver un $x \in E$ qui soit vecteur propre de chaque u_i.

I.5.21 On pourra distinguer deux cas, suivant que u soit inversible ou ne le soit pas.

I.5.22 En supposant que $AB - BA = B$, calculer $AB^k - B^k A$ pour tout entier $k \geqslant 1$.

I.5.23 Pour 1, observer que $P_s^{n!} = I_n$.

I.5.24 Se rappeler que $\lambda - q = \lambda$ pour tout $\lambda \in K$.

I.5.29 On trouve

$$D = \begin{pmatrix} 5 & -4 & 4 & -4 \\ -4 & 9 & -8 & 6 \\ 2 & 6 & -5 & 2 \\ 12 & -8 & 8 & -9 \end{pmatrix},$$

$$N = \begin{pmatrix} -4 & -1 & 0 & 1 \\ 8 & 2 & 0 & -2 \\ 4 & 1 & 0 & -1 \\ -8 & -2 & 0 & 2 \end{pmatrix}.$$

Une réduite de Jordan de A est :

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

I.5.31 Pour la nécessité, on pourra utiliser la réduction de Jordan.

I.5.32 Soit $P := X^k - (a_{k-1}X^{k-1} + \cdots + a_1 X + a_0)$ le « polynôme caractéristique » de la relation de récurrence considérée. La condition nécessaire et suffisante est que, pour chaque racine λ de P, ou bien $|\lambda| < 1$, ou bien $|\lambda| = 1$ et λ est racine simple de P.

I.5.33 La condition nécessaire et suffisante est que, si l'on applique à u le théorème de Jordan de la page 253, il n'y ait qu'un seul bloc correspondant à une valeur propre donnée. Cela revient à dire que $\mu_u = \chi_u$, ou encore, compte tenu de l'exercice I.5.31, qu'il existe un vecteur $x \in E$ tel que $\big(x, u(x), \ldots, u^{n-1}(x)\big)$ soit une base de E.

I.5.35 Pour la suffisance, on peut utiliser la réduction de Jordan.

I.5.36 Une première solution consiste à se ramener au cas d'un bloc de Jordan.

Une deuxième solution est de considérer, pour toute matrice $M \in M_n(\mathbb{C})$, l'égalité $AM = MA$ comme un système linéaire homogène en les coefficients de M, puis de raisonner topologiquement, en appliquant l'exercice 7 de la page 412.

I.5.37 On montre d'abord que u est nilpotent (resp. diagonalisable) dès que A possède cette propriété. Notons par exemple u_A l'endomorphisme u associé à A. Ensuite, lorsque u est diagonalisable, on écrit la décomposition de Dunford $A = D + N$ de A, et l'on obtient la décomposition de Dunford $u_A = u_D + u_N$ de u_A.

I.5.38 Il y a d'abord les couples (a, M), où M est nilpotente et a est quelconque. Il y a d'autre part les couples (a, M) obtenus comme suit. D'abord a est une racine de l'unité, *i.e.* il existe un entier $p \geqslant 2$ tel que $a^p = 1$. Ensuite M est semblable à une matrice diagonale par blocs de la forme $\mathrm{Diag}(A, aA, \ldots, a^{p-1}A)$ ou $\mathrm{Diag}(A, aA, \ldots, a^{p-1}A, N)$, A étant une matrice carrée inversible et N étant une matrice carrée nilpotente.

I.6 : Groupes classiques

I.6.1 La dimension de E/H est 1.

I.6.2 Raisonner par opérations élémentaires. Pour $\mathbf{Tsu_n}(K)$, il suffit d'ajouter les matrices diagonales. Remarquer qu'une matrice commute avec $T_{i,j}(\alpha)$ si, et seulement si, elle commute avec $E_{i,j}$.

I.6.3 La classe de A est dans le centre du groupe concerné si, et seulement si, les matrices $AT_{i,j}(\alpha)$ et $T_{i,j}(\alpha)A$ sont proportionnelles pour tous $i \neq j$ et $\alpha \in \mathbb{K}$.

I.6.4 (i) G/H abélien si, et seulement si, tous les commutateurs de G/H sont triviaux, c'est-à-dire si ceux de G sont dans H. (ii) Comme $D(GL_n(K)) \subset \mathrm{Ker}\,\varphi$, on peut factoriser φ.

I.6.5
Lorsque $K = \mathbf{F}_2$, $\det A \neq 0 \Leftrightarrow \det A = 1$. Comme $GL_2(\mathbf{F}_2)$ est non commutatif et a six éléments, il est isomorphe à $\mathfrak{S}_3$, dont le groupe dérivé est le groupe alterné (voir aussi l'exercice I.6.11).

I.6.6 Il s'agit de traductions mécaniques d'énoncés d'un langage dans un autre.

I.6.7 Oui si, et seulement si, $r < \dim E$.

I.6.8 On trouve $\mathrm{card}\ L_r(E) = (q^n - 1)\cdots(q^n - q^{n-r+1})$. Le cardinal du stabilisateur est $q^{r(n-r)}(q^{n-r} - 1)\cdots(q^{n-r} - q^{n-r-1})$, qui est égal à $(q^n - q^r)\cdots(q^n - q^{n-1})$.

I.6.9 Dans des bases adaptées, on trouve des matrices de la forme
$M_{X,A_1,A_2} := \begin{pmatrix} A_1 & X \\ 0 & A_2 \end{pmatrix}$,
$A_1 \in GL_r(K)$, où $X \in M_{r,n-r}(K)$ et $A_2 \in GL_{n-r}(K)$. Les trois actions demandées ont pour matrice respective A_1, A_2 et M_{0,A_1,A_2}.

I.6.10 (ii) On retrouvera les résultats de l'exercice I.2.0 de la page 88.

I.6.11 (i) Tout découle de l'égalité $\mathbf{F}_2^* = \{1\}$. (ii) Chacun des trois vecteurs non nuls engendre l'une des trois droites. (iii) Comparer à l'exercice I.6.5.

I.6.12 (ii) Pour tout x, montrer que son image est invariante par G ; et que, si x est invariant par G, son image est x. Utiliser le fait que la trace d'un projcteur est égale à son rang multiplié par l'unité du corps.

I.6.13 Le noyau et l'image de φ sont stables. Soit λ une valeur propre de φ ; appliquer la première question à $\varphi - \lambda \operatorname{Id}_E$.

I.6.14 (i) Traduction mécanique, puis propriété du rang d'une famille dans un espace vectoriel. (ii) Pour l'implication non évidente, considérer le sous espace affine X' engendré puis expliciter un $g \neq \operatorname{Id}_X$ qui induise l'identité sur X'.

I.6.15 (i) Il est clair que $\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \in G$.
Pour la réciproque, considérer $(0, \dots, 0, 1)$.
(ii) et (iii) Utiliser la description matricielle.

I.6.16 (i) Utiliser l'exercice 6 de la page 271.

I.6.17 (i) $SL_n(K)$ est fermé dans un connexe ; $GL_n^+(\mathbb{R})$ est l'image réciproque de $\mathbb{R}_+^*$ par $\det$. (iii) Pour montrer que $SL_n(\mathbb{Q})$ est dense dans $SL_n(\mathbb{R})$, approcher $A \in SL_n(\mathbb{R})$ par une suite de matrices A_k rationnelles ; à partir d'un certain rang, $\det A_k \neq 0$ et l'on peut considérer les $D_1(1/\det A_k)A_k$.

I.6.18 Les conditions $\lambda_i + \dfrac{a_i}{p} = \lambda_j + \dfrac{a_j}{p}$ avec $i \neq j$ et $p \geqslant 1$ définissent une union dénombrable d'hyperplans, qui est donc de mesure nulle.

I.6.19 Écrire les développements limités à l'ordre 1 de $(A+H)(B+K)$ et de $(A+H)^2$.

I.6.20 Puisque D est inversible, le terme constant de son polynôme caractéristique est non nul et l'on déduit du théorème de Cayley-Hamilton que $D^{-1} \in K[D]$.

I.6.21 (ii) Toute matrice orthogonale est de la forme $\pm A$, où $A \in SO_n(\mathbb{R})$.

I.6.22 Un élément du centre est diagonalisable dans toute base orthonormale.

I.6.23 Existence : raisonner par opérations élémentaires. Unicité dans le cas inversible : écrire que $Q_2 = Q_1 R$ où $R = R_1 R_2^{-1}$.

I.6.24 Dans la décomposition $M = OR$, décomposer $R = DT$. Attention ! L'unicité de la décomposition $M = OR$ et celle de la décomposition $R = DT$ n'entraînent pas de manière purement formelle l'unicité de la décomposition $M = ODT$ (mais c'est tout de même facile).

I.6.25 On commencera par établir que l'application $A \mapsto A^2$ de $H_n^{++}(\mathbb{C})$ dans lui-même est bijective ; on s'inspirera alors de la preuve du théorème 24 de la page 283.

I.6.26 La réponse est oui, sauf si $\dim E = 2$ et a, b, c deux à deux distincts.

I.6.27 (i) Par exemple $x_1 = \cos t_1$, $x_i = (\sin t_1) y_i$ pour $i \geqslant 2$ et l'on recommence avec $\sum y_i^2 = 1$. (ii) Si $M \in SU_2(\mathbb{C})$ a pour lignes L_1, L_2, alors $L_1 \perp \overline{L_2}$ pour la forme bilinéaire usuelle $aa' + bb'$ et donc $L_1 = (\lambda, \mu)$ et $\overline{L_2} = (-\mu, \lambda)$; on calcule alors le déterminant.
L'homéomorphisme associe à cette matrice le point $(\operatorname{Re}\lambda, \operatorname{Im}\lambda, \operatorname{Re}\mu, \operatorname{Im}\mu)$.

I.6.28 (i) Sa limite est dans $O_n(\mathbb{R})$ et la sous-suite correspondante de $(S_k)_{k\in\mathbb{N}}$ converge vers une matrice symétrique positive.

I.6.29 On trouve $\begin{pmatrix} \cos a & -\sin a \\ \sin a & \cos a \end{pmatrix}$.

I.6.30 Utiliser le théorème 21 pour la première assertion et l'orthodiagonalisabilité pour les deux autres.

I.6.31 (i) Considérer $q \circ \varphi$ comme un polynôme en t et effectuer la division euclidienne par le polynôme unitaire $t^2 - (x^2 + y^2 + z^2)$, ce qui ramène au cas de $A(x, y, z) + B(x, y, z)t$ identiquement nul sur le cône de lumière. (iii) Utiliser le théorème de la moyenne pour les intégrales de fonctions vectorielles.

I.6.32 (i) L est connexe car paramétré par $a \in \mathbb{R}$ et $SO_{1,1}(\mathbb{R})$ ne l'est pas car la fonction $(a_{i,j}) \mapsto a_{1,1}$ a pour image $\mathbb{R}^*$.

I.6.33 Expliciter les homographies associées à $T_{1,2}(\alpha)$ et $D_i(\lambda)$.

I.6.34 L'équation au point fixe $h(z) = z$ équivaut à l'équation du second degré $z(cz + d) = az + b$ qui admet une racine double z_0 ou deux racines distinctes z_1, z_2. Dans le premier cas, conjuguer h par $c/(z - z_0)$, dans le second cas par $(z - z_1)/(z - z_0)$.

I.6.35 (i) Oui. (ii) Non.

I.6.36 On trouve en général :

$\{x, 1-x, 1/x, x/(x-1), 1/(1-x), 1-1/x\}$

et, dans les cas particuliers $x \in \{0, 1, \infty, -1, 1/2, \frac{-1\pm \mathrm{i}\sqrt{3}}{2}\}$, une orbite plus petite (stabilisateur non trivial).

I.6.37 Les doubles transpositions $(ab)(cd)$ (et l'identité).

I.6.38 Même réponse pour le noyau.

I.6.39 L'action de g étant continue, elle laisse stable l'adhérence $\overline{\mathcal{H}}$ de $\mathcal{H}$, donc aussi sa frontière $\partial\mathcal{H} := \overline{\mathcal{H}} \setminus \mathcal{H} = \mathbf{P}^1(\mathbb{R})$.

I.6.40 Conjuguer par l'homographie qui envoie la droite réelle achevée sur le cercle unité.

I.6.41 Ce sont des arcs de cercles de frontière sur le cercle unité et orthogonaux à ce dernier.

I.6.42 (ii) Se ramener au cas de $(\mathrm{i}, \lambda\mathrm{i})$ avec $\lambda > 0$.

I.6.43 Ce sont les demi-cercles de centre réel passant par M_0 et ne rencontrant pas Δ ; et éventuellement une demi-droite verticale.

I.7 : Polynômes à plusieurs indéterminées

I.7.1 Il l'est si, et seulement si, $n \leqslant 2$.

I.7.2 (i) On peut remplacer $X_i^{q(k-1)+r}$ par X_i^r si $q \geqslant 1$.
(ii) $x^E x^F = x^{E \cup F}$.

I.7.3 (i) Si e, f sont idempotents, alors $1-e$ et ef le sont. Si $E \neq F$, et si $i \in E \setminus F$ ou symétriquement, alors $x_i(1-x_i)$ est facteur de $u_E u_F$. On trouve que $\sum u_E = 1$ en développant $\prod(x_i + (1-x_i))$. On a $u_F = \sum\limits_{E \supset F} (-1)^{\text{card } E - \text{card } F} x^E$ et $\sum x_i = \sum(\text{card } E) u_E$, d'où $(\sum x_i)^k = \sum(\text{card } E)^k u_E$.

I.7.4 Utiliser $\deg_{X_{n,n}} \text{Det}_n$ et discuter plus généralement la réductibilité de $aX + b \in A[X]$ lorsque A est intègre.

I.7.5 C'est une fraction rationnelle de dénominateur $\prod\limits_{1 \leqslant i,j \leqslant n} (X_i + Y_j)$ et dont le numérateur s'annule lorsque $X_i = X_{i'}$ avec $i \neq i'$ ou $Y_j = Y_{j'}$ avec $j \neq j'$.

I.7.6 Pour $n = 2$, la seule nouvelle relation est $f + f'' = 0$.

I.7.7 Si $P = \sum P_i$ (composantes homogènes), alors $P(f, f') = \sum P_i(1, 2x) e^{ix^2}$ ne peut s'annuler identiquement que si tous les $P_i(1, 2x)$ s'annulent identiquement.

I.7.8 Pour toute relation P, considérer la division euclidienne de P par le polynôme P_0 *non unitaire* en X_2.

I.7.9 C'est vrai si le corps K est infini, mais pas autrement.

I.7.10 On a $\omega(PQ) = \omega(P) + \omega(Q)$ et $\deg(PQ) = \deg P + \deg Q$ avec $\omega(P) \leqslant \deg P$, $\omega(Q) = \deg Q$. L'hypothèse dit que $\omega(PQ) = \deg PQ$.

I.7.11 C'est un sous-corps qui contient les X_i/X_j (immédiat) ; réciproquement, si P et Q sont homogènes de même degré,

$$\frac{P(X_1, \ldots, X_n)}{Q(X_1, \ldots, X_n)} = \frac{P(1, X_2/X_1, \ldots, X_n/X_1)}{Q(1, X_2/X_1, \ldots, X_n/X_1)}.$$

I.7.12 Définir un automorphisme de l'espace vectoriel S_1 tel que $L_i \mapsto X_i$, et l'étendre en un automorphisme de la K-algèbre $K[X_1, \ldots, X_n]$.

I.7.13 (i) Pour tout polynôme F en r indéterminées de valuation k, les polynômes $F(P_1 + Q_1, \ldots, P_r + Q_r) - F(P_1, \ldots, P_r)$ et $F(P_1, \ldots, P_r) - \text{In}(F)(P_1, \ldots, P_r)$ sont de valuations strictement plus grandes que k.

I.7.14 Calculer $\det(TI_n - \lambda M)$.

I.7.15 Il est nul.

I.7.16 Pas nécessairement.

I.7.17 On prouve facilement $(ii) \Rightarrow (i) \Rightarrow (iii) \Rightarrow (ii)$.

I.7.18 Utiliser la formule de Leibniz et une récurrence sur $|\alpha|$.

I.7.19 (i) La dérivée partielle de σ_k par rapport à X_i est un polynôme symétrique élémentaire en les autres X_j.
(ii) On trouve la même formule qu'en calcul différentiel ordinaire.
iii) On trouve
$\sum X_i \dfrac{\partial P}{\partial X_i}(TX_1, \ldots, TX_n)$.

I.7.20 De l'exercice précédent (ii), on déduit les mêmes formules qu'en calcul différentiel.

I.7.21 Comme en calcul différentiel, on invoque la formule de composition des matrices jacobiennes.

I.7.22 Sélectionner les termes de degré k dans la formule de Taylor.

I.7.23 On a :
$P(X_1 + Y_1, \ldots, X_n + Y_n) \in R[X_1, \ldots, X_n]$
où $R := K[Y_1, \ldots, Y_n]$;
les $\Delta_\alpha P(Y_1, \ldots, Y_n)$ sont les coefficients.
Les identités s'obtiennent en considérant le produit de $P(X_1 + Y_1, \ldots, X_n + Y_n)$ par $Q(X_1 + Y_1, \ldots, X_n + Y_n)$
et $P(X_1 + Y_1 + Z_1, \ldots, X_n + Y_n + Z_n)$.

I.7.24 Partant d'un multiindice quelconque, celui indiqué en est la version triée. La dimension recherchée est le nombre de solutions de $\sum \alpha_i = k$
avec $\alpha_1 \geqslant \alpha_2 \geqslant \cdots \geqslant \alpha_n \geqslant 0$.

I.7.25 Écrire la matrice de p dans une base adaptée à la somme directe $\operatorname{Im} p \oplus \operatorname{Ker} p$.

I.7.26 On a
$\prod_{i \neq j} (X_{\sigma(i)} - X_{\sigma(j)}) = \varepsilon(\sigma) \prod_{i \neq j} (X_i - X_j)$,
où $\varepsilon(\sigma) = \pm 1$. On vérifie que $\varepsilon(\tau) = -1$ pour une transposition et que $\varepsilon(\sigma_1\sigma_2) = \varepsilon(\sigma_1)\varepsilon(\sigma_2)$.

I.7.27 (i) Si P/Q (forme réduite) est symétrique on voit que $\Phi_\sigma(P) = \lambda_\sigma P$ et $\Phi_\sigma(Q) = \lambda_\sigma Q$, où $\sigma \mapsto \lambda_\sigma$ est un morphisme de groupes. S'il est trivial on a fini, sinon on écrit $P/Q = PQ/Q^2$.

I.7.28 On compte les $\sigma_1^{k_1} \cdots \sigma_n^{k_n}$ de degré k, donc les solutions de $k_1 + \cdots + nk_n = k$.

I.7.29 On a $\Delta_2^2 = \sigma_1^2 - 4\sigma_2$. Pour calculer Δ_3^2, on remarque qu'en remplaçant chaque X_i par $X_i - \sigma_1/3$, on ne change pas le discriminant mais qu'on se ramène à un polynôme $X^3 + pX + q$ pour lequel le résultat est essentiellement $4p^3 + 27q^2$.

I.7.30 Si $\alpha_i < \alpha_j$ avec $i < j$ « minimaux », alors la transposition correspondante fait apparaître un terme d'exposant plus grand, contredisant la dominance.

I.7.31 Utiliser les formules de Newton (section 1.4 de la page 329).

I.7.32 (i) Poser $K' := K[X]/<Q>$, où Q est un facteur irréductible de P.

I.7.33 (ii) Développer en série entière la décomposition en éléments simples de cette fraction rationnelle.

I.7.34 Pour (i) comme pour (ii), la méthode est la même que dans le texte.

I.7.35 Les calculs du texte s'appliquent tels quels car ils produisent des polynômes unitaires à coefficients entiers. Pour la dernière question, remarquer que tout nombre algébrique s'écrit x/d où x est un entier algébrique et $d \in \mathbb{N}^*$.

I.7.36 Si $P = \prod_{i=1}^{n}(X - x_i)$, démontrer que chaque $P_k := \prod_{i=1}^{n}(X - x_i^k)$ $(k \in \mathbb{N})$ est un polynôme à coefficients entiers, dont les coefficients sont uniformément majorés ; il n'y a donc qu'un nombre fini de P_k.

I.7.37 (i) Pour démontrer la surjectivité, on pourra comparer les dimensions de ces deux espaces vectoriels. (ii) Pour tout $x \in K^*$, qui est un groupe à $(q-1)$ éléments, $x^{q-1} = 1$.

I.7.38 (i) Affiner la méthode utilisée pour prouver le lemme 17 de la page 336.

I.7.39 Appliquer le théorème 18 à P et $\Phi_\sigma(P)$.

I.7.40 Tout polynôme nul sur une partie dense est nul (continuité).

I.7.41 (i) D'après le module I.5, il contient le complémentaire du lieu d'annulation de Sep ; il faudra justifier la non-nullité de ce dernier.
(ii) Si D est la diagonale des valeurs propres λ_i, $\chi_A(D)$ est la diagonale des $\chi_A(\lambda_i)$ et, pour toute matrice inversible P, $\chi_A(PDP^{-1}) = P\chi_A(D)P^{-1}$.
(iii) Les n^2 coefficients de $\chi_A(A)$ sont des polynômes en les coefficients de A.

I.7.42 (i) Le groupe S_n est représenté dans le groupe G par les matrices de permutation et l'action ci-dessus étend celle de G.
(ii) Découle de l'égalité $\chi_{PMP^{-1}} = \chi_M$.
(iii) L'algèbre des polynômes invariants par G est librement engendrée par les c_i.

I.7.43 Dans le développement de $\det S_{p,1}(A, B)$ par rapport à sa première colonne, chaque coefficient a_i est multiplié par $(-1)^i$ et par un déterminant mineur suivant, qui vaut $\lambda^{p-i}\mu^i$.

I.7.44 La matrice du système d'équations est à peu de choses près la matrice de Sylvester.

I.7.45 On a $\mathrm{Res}_{p,q}\big(\Phi(A), \Phi(B)\big) = \varphi\big(\mathrm{Res}_{p,q}(A, B)\big)$ mais l'égalité correspondante pour les résultants n'en découle que si les coefficients dominants de A et de B n'appartiennent pas au noyau de φ.

I.7.46 Calculer $\mathrm{Res}_{p,q}(\lambda A, \mu B)$.

I.7.47 Ils sont égaux à un facteur près.

I.7.48 Utiliser la formule donnant le discriminant en fonction des racines et le théorème 24.

I.7.49 Il s'agit d'éliminer x, d'où une courbe dans le plan des y, z ; et ainsi de suite.

I.7.50 Il s'agit d'éliminer u et v. On doit trouver $x^2 - y^2 z = 0$.

I.7.51 Utiliser le théorème 24 de la page 341.

I.7.52 Si $P(x) = Q(y) = 0$ et $x \neq 0$ et si $z = xy$, éliminer x entre les équations $P(x) = 0$ et $x^{\deg Q} Q(z/x) = 0$.

I.7.53 On trouve $M(t,h)$ en résolvant un système de déterminant :

$$a(t)b(t+h) - a(t+h)b(t) = h(a(t)b'(t) - a'(t)b(t)) + \mathrm{o}(h).$$

Calculer alors un développement limité à l'ordre 1 en h des coordonnées de $M(t,h)$.

I.7.54 On est ramené à éliminer t entre les équations $Yt^4 - 2(X+1)t^3 - 2(X-1)t - Y = 0$ et $4Yt^3 - 6(X+1)t^2 - 2(X-1) = 0$.

I.7.55 Une équation du rayon réfléchi est $\dfrac{y - \sin\theta}{x - \cos\theta} = \tan 2\theta$, que l'on rend rationnelle en le paramètre en posant $t := \tan(\theta/2)$.

I.8 : Structures discrètes et récursivité

I.8.1 Le mot vide est dans L_1 et $x \in L_1 \Rightarrow axb \in L_1$. Le mot vide est dans D et $x, y \in D \Rightarrow axby \in D$.

I.8.2 Penser à aba.

I.8.3 La relation n'est pas transitive : $abcd \to bacd \to badc$ mais $abcd$ ne se recombine pas directement en $badc$. L'algorithme épluche w et w' par la droite et par la gauche tant qu'ils ont les mêmes caractères ; puis teste si les mots restant sont miroir l'un de l'autre.

I.8.4 Prendre $\varphi : X \to \mathbb{N}$ la fonction constante 1.

I.8.5 Remarquer que, si u est facteur de v et $|u| = |v|$, alors $u = v$.

I.8.6 Dans chaque cas, appliquer le lemme 4 de la page 357.

I.8.7 (i) Invoquer le lemme 4 de la page 357. (ii) Le stabilisateur de w sous l'action de $\mathbb{Z}/n\mathbb{Z}$ est $p\mathbb{Z}/n\mathbb{Z}$.

I.8.8 Vérifier que L^* est un sous-monoïde et que tout sous-monoïde contenant L contient L^*.

I.8.9 Pour chacun de ces langages, vérifier qu'il est égal à $\{a,b\}^*$.

I.8.10 Notons $X := \{0,\dots,9\}$ et $Y := \{1,\dots,9\}$. Alors le langage des nombres sans virgule (entiers) en écriture propre (pas de 0 en tête) est $A_1 := 0 + YX^*$. Celui des nombres avec virgule en écriture propre (pas de 0 à droite) est $A_2 := A_1.v.X^*Y$, où v désigne la virgule. Le langage recherché est $A_1 + A_2$.

I.8.11 $(\varepsilon + L)^+ = L^*$.

I.8.12 Utiliser l'exercice précédent.

I.8.13 Notant $\Phi(X) := L_1X + L_2$, l'application Φ est croissante (pour l'inclusion) et vérifie $\Phi(\bigcup X_i) = \bigcup \Phi(X_i)$, ce qui suffit à montrer la formule.

I.8.14 Il faudra définir le produit de deux matrices à l'aide des opérations sur les langages.

I.8.15 L'un est égal à L, l'autre y est inclus, peut-être strictement.

I.8.16 La plus difficile est celle qui porte sur $x^{-1}(L_1L_2)$. Écrire $xy = w_1w_2$ avec $w_i \in L_i$, et raisonn er selon la première lettre de w_1 (ou sa vacuité).

I.8.17 Oui (quatre résiduels). Non.

I.8.18 Si $u \in P$ n'est pas le mot vide, il a au moins deux lettres et l'on discute selon que $u = aav$, $u = abv$, $u = bav$ ou $u = bbv$. Pour la rationalité, appliquer la proposition 7 de la page 359.

I.8.19 Soit $x \in L_0$, codant l'entier n multiple de 3. Si $x = y0$, resp. $x = y1$, alors y code p tel que $2p = n$, resp. $2p + 1 = n$. Par exemple dans le deuxième cas, $2p + 1 \equiv 0 \pmod 3$ équivaut à $p \equiv 1 \pmod 3$, donc à $y \in L_1$.

I.8.20 On sait calculer les résiduels et déterminer la propreté inductivement.

I.8.21 Sens direct : c'est vrai pour $w := \varepsilon$; si c'est vrai pour w_1 et pour w_2, c'est vrai pour aw_1bw_2. Réciproque : si $w \neq \varepsilon$ vérifie ces conditions, $w' := aw_1b$ est le premier préfixe non vide de w tel que $|w'|_a = |w'|_b$.

I.8.22 Le langage $L' = \{a^nba^n \mid n \in \mathbb{N}\}$ n'est pas rationnel, donc L non plus.

I.8.23 Si $n := \operatorname{card} S$, respectivement $2^{n(n-1)/2}$ et 2^{n^2}.

I.8.24 Pour aller de liste vers matrice : pour tout i, mettre à 0 tous les $m_{i,j}$ puis, parcourant la liste $T_G[i]$, mettre $m_{i,j}$ à 1 pour tout j rencontré. Réciproquement : pour tout i, créer la liste vide $T_G[i]$ puis, pour tout j, si $m_{i,j} = 1$ ajouter j à cette liste.

I.8.25 (iii) En remplaçant les successeurs par des voisins, on obtient la réunion des composantes connexes des éléments de E.

I.8.26 Après k exécutions, $X[i, j]$ dit s'il existe un chemin de i à j dont tous les sommets intermédiaires ont un numéro $\leqslant k$.

I.8.27 L'existence de μ équivaut à la présence du monôme $\prod X_{i,\mu(i)}$ dans le déterminant.

I.8.28 Si l est la longueur maximum d'un chemin, un majorant est $2l - 1$.

I.8.29 Le coefficient d'indices (i, j) de M^k est ici le nombre de chemins de longueur k reliant i à j.

I.8.30 S'il existe un chemin de s à t, il en existe un de lonngueur majorée par $\operatorname{card} A$ et par $\operatorname{card} S - 1$.

I.8.31 Si $n := \operatorname{card} X$ et $m := \operatorname{card} Q$, respectivement 2^{2m+m^2n} et $m2^mm^{mn}$.

I.8.32 Le complémentaire de $L(\mathcal{A})$.

I.8.33 La réunion de $L(\mathcal{A}_1)$ et de $L(\mathcal{A}_2)$.

I.8.34 L'intersection de $L(\mathcal{A}_1)$ et de $L(\mathcal{A}_2)$.

I.8.35 Le langage $\{abba\}$.

I.8.36 Prendre comme modèle celui de l'exercice précédent.

I.8.37 Le premier automate a trois états, correspondant aux résiduels a^*b^*, b^* et $\varnothing$.

I.8.38 (i) Penser au montage en parallèle (pour la somme) et au montage en série (pour le produit).

I.8.39 Aucun élément de L n'admet de facteur itérant.

I.8.40 La réciproque est fausse.

I.8.41 Tout cycle $e_{i_1} \to \cdots \to e_{i_k} \to e_{i_1}$ donne lieu à une somme nul d'éléments de la famille, donc à une relation linéaire. Tout chemin $i_0 \to i_1 \to \cdots \to i_k$ permet d'écrire $e_{i_k} - e_{i_0}$ comme somme d'éléments de la famille ; or les $e_i - e_j$ (i, j quelconques) engendrent H. Attention, les réciproques sont nettement plus difficiles !

I.8.42 L'ensemble ordonné associé à un arbre est l'ensemble de ces sommets avec la relation de descendance.

I.8.43 On trouvera c_{n-1} arbres généraux de taille n et c_n arbres binaires de taille n, où $c_0 = c_1 = 1$, $c_2 = 2$ et $c_3 = 5$ (nombres de Catalan, voir l'exercice I.8.45 ci-dessous).

I.8.44 On a $\eta(t) = |t|$ pour les arbres *filiformes* (un seul nœud à chaque niveau) et $|t| = 2^{\eta(t)} - 1$ pour les arbres parfaitement équilibrés (tous les niveaux sont remplis). Il y a 2^{n-1} arbres filiformes de taille n (pour $n \geqslant 1$).

I.8.45 (i) $a_n = b_{n-1}$ (enlever la racine) ; b_n est la somme sur k des nombres de forêts de k arbres, ceux-ci ayant des tailles $i_1, \ldots, i_k$ telles que $i_1 + \cdots + i_k = n - 1$.
(ii) Pour fabriquer un arbre binaire de taille n, il faut choisir les tailles i, j de ses sous-arbres telles que $i + j = n - 1$ puis un arbre parmi c_i et un autreparmi c_j.
(iii) permet de compter les expressions parenthésées.
(iv) Si le mot w_1, resp w_2 correspond à l'arbre t_1, resp. t_2, alors aw_1bw_2 correspond à $N(t_1, t_2)$.

I.8.46 C'est le calcul dichotomique de a^n (module d'algorithmique de [L1]).

I.8.47 Le nombre C_n de ces appels vérifie une relation de récurrence.

I.8.48 Il faut mémoriser des listes.

I.8.49 $C(n)$ s'obtient à partir de $C(n-1)$ par $(a, b) \mapsto (b, a)$.

I.8.50 Considérer un polynôme comme un tableau de coefficients muni d'un indice de début et d'un indice de fin.

I.8.51 Il y a $2^n - 1$ déplacements.

I.8.52 On en déduit une contradiction : il n'existe pas de telle fonction **`termine`**.

I.8.53 L'ordre lexicographique sur $\mathbb{N} \times \mathbb{N}$ est un bon ordre.

I.8.54 Contrairement à la fonction d'Ackermann-Peter, celle de Morris peut boucler.

I.8.55 Similaire au calcul de la taille, en remplaçant $x+y$ par $1+\max(x,y)$.

I.8.56 Appelons A et B ces fonctions, et supposons que $B(A(f_1))=f_1$ et $B(A(f_2))=f_2$. Il faut alors calculer $B(A(R(f_1)::f_2))$ en appliquant les règles inductives.

I.8.57 La traduction est mécanique.

I.8.58 Règles inductives : $\chi(F)=0$ et $\chi(N(t_1,t_2))=\max(\chi(t_1),1+\chi(t_2))$. Les mauvais arbres sont déséquilibrés à droite.

II.1 : Espaces vectoriels normés

II.1.7 La réponse est non.

II.1.11 On pourra considérer les $u^k\circ v-v\circ u^k$ pour $k\in\mathbb{N}^*$.

II.1.13 On trouve :

$$|||A|||_1=\max\bigl(\|C_1\|_1,\dots,\|C_n\|_1\bigr),$$

où $C_1,\dots,C_n$ désignent les colonnes de A.

II.1.14 Attention, l'hypothèse $|||u|||\leqslant 1$ est nécessaire, comme le lecteur le montrera.

II.1.16 Réponse 1 : f doit être injective (*i.e.* strictement monotone).
Réponse 2 : f doit être bijective.

II.1.17 On pourra considérer la suite de terme général $1+X+\cdots+X^n$.

II.1.18 Utiliser le théorème du point fixe.

II.1.19 On commencera par prouver que l'image de f est fermée, puis on appliquera le théorème d'inversion locale.

II.1.25 Minimiser $d\bigl(x,f(x)\bigr)$ sur X. Attention : d'une part le théorème du point fixe ne s'applique pas, et d'autre part le résultat est en général faux si X n'est pas compact.

II.1.28 On pourra fixer M_2,M_3 et faire varier M_1 sur $\mathcal{C}_1$, en paramétrant ce cercle.

II.1.32 Pour 1, considérer AXX^*, où X est un vecteur propre de A. Pour 2, se ramener au cas où A est triangulaire, et considérer $D^{-1}AD$, où $D=\mathrm{Diag}(1,t,\dots,t^{n-1})$, le réel $t>0$ étant bien choisi. Pour 4, on pourra utiliser 1 et 2.

II.1.35 On pourra montrer d'abord que l'ensemble D des nombres dyadiques, c'est-à-dire de la forme $a/2^n$, où $a\in\mathbb{Z}$ et $n\in\mathbb{N}$, est dense dans $\mathbb{R}$.

II.1.39 Dans $\mathbb{R}^2$, par exemple, il y a suffisamment d'horizontales et de verticales pour « éviter » les points de S.

II.1.41 Pour 2, on ne peut pas aller continûment de l'origine à un point de C.

II.1.43 Pour 2, on pourra utiliser la preuve de la proposition 53 de la page 436.

II.1.44 On pourra vérifier d'abord que $E=V^\perp\oplus V^{\perp\perp}$ et $E=V^\perp\oplus\overline{V}$.

II.2 : Suites et séries de fonctions

II.2.9 Montrer que l'adhérence de $\mathcal{A}$ dans $\mathcal{C}(J,\mathbb{R})$ est un sous-anneau de $\mathcal{C}(J,\mathbb{R})$ contenant $\mathbb{Q}$.

II.2.10 Considérer sur E d'une part la norme de la convergence uniforme, d'autre part une norme « adaptée » à la base obtenue en 1.

II.2.15 On trouve 0 si $n<0$ et $2\pi/n!$ si $n\geqslant 0$.

II.2.16 La somme de la première série est 1. Pour les deux séries, penser à Fubini.

II.2.24 Il y a convergence si, et seulement si, $a<1$.

II.2.29 Penser à la dérivation.

II.2.33 Le rayon de convergence est 1.

II.2.34 Le rayon de convergence est 1.

II.2.40 Considérer les séries entières $\sum_{n\geqslant 0} z^{8n+k}/(8n+k)$, pour $k=1,4,5,6$.

II.2.41 Utiliser une interversion des signes $\sum$ et $\int$.

II.3 : Intégration

II.3.1 La fonction f est constante égale à 1.

II.3.2 $L_1=2\sqrt{2}-2$, $\quad L_2=1/3$, $L_3=\pi/4$, $\quad L_4=(\ln 2)/2$, $L_5=1/2$.

II.3.4 On pourra faire une intégration par parties.

II.3.5 On déduit $\overset{*}{\int_0^1} 1_U=1$ de la densité de U dans $[0,1]$. On montre que $\overset{*}{\int_0^1} 1_U\leqslant \log U\leqslant \dfrac{1}{2}\cdot$

II.3.6 On a $\log U=\frac{1}{2}$ et $\log K=\frac{1}{2}$.

II.3.7 On utilise des sous-ensembles élémentaires fermés $K_n\subset L_n$ tels que $\log K_n\geqslant \log A_n-\varepsilon 2^{-n}$.
Si la suite K_n n'est pas décroissante on utilise les $L_n:=\mathbb{K}_1\cap\ldots\cap K_n$ en montrant que
$\log L_n\geqslant \log A_n-\varepsilon/2-\ldots-\varepsilon/2^n$.

II.3.8 i) Soit $x>0$,
poser $I(x):=\displaystyle\int_a^b \theta(xt)\,\mathrm{d}t$, et montrer que :

$$xI(x)=n_x\int_0^T \theta(u)du+\int_{xa+n_xT}^{xb}\theta(u)du,$$

avec $n_x:=E(x\frac{b-a}{T})$.
ii) Vérifier le résultat lorsque f est en escalier sur $[a,b]$. Si f est intégrable sur $[a,b]$, utiliser la proposition définition 56 page 43.
iii) Appliquer les résultats de (ii) aux fonctions : $f:t\in[0,\pi]\to\frac{t(t-2\pi)}{\sin(\frac{t}{2})}$
et $\theta:u\in\mathbb{R}\to\sin u$.

II.3.9 Montrer que pour toute subdivision $\mathbf{u}$ de $[a,b]$, $V(f,\mathbf{u}) \leqslant \displaystyle\int_a^b |f'(t)|\,\mathrm{d}t$ et que pour tout $\varepsilon > 0$, il existe une subdivision pointée de $[a,b]$ telle que $V(f,\mathbf{u}) \geqslant \displaystyle\int_a^b |f'(t)|\,\mathrm{d}t - \varepsilon$.

II.3.10 Utiliser : $\big\|f(x) - f(a)\big\| \leqslant V(f)$, pour tout $x \in [a,b]$,

II.3.11 Soit $\{x_k\}_{k\in[\![1,n]\!]}$ l'ensemble des points de discontinuité de f. Montrer que $V(f) = \displaystyle\int_a^b |f'(t)|\,\mathrm{d}t + \sum_{k=1}^{n}\big(\sigma_k^+ + \sigma_k^-\big)$ où $\sigma_k^+ := |f(x_k) - f(x_k^+)|$ et $\sigma_k^- = |f(x_k) - f(x_k^-)|$.

II.3.12 (i) Par un changement de variable : on se ramène au cas $a = 0, b = 1$. On écrit :

$$R_n(f) = \frac{1}{n}\sum_{k=0}^{n-1}\int_0^1 \Big(f\big(\frac{x+k}{n}\big) - f\big(\frac{k}{n}\big)\Big)\,\mathrm{d}x$$

(ii) (resp. (iii)) Appliquer la formule de Taylor avec reste intégral à l'ordre 1 (resp. 2) à $f\big(\frac{x+k}{n}\big) - f\big(\frac{k}{n}\big)$ et utiliser la formule de la moyenne.

II.3.13 On utilisera la proposition 62 et la proposition-définition 27. Pour tout $\varepsilon > 0$, il existe une application en escalier $\varphi_\varepsilon : I \mapsto \mathbb{C}$ et une fonction en escalier $\psi_\varepsilon : I \mapsto \mathbb{R}$ telles que :

$$|f - \varphi_\varepsilon| \leqslant \psi_\varepsilon \text{ et } \int_a^b \psi_\varepsilon \leqslant \varepsilon.$$

II.3.14 2.a) On montrera que $\|f\|$ n'est pas $\mathscr{R}$-intégrable.

2.b) On a $\displaystyle\int_0^1 u_n \circ f = 0$.

2.c) On montrera que les sommes de Riemann « convergent vers 0 ».

II.3.15 Considérer les suites $(\frac{1}{\pi/2+2n\pi})_{n\in\mathbb{N}}$ et $(\frac{1}{3\pi/2+2n\pi})_{n\in\mathbb{N}}$.

II.3.16 La limite existe et vaut $f(0)$.

II.3.17 1 Non convergente en $+\infty$.
2 Non convergente en 0.
3 Convergente. 4 Convergente.

II.3.18 Le changement de variable :

$$t = \frac{b+a}{2} + u\frac{b-a}{2}$$

permet de montrer que l'intégrale vaut π.

II.3.19 Un changement de variable $u = \sqrt{1-t}$ permet de montrer que l'intégrale vaut $-4 + 4\ln 2$.

II.3.20 On pourra utiliser un changement de variable. L'intégrale est convergente si, et seulement si, l'une des deux conditions suivantes est vérifiée :
- soit $\lambda < 1$,
- soit $\lambda = 1$ et $\mu > 1$.

II.3.21 Les intégrales sont absolument convergentes pour $\alpha > 1$, semi-convergentes pour $0 < \alpha \leqslant 1$ et divergentes pour $\alpha \leqslant 0$.

II.3.22 On pourra intégrer par parties en posant $u = \sin\Big(\frac{t^3}{3} + xt\Big)$ et $v = \frac{1}{t^2+x}$.

II.3.23 1. a) Considérer $x \mapsto x^{-\varepsilon}/g(x)$.

1.c) Calculer $\displaystyle\int_a^x g(t)\,\mathrm{d}t$ par intégration par parties.

2. Poser $g : x \mapsto 1/\ln x$.

II.3.24 1. On modifie $x \mapsto \frac{1}{x^2}$ aux points de $\mathbb{N}^*$. 2. On choisit convenablement f affine par morceaux et continue telle que $f(n) = 1$ pour tout $n \geqslant 2$.

II.3.25 Calculer $I_n := \displaystyle\int_1^{n+1} \frac{x - E(x) - \frac{1}{2}}{x}$ et montrer que $\displaystyle\int_1^{+\infty} \frac{x - E(x) - \frac{1}{2}}{x} = \lim_{n\to+\infty} I_n$.

II.3.26 i) Étudier la série $\sum\Big(f(a+k) - \displaystyle\int_{a+k}^{a+k+1} f(t)\,\mathrm{d}t\Big)$ en remarquant :

$$f(a+k) - \int_{a+k}^{a+k+1} f(t)\,\mathrm{d}t = \int_{a+k}^{a+k+1} (t - (a+k+1))f'(t)\,\mathrm{d}t.$$

ii) Soient $x > a+1$ et $n := E(x-a-1)$, montrer que

$$\left|\int_{a+n+1}^{x} f(t)\,\mathrm{d}t\right| \leqslant |f(a+n+1)| + \int_{a+n+1}^{x} |f'(t)|\,\mathrm{d}t.$$

iii) Pour tout $\alpha \leqslant 1$ la série diverge.

II.3.27 (iii) Faire le changement de variable $u := at$ dans l'intégrale $\displaystyle\int_0^{+\infty} \frac{e^{-at} - e^{-bt}}{t}\,\mathrm{d}t$ pour se ramener à $\displaystyle\int_0^{+\infty} \frac{e^{-u} - e^{-\frac{b}{a}u}}{u} du$.

II.3.28 Développer la fonction $\frac{\ln(1-t)}{t}$ en série entière et utiliser la proposition 124 de la page 638. On en déduit $I = \zeta(3)$.

II.3.29 (i) Montrer (en faisant le changement de variable $t = xu$) :

$$f(x) = \int_0^{\frac{1}{x}} (1 + u^\alpha)^{-\frac{1}{\alpha}} du.$$

(ii) Montrer :

$$\begin{aligned} f(x) + \ln x &= \int_0^x \frac{1}{v}\Big((1+v^\alpha)^{-\frac{1}{\alpha}} - 1\Big)dv \\ &= \sum_{k=1}^{n} \frac{1}{\alpha k}\, a_k x^{\alpha k} + \mathrm{o}(x^{(n+1)\alpha-1}). \end{aligned}$$

Si $\alpha = 3$, on a :

$f(x) + \ln x = C + \frac{x^3}{9} - \frac{x^6}{27} + \frac{14\,x^9}{9^3} + \mathrm{o}(x^{11})$ avec :

$$C := \int_0^1 (1+t^3)^{-\frac{1}{3}}\,\mathrm{d}t + \int_0^1 \frac{1}{t}\Big((1+t^3)^{-\frac{1}{3}} - 1\Big)\,\mathrm{d}t$$

II.3.30 (i) Pour tous $x \in \mathbb{R}$ et $t \in [0, +\infty[$ on a :

$$\left|\frac{e^{-(t^2+i)x^2}}{t^2+i}\right| \leqslant \frac{1}{\sqrt{1+t^4}}.$$

(ii) Pour tous $x \in \mathbb{R}$ et $t \in [0, +\infty[$, on a :

$$\frac{\partial\Big(\frac{e^{-(t^2+i)x^2}}{t^2+i}\Big)}{\partial x} = -2xe^{-(t^2+i)x^2}$$

$$F'(x) = -2x\int_0^{+\infty} e^{-(t^2+i)x^2}\,\mathrm{d}t.$$

Si $x \neq 0$,

$$\begin{aligned} F'(x) &= -2e^{-ix^2}\int_0^{+\infty} e^{-u^2}\,\mathrm{d}u \\ &= -\sqrt{\pi}e^{-ix^2}. \end{aligned}$$

(iii) Pour tout $x > 0$,

$$F(x) - F(0) = -\int_0^x \sqrt{\pi}e^{-iu^2}\,\mathrm{d}u.$$

Prouver $\lim_{x\to+\infty} F(x) = 0$ et :

$$F(0) = \int_0^{+\infty} \frac{1}{u^2+i}\,\mathrm{d}u$$
$$= \sqrt{\pi}\int_0^{+\infty} e^{-iu^2}\,\mathrm{d}u.$$

On a :

$$\int_0^{+\infty} \frac{1}{u^2+i}\,\mathrm{d}u = (1-i)\int_0^{+\infty} \frac{u^2}{1+u^4}\,\mathrm{d}u$$
$$= \pi\sqrt{2}/2.$$

II.3.31 Question 2. Raisonner par l'absurde en supposant que, pour tout $n \in N^*$, $\alpha_n > \delta > 0$. Choisir des sous-ensembles élémentaires $X_n \subset A_n$ tel que $\mathrm{long}\, X_n > \alpha_n - \delta/2^n$ et poser $H_n := \bigcap_{i\in[\![1,n]\!]} X_i$. Montrer que chaque H_n n'est pas vide. On vérifiera pour cela que si $X \subset A_n \setminus X_n$ est un sous-ensemble élémentaire, on a $\mathrm{long}(X) < \delta/2^n$, puis que si $Y \subset A_n \setminus H_n$ est un sous-ensemble élémentaire, $\mathrm{long}(Y) < \delta$.

II.3.32 Soit $\varepsilon > 0$. On définit des sous-ensembles $\{A_n\}_{n\in\mathbb{N}}$ de $[a,b]$ par $A_n :=$

$$\left\{ x \in [a,b] \;\middle|\; \exists k \geqslant n, f_k(x) \geqslant \frac{\varepsilon}{2(b-a)} \right\}$$

On prouve ensuite, en utilisant le résultat de la question 2 de de l'exercice II.3.31 pour la famille de parties $\{A_n\}_{n\in\mathbb{N}}$, que, si n est assez grand, pour toute fonction en escalier ψ telle $0 \leqslant \psi \leqslant f_n$, on a $\int_a^b \psi \leqslant \varepsilon$.

II.3.33 1. $\lim_{x\to+\infty} F(x) = 0$.

2. $F'(x) = -\dfrac{e^{-x}}{\sqrt{x}}\displaystyle\int_0^{+\infty} e^{-u^2}\,\mathrm{d}u$.

3. Faire le changement de variable $x = u^2$.

II.3.34 La fonction f est continue sur $[0,+\infty[$ et dérivable sur $]0,+\infty[$.

II.3.35 La fonction f est définie et de classe $\mathcal{C}^\infty$ sur $]0,1[$.

II.3.36 On écrit $U := \bigcup_{i\in\mathbb{N}} I_i$ où les I_i sont des intervalles *ouverts* de $[a,b]$ et où la réunion est *disjointe*. On pose $U_n := \bigcup_{i\in[\![0,n]\!]} I_i$ et l'on considère la suite $(1_{U_n})_{n\in\mathbb{N}}$.

II.4 : Séries de Fourier

II.4.1 1. $c_n = (-1)^n \dfrac{\sinh(\pi)}{\pi(1+n^2)}$, $a_n = 2c_n$, $b_n = 0$.

2. Pour $n \neq 0$, $c_n = (-1)^n \dfrac{2}{n^2}$, $a_n = 2c_n$, $b_n = 0$.

3. Si n est pair $c_n = a_n = b_n = 0$. Si n est impair, $c_n = \dfrac{-4i}{\pi^3 n^3}$, $a_n = 0$, $b_n = 2ic_n$.

4. $c_n = \dfrac{2}{\pi(1-4n^2)}$, $a_n = 2c_n$, $b_n = 0$.

5. Si n est impair, $c_n = a_n = b_n = 0$. Si $n = 2k$ est pair, $c_n = \dfrac{(-1)^k}{\pi(1-n^2)}$, $b_n = 2c_n$, $a_n = 0$.

II.4.2 1. Définir les coefficients de Fourier par une intégrale sur $[0,T]$ et faire un changement de variable de la forme $t = u + T/2$.

2. Définir les coefficients de Fourier par une intégrale sur $[-T/4, 3T/4]$ et faire un changement de variable de la forme $t = T/2 - u$.

II.4.3 Linéariser $\sin^3$ en utilisant les formules de Moivre.

II.4.4 La série est normalement convergente, de même que la série des dérivées. La fonction f est donc de classe $\mathcal{C}^1$. C'est sa propre série de Fourier.

II.4.5
1. La fonction f est de classe $\mathcal{C}^\infty$ et ses coefficients de Fourier sont donc à décroissance rapide.
2. Les racines de l'équation caractéristique $rx^2-(1+r^2)x+r$ sont r et $1/r$.
3. $a_n = 2r^n$.
4. La fonction f étant de classe $\mathcal{C}^\infty$, elle est sommes de sa série de Fourier.
5. Il s'agit de la partie réelle d'une série géométrique de raison $r^n \mathrm{e}^{int}$.

II.4.6 Choisir d'abord p suffisamment grand pour que $|u_k-\ell|<\varepsilon/2$ pour $k\geqslant p$ et choisir ensuite n suffisamment grand pour $\dfrac{1}{n}\displaystyle\sum_{k=1}^{p}|u_k-\ell|$ soit inférieur à $\varepsilon/2$.

II.4.7 Utiliser la formule de Parseval puis écrire la série comme somme de la série des termes pairs et de la série des termes impairs.

II.4.8
1. Revenir à la définition.
2. Utiliser la version polarisée de la formule de Parseval puis montrer que $\overline{c_k}(\overline{g}\cdot \mathrm{e}_n) = c_{n-k}(g)$.

II.4.9 Utiliser le théorème de Dirichlet au point $x=1/4$.

II.4.10 Une intégration directe à partir de la définition de $c_n(f)$ permet de déterminer le coefficient. On utilise ensuite la formule de Parseval.

II.4.11
1. La série est normalement convergente.
2. $a_n = (-1)^n \dfrac{2a\sinh(a\pi)}{\pi(n^2+a^2)}$ et $b_n = -(-1)^n \dfrac{2n\sinh(a\pi)}{\pi(n^2+a^2)}$.
3. $af+f' = \dfrac{\pi \mathrm{e}^{at}}{2\sinh(a\pi)} - \dfrac{1}{2a}$.
4. $f(t) = \dfrac{\pi\cosh(at)}{2a\sinh(a\pi)} - \dfrac{1}{2a^2}$.
5. $$\sum_{n=1}^{+\infty}\frac{(-1)^n}{n^2+a^2} = f(0) = \frac{\pi}{2a\sinh(a\pi)} - \frac{1}{2a^2}.$$

II.5 : Fonctions de plusieurs variables

II.5.1
1. Non inversible.
2. Non inversible.
3. Inversible.

II.5.2 Le déterminant de la matrice Jacobienne est égal à $1-\cos(x/2)\cos(y/2)/4$ qui est toujours supérieur à $3/4$. La matrice Jacobienne est donc inversible en tout point de $\mathbb{R}^2$.

II.5.3 L'application admet un inverse local en tout point $(x,y)\in\mathbb{R}^2$ tel que $y\neq 0$.

II.5.4 Appliquer le théorème d'inversion locale à l'application $f(x,y) = \big(x+y+\sin(xy), \sin(x^2+y)\big)$ au point $(0,0)$. Il y a une solution pour a suffisamment proche de 0.

II.5.5 $\dfrac{\partial F}{\partial y}(0,1) = 3 \neq 0$; on peut appliquer le théorème des fonctions implicites. On a alors $\varphi'(0)=1$.

II.5.6 $\dfrac{\partial F}{\partial y}(1/2,1/2) = 1 \neq 0$; on peut appliquer le théorème des fonctions implicites.
On a alors $g'(1/2) = -3$.

II.5.7 Appliquer le théorème des fonctions implicites à l'équation $f(c,x) := P_c^{\circ k}(x) - x = 0$ en observant que $\dfrac{\partial f}{\partial x}(c,x) = (P_c^{\circ k})'(x)$.

II.5.8 Appliquer le théorème d'inversion locale à l'application $G : M_2(\mathbb{R}) \to M_2(\mathbb{R})$ définie par $G(M) := M^2$.

II.5.9 Pour $t \in I$, on note P_t le polynôme caractéristique de la matrice $A(t)$. On pourra appliquer le théorème des fonctions implicites à l'équation $\big(P_t(\lambda_1), \dots, P_t(\lambda_n)\big) = (0, \dots, 0)$.

II.5.10 X, Y_c pour $c \neq 0$ et Z sont des sous-variétés lisses de $\mathbb{R}^2$.

II.5.11 X_c pour $c \neq 0$ et Y sont des sous-variétés lisses de $\mathbb{R}^3$. En revanche, Z n'est pas une sous-variété lisse de $\mathbb{R}^3$: considérer le point $(0,0,0)$.

II.5.12 Le cône n'est pas une sous-variété lisse de $\mathbb{R}^3$: considérer le point $(0,0,0)$.

II.5.13 Si Φ est la forme polaire de Q, alors la différentielle $D_{\mathbf{x}}Q$ est l'application $\mathbf{h} \mapsto 2\Phi(\mathbf{x},\mathbf{h})$. Si $\mathbf{x} \neq \mathbf{0}$, cette application n'est pas identiquement nulle puisque Q est non dégénérée. Le noyau de la différentielle est l'orthogonal de $\mathbf{x}$ pour la forme quadratique Q.

II.5.14 Non, car il y a un point de rebroussement en $t = 0$.

II.5.15 L'espace tangent est l'image de $\mathbb{R}^2$ par l'application linéaire $D_{(r,s)}\varphi$, c'est-à-dire le plan dirigé par les vecteurs :
$$(-\sin r \cos s, -\sin r \sin s, \cos r)$$
et $(-\sin s, \cos s, 0)$.

II.5.16 Si au voisinage de $\mathbf{a}$, X est donné par l'équation $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ et au voisinage de $\mathbf{b}$, Y est donné par l'équation $\mathbf{G}(\mathbf{y}) = \mathbf{0}$, alors au voisinage de $(\mathbf{a},\mathbf{b})$, $X \times Y$ est donné par $\mathbf{H}(\mathbf{x},\mathbf{y}) = 0$ avec $\mathbf{H}(\mathbf{x},\mathbf{y}) := \big(\mathbf{F}(\mathbf{x}), \mathbf{G}(\mathbf{y})\big)$.

II.5.17
1. $D_M F(H) = {}^tM \cdot H + {}^tH \cdot M$.
2. $\operatorname{Ker} D_M F = \big({}^tM\big)^{-1} \cdot E_a$.
3. Appliquer le théorème des fonctions implicites à l'équation $F(M) = \mathrm{I}_n$.

II.5.18 f' est de classe $\mathcal{C}^1$ et satisfait les conditions de Cauchy-Riemann.

II.5.19 Les polynômes de Taylor d'ordre 2 sont :
1. $x + 2y$.
2. $1 - x^2 - y^2$.
3. $1 - x^2 - y^2$.
4. $1 + xy$.
5. $1 + (x-1)^2 - y^2/2$.

II.5.20
1. $(x,y) = (1,2)$ est un point selle et $(x,y) = (3,6)$ est un minimum.
2. $(x,y) = (\pi/2 + k\pi, 1)$ est un point selle.

3. $(x,y,z)=(\pi/4+k\pi/2,\ell\pi,0)$ est un point selle.
   $(x,y,z)=(k\pi/2,\pi/2+\ell\pi,0)$ est un point selle si $\cos(2x)\sin(y)=1$ et un minimum local si $\cos(2x)\sin(y)=-1$.
4. L'ensemble des points critiques est le plan d'équation $x+y+z=0$ qui ne contient que des minima.

II.5.21 1. $(x,y)=(0,0)$ est un point selle.
2. $(x,y)=(0,0)$ est un point critique dégénéré. Ce n'est ni un maximum local, ni un minimum local.
3. $(x,y)=(0,0)$ est un point selle et $(x,y)=(1,1)$ est un minimum local.
4. $(x,y)=(0,0)$ est un minimum local, $(x,y)=(0,\pm\sqrt{3}/3)$ est un point selle et $(x,y)=(\pm\sqrt{2}/2,0)$ est un maximum local.

II.5.22 Si $F(x,y,z):=x^2+y^3+xyz^3-3$, on a $\dfrac{\partial F}{\partial y}(1,1,1)=4\neq 0$.
Le polynôme de Taylor d'ordre 2 de φ est :
$$1-\frac{3}{4}h-\frac{3}{4}k-\frac{31}{64}h^2-\frac{27}{32}hk-\frac{39}{64}k^2$$
avec $h=x-1$ et $k=z-1$.

II.5.23 1. Fermée et exacte.
2. Fermée et exacte.
3. Fermée et exacte.
4. Non fermée, non exacte.

II.5.24 $U=\mathbb{R}^2\setminus\{(0,0),(-1,0)\}$. On pourra calculer $\displaystyle\int_{C(0,r)}\omega$ selon les valeurs de r.

II.5.25 Utiliser les relations de Cauchy-Riemann.

II.5.26 1. L'aire de la cardioïde est $3\pi/2$.
2. L'aire de la boucle du folium de Descartes est $3a^2/2$.
3. L'aire de l'astroïde est $3\pi/8$.

II.6 : Fonctions analytiques

II.6.1 La fonction $f:\mathbb{R}\to\mathbb{R}$ définie par $f(x)=e^{-1/x^2}$ si $x\neq 0$ et $f(0)=0$ est $\mathcal{C}^\infty$ mais n'est pas analytique en 0.

II.6.2 Utiliser la formule de Taylor-Lagrange.

II.6.3 $f(z)=\displaystyle\sum_{n\geqslant 0} z^n/(2n)!$. Les zéros de f sont les points de la forme $(\pi/2+k\pi)^2$ avec $k\in\mathbb{N}$.

II.6.4 1. On pourra montrer que si $f(z):=\displaystyle\sum_{n\geqslant 0} b_n z^n$, alors
$a_n=-\alpha^{-n-1}\displaystyle\sum_{k=0}^{n} b_k\alpha^k$ et donc que $a_n\sim f(\alpha)/\alpha^n$.
2. Appliquer la question précédente avec $f(z):=\dfrac{z+z_0-\alpha_0}{P(z+z_0)}$ et $\alpha:=\alpha_0-z_0$.

II.6.5 $\displaystyle 1+\sum_{p=1}^{+\infty}\frac{z^{3p}}{(2\cdot 3)\cdots((3p-1)3p)}$ et $\displaystyle z+\sum_{p=1}^{+\infty}\frac{z^{3p+1}}{(3\cdot 4)\cdots(3p(3p+1))}$ forment une base de l'espace des solutions. Leurs rayons de convergence sont infinis. Toute autre solution est combinaison linéaire à coefficients constants de ces deux fonctions. (S'inspirer du module sur les équations différentielles, en particulier la notion de système fondamental de solutions.)

II.6.6 Développer l'exponentielle en série entière et justifier la permutation de la série et de l'intégrale.

II.6.7 On pourra choisir pour D la réunion des disques de convergence des développements en série entières au voisinage de tous les points de I.

II.6.8 Il n'y a que les fonctions constantes.

II.6.9
1. On pourra commencer par montrer que $\{m\tau_1 + n\tau_2 \mid m, n \in \mathbb{Z}\}$ est dense dans $\mathbb{R}$.
2. Non à cause de l'existence de fonctions analytiques périodiques.

II.6.10 L'ensemble des $2\pi\dfrac{k}{2^n}$, $k, n \in \mathbb{N}$, est dense dans $\mathbb{R}$.

II.6.11 La deuxième condition dit que la distance de $z \in U$ à la frontière de U est $\geqslant 1/n$.

II.6.12
1. Il y a 3 composantes connexes.
2. $\operatorname{Log}(z_1 z_2) = \operatorname{Log}(z_1) + \operatorname{Log}(z_2) + C$ avec $C = 0$, $C = -2\pi$ o u $C = 2\pi$ suivant la composante connexe de Ω où (z_1, z_2) se trouve.

II.6.13 Dériver les fonctions concernées.

II.6.14 Dériver les fonctions concernées.

II.6.15
$$F^{-1}(w) = w + \frac{1}{3}w^3 + \frac{2}{15}w^5 + \mathrm{O}(w^7).$$

II.6.16
1. Utiliser l'identité $F^n \circ G = w^n$.
2. En tant que dérivée d'une fonction, $\dfrac{(F^n)'}{F^p}$ a un résidu nul en 0 pour $p > n$.
3. Les $b'_{k,n} = \dfrac{n}{k} c_{k-n,k}$ sont solutions du même système que les $b_{k,n}$.

II.6.17 Considérer par exemple $x \mapsto 1/(1 + x^2)$.

II.6.18 Appliquer le principe du module maximum.

II.6.19 Appliquer le principe du module maximum dans un disque $D(0, r)$ où $r \nearrow 1$.

II.6.20 On pourra considérer la fonction $g : U \to \mathbb{C}$ définie par $g(z) = \exp\big(f(z)\big)$.

II.6.21
1. Utiliser le théorème des zéros isolés.
2. C'est un produit de fonctions analytiques.
3. Commencer par factoriser par le produit des $(z - a_j)^{m_j}$.
4. Invoquer l'exercice 11 de la page 790.

II.6.22 On pourra montrer que f est bornée sur $\mathbb{C}$.

II.7 : Équations différentielles

II.7.1 Écrire $x = \sqrt{E_0}\sin u$, où u est de classe $\mathcal{C}^1$; on a alors $\dot{u}^2 = 1$.

II.7.2 Par exemple, la fonction égale à $-\sqrt{E_0}$ sur $]-\infty, -\pi/2]$, à $\sqrt{E_0}\sin$ sur $[-\pi/2, \pi/2]$ et à $+\sqrt{E_0}$ sur $[\pi/2, +\infty[$ est solution ; donc pas d'unicité.

II.7.3 Prendre $f(t) = e^{-t^{-2}}$, $f(0) = 0$ et $g(t) = \pm f(t)$, selon le signe de t.

II.7.4 Soit p le plus petit entier tel que le wronskien de p quelconques parmi les f_i est identiquement nul sur J ; quitte à réordonner les indices, on peut donc supposer que $W_{p-1}(f_1, \dots, f_{p-1})$ est non identiquement nul, donc non nul sur un intervalle $J' \subset J$. On peut alors appliquer l'exercice 4 de la page 809 à $f_1, \dots, f_{p-1}, f_i$.

II.7.5 On écrit $x = \lambda e^{\mathrm{i}t} + \mu e^{-\mathrm{i}t}$, ce qui donne $\dot{\lambda} = (b/2\mathrm{i})e^{-\mathrm{i}t}$ et $\dot{\mu} = (-b/2\mathrm{i})e^{\mathrm{i}t}$, puis on invoque les conditions initiales.

II.7.6 Pour $\omega = 1$, l'équation $\dot{y} + \mathrm{i}y = e^{\mathrm{i}t}$ se résoud comme dans le texte ; mais ensuite la résolution de l'équation $\dot{x} - \mathrm{i}x = \frac{e^{\mathrm{i}t} - e^{-\mathrm{i}t}}{2}$ fait apparaître un terme en $te^{\mathrm{i}t}$.

II.7.7 Unicité : deux solution diffèrent par une solution de $y' - ky = 0$.

II.7.8 Poser $g := f + f'$ et résoudre $f + f' = g$.

II.7.9 Non (penser aux zéros isolés).

II.7.10 Par la première méthode : une solution particulière est un polynôme de degré 1, la solution générale comporte une combinaison linéaire de e^t, te^t. Par la seconde méthode : le seul terme de la décomposition en éléments simples est $\dfrac{1}{(x-1)^2}$.

II.7.11 Par la première méthode : utiliser la décomposition en éléments simples :

$$\frac{1}{X^2+1} = \frac{(-1/2\mathrm{i})}{X+\mathrm{i}} + \frac{(1/2\mathrm{i})}{X-\mathrm{i}}.$$

Par la deuxième méthode : si $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, alors

$$e^{At} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

II.7.12 On pourra vérifier que le membre droit *n'est pas solution* de (S_0).

II.7.13 Si B est une matrice telle que $BB' = B'B$, alors de $(e^B)' = e^B B'$.

II.7.14 Pour la réciproque, considérer $\begin{pmatrix} 0 & g \\ f & 0 \end{pmatrix}$.

II.7.15 (ii) Invoquer le théorème des fonctions implicites (module II.5). (iii) Puisque B' commute avec B, les espaces propres de B sont stables par B' ; comme ce sont des droites, $B'U_i = \mu_i U_i$. En dérivant $BU_i = \lambda_i$, on voit que l'on peut remplacer U_i par $f_i U_i$ qui est conctant si $f_i' + f_i\mu_i = 0$.

II.7.16 Posant $\mathcal{X}(t) := R(t, t_0)$, écrire $\mathcal{X}(t) - I_n = \displaystyle\int_{t_0}^{t} A(s)\mathcal{X}(s)\,\mathrm{d}s$ et invoquer le critère de Cauchy pour la convergence de l'intégrale lorsque $t \to +\infty$.

II.7.17 S'inspirer de l'exercice 21 de la page 842.

On trouve les fonctions de $x^2 - \dot{x}^2$.

II.7.18 La solution générale (sans conditions aux limites) est facile à exprimer.

II.7.19 Il faudra distinguer deux cas selon que $\pm \mathrm{i}\omega$ est ou non solution de l'équation caractéristique.

II.7.20 Le changement $t \mapsto -t$ transforme $\dot{I}$ en $-\dot{I}$ sans affecter I et $\ddot{I}$.

II.7.21 On peut soit vectorialiser et invoquer le critère sur les valeurs propres de la matrice d'un système ; soit utiliser la description explicite des solutions de la section 1.2.

II.7.22 Écrire l'équation satisfaite par y et faire disparaitre le terme en y'.

II.7.23 C'est une conséquence directe de (ii).

II.7.24 C'est une conséquence directe de la dichotomie de la question (iv).

II.7.25 Considérer le wronskien de f et d'une fonction trigonométrique bien choisie.

II.7.26 Étudier $f^2 + \dfrac{f'^2}{q}$.

II.7.27 Poser $g(u) := f(e^u)$.

II.7.28 Utiliser la méthode de variation des constantes.

II.7.29 Ils doivent être de module inférieur à 1.

II.7.30 Appliquer le corollaire 26 de la page 828 à l'équation différentielle satisfaite par $R(t,t_0)^{-1} R_\varepsilon(t,t_0)$.

II.7.31 Sur un intervalle plus petit, la limite au bord serait bornée et le lemme de sortie des compacts (théorème 54 de la page 861) s'appliquerait.

II.7.32 Introduire $G(t) := \displaystyle\int_{t_0}^{t} B(s)g(s)\,\mathrm{d}s$ et $C(t) := \exp \displaystyle\int_{t_0}^{t} B(s)\,\mathrm{d}s$.

II.7.33 Ordonner l'ensemble des couples (J,F), où F est une solution sur J du problème de Cauchy en (t_0, Y_0).

II.7.34 Utiliser le lemme de Gronwall.

II.7.35 Étudier la « fonctionnelle » Φ qui à f associe $\Phi(f) : t \mapsto 1 + \displaystyle\int_0^t f^2(s)\,\mathrm{d}s$.

II.7.36 On a $f(t,t_0,0) = 0$. Si $y_0 \neq 0$, on cherche $y = \dfrac{1}{a - t^2}$ avec $y_0 = \dfrac{1}{a - t_0^2}$, ce qui détermine a ; puis déterminer le plus grand intervalle contenant t_0 sur lequel de dénominateur de $f(t,t_0,y_0) = \dfrac{y_0}{1 + y_0(t_0^2 - t^2)}$ ne s'annule pas.

II.7.37 Chercher les solutions dans les deux zones $y > t$, $y < t$, puis envisager des recollements.

II.7.38 Soit b la borne supérieure de l'intervalle de définition de la solution maximale f. Lorsque t tend vers b, f' est bornée ; si $b < +\infty$, f y admet une limite et l'on peut reposer le problème de Cauchy en b.

II.7.39 (i) Commencer par le cas où $f(t_0) < g(t_0)$ et raisonner comme dans la preuve du théorème 56 de la page 862.

II.7.40 (ii) Par symétrie, se ramener aux résultats sur $y' = y^2 - t$. (iii) À droite de t_0, la courbe croît ; elle ne peut rester confinée dans un compact si $\beta \leqslant 0$. (iv) Lorsque $y \geqslant 0$, f' est croissante et $f' \geqslant f'(0) = y_1^2$. (v) Les trajectoires ne peuvent se couper. (vi) Lemme de Gronwall.

II.7.41 Invoquer selon le cas le lemme de sortie des comppacts (théorème 54 de la page 861) ou le théorème de comparaison 56 de la page 862.

II.7.42 Dans le cas contraire, on peut trouver $t_1 < t_2 < t_3$ dans I tels que, par exemple $x(t_1), x(t_3) < x(t_2)$; à l'aide du théorème des valeurs intermédiaires, en déduire une contradiction.

II.7.43 Si $|\ddot{\rho}| \leqslant M$ et si $\dot{\rho}(t_0) \leqslant \alpha < 0$, alors $\dot{\rho} \leqslant \alpha/2$ sur $\left[t_0 - \dfrac{|\alpha|}{2M}, t_0 + \dfrac{|\alpha|}{2M}\right]$.

II.7.44 On a $U(x) = E_0 + \dfrac{1}{2}\alpha(x-x_0)^2 + \mathrm{o}\big((x-x_0)^2\big)$; on peut donc appliquer un changement de variable u tel que $U = E_0 + \dfrac{1}{2}u^2(x)$.

II.7.45 Utilser le fait que $\dot{\theta}^2 - \cos\theta$ décroît.

II.7.46 Il faut utiliser l'estimation $|\sin\theta - \theta| \leqslant |\theta|^3/6$.

II.7.47 On a $\dot{r}^2/2 - C/r = E$ car $\overline{U}(r) = U(r) = -C/r$. Cette équation peut être résolue explicitement.

II.7.48 La période de rotation vaut :
$$T = \int_{-\pi}^{\pi} \frac{dt}{d\theta}\,\mathrm{d}\theta = \int_{-\pi}^{\pi} \frac{p^2\ \mathrm{d}\theta}{J(1+e\cos\theta)^2}.$$

II.7.49 Tracer les courbes $xye^{-(x+y)} = C$ pour $C > 0$. La périodicité vient de ce que le système est autonome.

II.7.50 Étudier $f := \rho^2 - \lambda$.

II.7.51 (i) Les courbes intégrales ne se croisent pas. (ii) On doit trouver des cycles ou des spirales.

II.8 : Méthodes numériques

II.8.1 Majorer $\sum_{j\geqslant 1} d_j b^{-j}$ en précisant à quelle condition la borne est atteinte.

II.8.2 Prendre pour x_{n+1} la partie entière de $\dfrac{1}{2}\left(x_n + \dfrac{B}{x_n}\right)$, où $B := b^{2p}$.

II.8.3 On trouve : cb^k, c^2b^{2k} et e^{cb^k}.

II.8.4 Une convergence linéaire.

II.8.5 Comparer les logarithmes.

II.8.6 Utiliser l'égalité $(n+1)I_nI_{n+1} = \dfrac{\pi}{2}$ (voir preuve du lemme 2 de la page 891). Le terme dans la limite au membre droit de la formule de Wallis est $\dfrac{1}{n+1}\left(\dfrac{4^n}{\binom{2n}{n}}\right)^2$.

II.8.7 On suppose $a < b$ et $f(a) < 0 < f(b)$. Définie récursivement, $Z(u,v,\varepsilon)$ doit rendre $w := \dfrac{u+v}{2}$ si $v - u < \varepsilon$ ou si $f(w) = 0$; $Z(u,w,\varepsilon)$ si $f(w) > O$; et $Z(w,v,f)$ si $f(w) < 0$.
On appelle $Z(a,b,\varepsilon)$.

II.8.8 On a $f(0) > 0 > f(\sqrt{2})$. Ici $f' < 0$ et $f'' > 0$ sur l'intervalle ouvert, on pose donc $x_0 := \sqrt{2}$. On itère avec la formule $x_{i+1} := \dfrac{3}{6 - x_i^2}$.

II.8.9 Prendre $t := 1/p$.

II.8.10 Une racine dans $[-2,-1]$, une racine dans $[-1,0]$, une racine dans $[1,2]$ (par étude des variations).
On peut itérer $x \mapsto \dfrac{-1+x^5}{5}$ à partir de 0.

II.8.11 Examinons le cas simplifié où $\varphi(x) = x + \lambda x^k$. On veut avoir $|\varphi'| < 1$ au voisinage de 0, où $\varphi'(x) = 1 - \lambda x^{k-1}$, ce qui donne les conditions k impair et $\lambda < 0$.

II.8.12 Points $-1, 0, 1$: l'erreur effective est $\left\|x^4 - x^2\right\|_\infty = 0{,}25$; l'erreur théorique est $\dfrac{8}{3\sqrt{3}} \approx 1{,}54$. Points de Tchebychev : l'erreur effective est $\left\|x^4 - 3x^2/4\right\|_\infty = 0{,}25$; l'erreur théorique est 1.

II.8.13 On ne comptera pas le coût des évaluations $f(a_i)$. La formule (9) demande $n(n+1)/2$ soustractions, n additions, $(n+1)$ divisions et $(n+1)(n-1)$ multiplications. Le programme itératif donné à la fin de la section 3.1.2 exécute $n(n-1)/2$ fois 2 soustractions et 1 division.

II.8.14 L'intégrale vaut $\ln(1+\alpha^2) - 2 + 2\alpha \arctan \dfrac{1}{\alpha}$. Pour appliquer le théorème, on utilise l'encadrement :
$\ln(t^2) \leqslant \ln(\alpha^2 + t^2) \leqslant \ln(1+t^2)$.

II.8.15 Pour ce dernier, montrer qu'à chaque dérivation $\sum \mu_i$ diminue de 1 en traitant séparément racines simples et multiples.

II.8.16 Rapprocher cette question de l'interpolation de Lagrange-Hermite.

II.8.17 L'ordre exact est 3.

II.8.18 On trouvera des poids négatifs.

II.8.19 On a $I = \dfrac{\pi}{4} \approx 0{,}785$. Point milieu : $I \approx 0,8$, erreur $0{,}015$, erreur théorique $0{,}021$. Trapèzes : $I \approx 0{,}75$, erreur $0{,}035$, erreur théorique $0{,}042$. Simpson : $I \approx 0{,}753$, erreur $0{,}032$. Gauß : $I \approx 0{,}785$, erreur $0{,}00013$.

II.8.20 On a 4 paramètres à déterminer, on peut espérer l'ordre 3. Seule la méthode de Gauß convient.

II.8.21

$$\int_{-1}^{1} f(x)\frac{\mathrm{d}x}{\sqrt{1-x^2}} \approx \frac{\pi}{n+1}\sum_{j=0}^{n} f\left(\cos\frac{2j+1}{2n+2}\pi\right).$$

II.8.22 Penser au produit scalaire $\displaystyle\int_0^1 fg$ sur $\mathbb{R}[X]$.

II.8.23 On a le majorant :

$$\frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)\,\|\delta A\|\,/\,\|A\|}\left(\frac{\|\delta B\|}{\|B\|}+\frac{\|\delta A\|}{\|A\|}\right).$$

II.8.24 Utiliser les formules de l'exercice I.3.14 de la page 140 pour obtenir une formule de récurrence pour le déterminant.

II.8.25 Utiliser l'exercice I.3.14 de la page 140.

Le coût est en $\mathrm{O}(n)$.

II.8.26 Une matrice de rayon spectral nul est nilpotente.

II.8.27 Si $\rho(J)=0$.

II.8.28 Changer l'indice k en $n+1-k$.

Index

A

Abel
  règle d'—, 621
  théorème radial d'—, 506
  transformation d'—, 490
absolument convergente (série —), 431
accélération centrale, 871
accroissements finis
  inégalité des —, 604
  théorème des —, 602
Ackermann-Peter (fonction d'—), 390
action
  d'un groupe, 58
  de groupes, 180
  fidèle, 59
  simple, 60
  simplement transitive, 68
  simplement 3-transitive, 298
  transitive, 68
adhérence, 405
  valeur d'—, 402
adhérent (point —), 405
adjoint d'un endomorphisme, 161, 199
affine
  groupe —, 64, 269
  transformation —, 64
Alembert-Gauß (théorème de —), 789
algèbre
  de Banach, 443
  K—, 26
  normée, 443
  norme dans une —, 55
  propriété universelle de l'— des polynômes, 29
  sous— engendrée par un élément, 28
  sur un corps, 26
algébrique
  élément —, 29
  nombre —, 343
algébriquement
  indépendants (éléments —), 316
  liés (éléments —), 316
algorithme
  de Babylone, 904
    pour les séries formelles, 41
  de Cholesky, 379
  de Mc Naughton-Yamada, 388
  de Warshall, 387
  QR, 172
alphabet, 355
alterné(e)
  application —, 105
  polynôme —, 349
analytique (fonction —), 766
anneau(x)
  de fractions, 9
  des germes de fonctions continues, 50
  extension d'un —, 22
  local, 53
  sous— engendré, 22
antihermitien(ne)
  endomorphisme —, 201
  matrice —, 187
antisymétrique
  application —, 105
  endomorphisme —, 169
  forme bilinéaire —, 111
Apollonius (théorèmes d'—), 178

application
alternée, 105
antisymétrique, 105
bilinéaire, 104
de Floquet, 846
en escalier, 550
linéaire induite par passage au quotient, 17
multilinéaire, 104
p-linéaire, 104
réglée, 590
semi-linéaire, 188
transposée, 99
approchée (unité —), 671
approximation polynomiale, 909
arbre binaire, 374
feuilles d'un —, 374
nœuds d'un —, 374
principe d'induction structurelle pour un —, 374
sous-arbres d'un —, 374
taille d'un —, 374
arbre(s)
arité d'un nœud d'un —, 374
correspondance de rotation pour les —, 375
descendance d'un sommet dans un —, 373
enraciné, 372
général, 373
hauteur d'un —, 374
infini, 373
niveau d'un nœud d'un —, 374
nœud externe d'un —, 374
nœud interne d'un —, 374
nœuds d'un —, 373
profondeur d'un —, 374
taille d'un —, 366, 374
Arden (lemme d'—), 359
aréolaire (vitesse —), 872
arêtes d'un graphe, 363
argument
d'un nombre complexe, 775
détermination de l'—, 775
arité d'un noœud dans un arbre, 374
arithmético-géométrique (inégalité —), 457
Arzela (théorème d'—), 632
attractif (point fixe —), 902
auto-adjoint (endomorphisme —), 169, 201
automate
calcul dans un —, 369
des parties, 370
des résiduels, 370
déterminisé d'un —, 371
états finaux d'un —, 368
états initiaux d'un —, 368
fini, 368
complet, 369
déterministe, 369
déterministe complet, 369
langage reconnu par un —, 369
relation de transition d'un —, 368
automorphisme
intérieur, 10
orthogonal, 165
unitaire, 200
autonome (système différentiel —), 865

B

Babylone (algorithme de —), 904
Banach
algèbre de —, 443
espace de —, 423
bandes (matrices —), 937
barrière
inférieure, 879
supérieure, 879
base
adaptée à un d-drapeau, 267
directe, 266
duale, 100
hilbertienne, 463
indirecte, 266
orthonormée, 145, 197
Bernstein (théorème de —), 529

Bertrand (séries de —), 618
Bessel (inégalité de —), 152, 198, 664
Beta (fonction —), 615
bidual, 97
bilinéaire (application —), 104
birapport, 300
bloc de Jordan, 253
Bolzano-Weierstraß (propriété de —), 432
Borel-Lebesgue (propriété de —), 435
bornée
    application —, 400
    partie —, 400
boucle, 367
bouts (théorème des —), 858
Burnside (formule de —), 87

C

calcul
    numérique, 881
    symbolique, 817
Cantor (ensemble de —), 399, 436, 471, 473, 567, 596
caractéristique (sous-espace —), 242
Catalan (nombres de —), 386, 389
Cauchy
    critère de —, 423, 430
    critère de — uniforme, 479, 490
    formules de —, 784
    inégalités de —, 787
    problème de —, 851
    règle de —, 504
    suite de —, 422
    théorème de —, 74
Cauchy-Lipschitz
    théorème de —, 859
    théorème de — linéaire scalaire, 830
    théorème de — linéaire vectoriel, 829
Cauchy-Schwarz (inégalité de —), 130, 194
Cayley (théorème de —), 61
Cayley-Hamilton (théorème de —), 230
central (élément —), 28
Cesáro (convergence au sens de —), 507
chaleur (équation de la —), 690
champs de vecteurs, 838
    point critique d'un —, 839
    point singulier d'un —, 839
    radial, 872
changement de variable, 607, 622
Chasles
    relation de —, 552, 560, 746
    relations de — pour la résolvante d'un système différentiel, 831
chinois (théorème —), 21
Cholesky
    algorithme de —, 379
    décomposition de —, 133, 195, 939
    méthode de —, 133
cinquième postulat d'Euclide, 305
clothoïde, 833, 915
codimension, 18
coeff, 662
coefficients
    binomiaux, 47
    de Fourier, 662
col, 741
commutateur, 265
compact (espace métrique —), 432
compagnon (matrice — d'un polynôme), 228
comparaison série-intégrale, 617
complet (espace métrique —), 423
composante connexe, 450
    par arcs, 452
composition des séries formelles, 44
concave (fonction —), 452
conditionnée (matrice mal —), 933
conditionnement, 933
conditions aux limites, 949
conditions initiales d'un système différentiel, 823
cône
    de lumière, 296
    isotrope, 115, 194
congruence, 6
    modulo un idéal, 14
    modulo un sous groupe, 11
congruentes (matrices —), 111

conjugaison, 60, 66, 113
classes de —, 66
conjugués
éléments —, 66
mots —, 385
sous-groupes —, 70
vecteurs —, 113
connexe (espace métrique —), 448
connexe par arcs (espace métrique —), 451
conservation de l'énergie, 801, 868, 872
consistance, 918
constante
d'Euler, 618
de Lebesgue, 909
continue
application —, 409
application uniformément —, 414
continuité d'une fonction définie par une intégrale, 642
contractante (application —), 410
contragrédiente (représentation —), 308
contravariantes (coordonnées —), 148
convergence
conditionnelle, 903
d'ordre α, 890
domaine de —, 501
dominée, 629, 631
encadrée, 631
linéaire, 889, 894, 902, 942
monotone, 630
normale, 492
quadratique, 884, 890, 902, 904
rayon de —, 501
simple, 475, 489
uniforme, 477, 489
convergente
méthode —, 918
série —, 429
série absolument —, 431
convexe
combinaison —, 445
enveloppe —, 445
fonction —, 452
partie —, 444
convolution (produit de —), 670
coordonnées
contravariantes, 148
covariantes, 148
homogènes, 271
courbe
de l'autoroute, 833
fermée, 744
paramétrée, 744
simple, 744
Cousin (lemme de —), 591
covariantes (coordonnées —), 148
critère
de Cauchy, 423, 430
uniforme, 479, 490
de Cauchy pour les intégrales impropres, 619
de Darboux, 573
critique (point —), 728, 739
crochet de dualité, 95
curviligne (intégrale —), 744
cycle, 79
cylindre de sécurité pour une équation différentielle, 852

D

Darboux
critère de —, 573
sommes de —, 570
théorème de —, 572
décomposition
d'Iwasawa, 309
de Cholesky, 133, 195, 939
de Dunford, 249, 250
de Dunford multiplicative, 279
de Gauß, 121
en carrés, 120
LU, 121, 937
polaire complexe, 309
polaire réelle, 283
QR, 161, 939
QR complexe, 309
QR d'une matrice carrée à coefficients réels, 173
spectrale, 234

defini(e) positif(ve)
endomorphisme —, 171, 202
forme quadratique —, 129
dégénérée (forme bilinéaire —), 109
degré
d'un polynôme, 23
partiel d'un polynôme, 319
total d'un polynôme, 319
demi-plan de Poincaré, 301
dense (partie —), 406
déplacement, 286
dérivabilité d'une fonction définie par une intégrale, 644
dérivation
d'une série formelle, 36
de la K-algèbre $K[X_1, \ldots, X_n]$, 324
dérivé(e)
d'une série entière, 514
groupe —, 265
partielle, 730
seconde, 740
dérivée
partielle, 705
dérivé(e)
partielle, 325
déterminant, 105
de Gram, 154
mineur principal, 121
détermination
de l'argument, 775
du logarithme, 775
déterminisé d'un automate, 371
développement à l'infini
série de Laurent, 35
développement d'une fraction rationnelle
série de Laurent, 33
série formelle, 34
diagonalisable
endomorphisme —, 231
matrice —, 231
diagonalisation simultanée de deux formes quadratiques, 184
diamètre, 400
dichotomie, 893
différences
divisées, 907
formule de Neville-Aitken, 908
formule de Newton, 907
finies, 905, 910
différentielle totale, 743
dilatation, 262
Dini
deuxième théorème de —, 481
premier théorème de —, 481
directe
base —, 266
somme — de sous-espaces, 214
somme — orthogonale, 146
directes (méthode dites — de résolution d'un système linéaire), 932
Dirichlet
noyau de —, 675
théorème de —, 675, 686
discrète (partie —), 407
discriminant, 338, 341
polynôme —, 314
distance, 143
d'un point à un sous-espace de dimension finie, 152
entre deux parties, 400
sur un ensemble, 397
distingué (sous-groupe —), 9
dominant
coefficient — d'un polynôme, 23
terme — d'un polynôme, 23
dominée (convergence —), 629, 631
drapeau, 88, 267
maximal, 267
droite hyperbolique, 303
dual(e)
base —, 100
d'un espace vectoriel, 95
dualité, 95
forme et vecteur orthogonaux, 98
Dunford
décomposition de —, 249, 250
multiplicative, 279
Dyck (langage de —), 354, 386, 389

E

écriture flottante
erreurs d'arrondi, 886
nombre de chiffres significatifs, 886
norme (**IEEE 754**), 885
égalité
de la médiane, 143
de Parseval, 673, 685
du parallélogramme, 143
élémentaire
sous-ensemble —, 564
élimination (et résultant), 342
emboîtés (segments —), 426
encadrée (convergence —), 631
endomorphisme
adjoint d'un —, 161
antihermitien, 201
autoadjoint, 201
diagonalisable, 231
hermitien, 201
défini positif, 202
positif, 202
induit, 224
normal, 203
scindé, 220
symétrique, 169
défini positif, 171
positif, 171
trigonalisable, 238
entier algébrique, 349
entière
fonction —, 788
série —, 500, 762
enveloppe, 344
convexe, 445
équation
aux classes, 65, 83
caractéristique, 811
d'un hyperplan, 97
de la chaleur, 690
de Liouville, 863
de Lotka-Volterra, 873
de Riccati, 862
de Van der Pol, 874
intégrale, 826
équation différentielle
à coefficients constants, 811
à coefficients périodiques, 846
conditions
aux limites pour une —, 834
initiales d'une —, 808, 851
continuité des solutions en fonction des conditions initiales d'une —, 860
courbe intégrale d'une —, 838, 850
cylindre de sécurité pour une —, 852
espace des phases d'une —, 836, 838, 850
explosion des solutions d'une —, 860
flot d'une — linéaire autonome, 845
homogène, 804
intégrale d'une —, 804
intégrale première d'une — linéaire autonome, 842
linéaire scalaire d'ordre n, 804
linéaire vectorielle d'ordre 1, 820
méthode d'Euler pour la résolution approchée d'une —, 855
méthode de variation des constantes pour résoudre une —, 809
méthode des coefficients indéterminés pour résoudre une —, 814
non homogène, 804
portant sur des séries formelles, 37
portrait de phases d'une — linéaire autonome, 837
résolution par quadrature d'une —, 809
solution d'une —, 804, 850
solution globale d'une —, 804, 851
solution maximale d'une —, 851
solutions approchées d'une —, 854
solutions continues et de classe $\mathcal{C}^1$ par morceaux d'une —, 852
sous forme résolue, 850
stabilité des solutions d'une — linéaire, 843
système fondamental de solutions d'une — homogène, 808
trajectoire d'une solution d'—, 836, 850

vectorialisation d'une —, 819
vectorielle d'ordre 1, 850
équivalentes (norme —), 421
ergodique (théorème —), 693
erreur, 909–912
de consistance, 917
escalier
application en —, 550
fonction en —, 550
espace
affine euclidien, 286
de Banach, 423
de Hilbert, 460
euclidien, 142
hermitien, 196
métrique, 397
compact, 432
complet, 423
connexe, 448
connexe par arcs, 451
projectif, 270
cartes d'un —, 271
tangent, 276, 287, 724
vecteur —, 276
espace vectoriel
normé, 394
préhilbertien complexe, 196, 660
préhilbertien réel, 142
espace-temps, 297
étoilée (partie —), 445
étoile
canonique, 279
d'un langage, 358
lemme de l'—, 371
propre d'un langage, 385
Euclide, 141
cinquième postulat d'—, 305
euclidienne
norme —, 142, 143
Euler
constante d'—, 618
fonction Gamma d'—, 498
formule d'—, 325
méthode d'—, 917
méthode d'— pour la résolution approchée d'une équation différentielle, 855
multiplicativité de l'indicatrice d'—, 54
exponentielle, 774
d'une série formelle, 45
de matrices, 276, 279
—polynôme, 813
exposant, 885
expression algébrique, 383
écriture postfixe d'une —, 383
expressions régulières, 361
extension (théorème d'—, ou de relèvement), 342
extremum, 728

F

factoriel (anneau —), 321
Faddeev-Le Verrier (méthode de —), 241
famille
normale, 792
φ-orthonormée, 114
fausse (méthode de la — position), 898
Féjer
noyau de —, 677
théorème de —, 678, 687
fermée
courbe —, 744
forme différentielle —, 747
partie, 403
filtrage par profils (*pattern matching*) en programmation récursive, 382
fixe (théorème du point —), 427
Floquet
application de —, 846
flot d'une équation linéaire autonome, 845
flottante
calcul en représentation —, 884
écriture — normalisée, 885
fonction
à variation bornée, 574
analytique, 766
Beta, 615
concave, 452

convexe, 452
d'Ackermann-Peter, 390
d'Airy, 654
de Morris, 390
en escalier, 550
entière, 788
Gamma, 498, 614, 634, 643, 645
holomorphe, 709, 785
polynomiale, 317, 335
rationnelle, 317
récursive, 375, 380
réglée, 590, 680
théorème des —s implicites, 708
fonctionnelle (programmation —), 376
forêt, 373
forme(s)
bilinéaire, 104
antisymétrique, 111
dégénérée, 109
non dégénérée, 109
rang d'une —, 111
symétrique, 111
différentielle, 743
différentielle exacte, 744
différentielle fermée, 747
hermitienne, 191
définie positive, 193
positive, 193
signature d'une —, 205
linéaire, 95
multilinéaire, 104
multilinéaire alternée, 105
p-linéaire, 104
polaire, 116, 191
d'une forme hermitienne, 191
d'une forme quadratique, 116
quadratique, 116
définie positive, 129
non dégénérée, 116
positive, 129
signature d'une — réelle, 132
quadratique définie, 116
sesquilinéaire, 187
formule(s)
de Burnside, 87
de Cauchy, 784
de Green-Riemann, 750
de la moyenne, 569, 609
de Parseval, 465, 673, 685
de Stirling, 890
de Taylor, 326, 608, 735
de Taylor avec reste intégral, 608, 736
de Taylor-Lagrange, 738
de Taylor-Young, 738
de Wallis, 951
du binôme généralisée pour les séries formelles, 46
Fourier
coefficients de —, 662
série de —, 662, 783
fraction(s) rationnelle(s)
corps des — en les n indéterminées $X_1, \ldots, X_n$, 314
forme irréductible d'une —, 322
homogène, 347
Fresnel (intégrales de —), 623, 656
Frobenius (théorème de —), 87
frontière (d'une partie), 407
Fubini
théorème de —, 485, 486, 648

G

Gamma, 793
fonction —, 498, 614, 634, 643, 645
Gauß
décomposition de —, 121
formule de —, 634
intégrale de —, 634
méthode de quadrature de —, 929
somme de —, 259
Gauß-Seidel (méthode de — de résolution d'un système linéaire), 943
géodésique, 303
géométrie
hyperbolique, 304
non euclidienne, 305

Gibbs
phénomène de —, 688
gradient, 147, 727
Gram
déterminant de —, 154
matrices de —, 154, 198
Gram-Schmidt (orthogonalisation de —), 155, 198
graphe, 362
acyclique, 366
arêtes d'un —, 363
biparti, 363
biparti complet, 363
chemin élémentaire d'un —, 364
chemin simple d'un —, 364
circuit dans un —, 364
circuit hamiltonien d'un —, 366
circuits euleriens d'un —, 365
complet, 363
composantes connexes d'un —, 364
cycle dans un —, 366
extrémités d'une arête dans un —, 363
induit, 363
liste d'adjacence d'un —, 387
liste d'adjacence d'un — non orienté, 387
matrice d'adjacence d'un —, 364
non orienté, 363
racine d'un —, 372
sommet isolé d'un —, 363
sommets adjacents d'un —, 364
sommets connectés d'un —, 364
sommets d'un —, 363
sommets voisins d'un —, 364
graphe orienté, 367
arc incident en un sommet d'un —, 367
chemin élémentaire dans un —, 368
chemin simple dans un —, 368
complet, 367
étiqueté, 369
fortement connexe, 387
induit, 367
matrice d'accessibilité d'un —, 368
matrice d'adjacence d'un —, 368
prédécesseur d'un sommet dans un —, 368
sommet accessible dans un —, 368
successeur d'un sommet dans un —, 368
Green-Riemann (formule de —), 750
Gronwall (lemme de —), 857
groupe
affine, 269
classique, 261
d'isotropie, 68
dérivé, 265
de Lorentz, 295, 306, 310
de Poincaré, 297
des fractions d'un monoïde, 9
des isométries affines, 286
des symétries, 286
des symétries directes, 286
du tétraèdre, 291
orthogonal, 165, 166, 280
orthogonal d'un espace euclidien, 167
projectif, 264, 270
simple, 292
spécial linéaire, 11
spécial orthogonal, 167
spécial projectif, 264
spécial projectif réel, 301
spécial unitaire, 200
unitaire d'un espace hermitien, 200, 280

H

Hadamard
formule d'—, 505
inégalité d'—, 158
Hanoï (tours de —), 379
Haussdorf (distance de —), 439
hermitien(ne)
endomorphisme —, 201
espace —, 196
forme —, 191
matrice —, 187
norme —, 196
Hessienne, 740

Hilbert, 141
cube de —, 399
espace de —, 460
matrice de —, 934
hilbertienne
base —, 463
Hölder (inégalité de —), 458, 459
holomorphe (fonction —), 709, 767, 785
homéomorphisme, 413
homogènes
coordonnées —, 271
polynômes —, 322
homographie, 271
de la droite projective, 296
homothétie (affine), 64
Householder
matrice de —, 173
méthode de —, 175, 939
hyperplan, 96

I

idéal(aux), 18
bilatère, 52
de $A[X]$, 25
engendré par une partie, 19
étrangers, 21
homogène, 347
maximaux, 19
premiers, 19
produit d'—, 19
racine d'un —, 54
somme d'—, 19
impair (sommet —), 363
implicitation de paramétrage, 344
impropre (intégrale —), 611
indicatrice (multiplicativité de l'— d'Euler), 54
indice de nilpotence, 246
indirecte (base —), 266
inductive (définition —), 354
inégalité(s)
arithmético-géométrique, 457
d'Hadamard, 158
de Bessel, 152, 198, 664
de Cauchy, 787
de Cauchy-Schwarz, 130, 194
de Hölder, 458, 459
de Minkowski, 131, 458, 459
des accroissements finis, 604
isopérimétrique, 695
inertie (théorème d'— de Sylvester), 131
infini (point à l'—), 298
initiale
forme — d'un polynôme, 323
intégrable
absolument —, 619
au sens de Riemann, 556
localement —, 600
intégrale(s)
curviligne, 744
de Cauchy, 580
de Fresnel, 623, 656
de Gauss, 634
de Stieltjes, 578
de Wallis, 891
impropre, 611
inférieure, 554
itérée, 829
semi-convergente, 620
supérieure, 554
intégration
des relations de comparaison, 624
par parties, 608
intérieur
d'une partie, 405
point — à une partie, 405
interpolation
de Lagrange-Hermite, 907, 952
linéaire, 895, 898
nœuds d'—, 906
polynomiale, 905
de Lagrange, 905
valeurs d'—, 906
invariant (sous-groupe —), 10
inversion (théorème d'— locale), 706, 779
inversion approchée (algorithme d'—), 883
irréductible
représentation induite —, 268

isobarycentre, 73
isocline, 863
isométrie, 286, 411
  affine directe, 286
  affine positive, 286
isopérimétrique (inégalité —), 695
isotrope
  cône —, 115, 190, 194
  vecteur —, 115, 190
itération d'un langage, 358
itérative
  méthode — de résolution d'un système linéaire, 941
  programmation —, 376
Iwasawa (décomposition d'—), 309

J

Jacobi
  méthode de — de résolution d'un système linéaire, 942
jacobien, 704
jauge, 591
Jordan
  bloc de —, 253
  mesure de —, 566
  réduction de —, 254
  réduite de —, 254
  théorème de —, 253

K

Kepler (loi de —), 871
Kleene (théorème de —), 371
König (lemme de —), 373
Kronecker (lemme de —), 350
Krull (théorème de —), 20

L

lacunaire (série —), 696
Lagrange-Hermite (interpolation de —), 907, 952
langage(s), 357
  concaténation de —, 358
  de Dyck, 354, 386, 389
  étoile d'un —, 358
  étoile propre d'un —, 385
  itération d'un —, 358
  principe d'induction structurelle pour les — rationnels, 361
  propre, 358
  rationnels, 361
  reconnaissable, 369
  reconnu par un automate, 369
  réguliers, 361
  résiduel d'un —, 359
Lebesgue
  constante de —, 909
  théorème de —, 597
Leibniz (règle de —), 324
lemme
  d'Arden, 359
  d'Ogden, 371
  de Gronwall, 857
  de König, 373
  de l'étoile, 371
  de Schur, 308
  de sortie des compacts, 861
  des noyaux, 221
limite
  d'une suite de série formelle, 39
  supérieure, 504
linéaire (représentation —), 267
Liouville
  équation de —, 863
  théorème de —, 788
lipschitzienne (application —), 410
localisation des racines d'un polynôme réel, 895
logarithme
  complexe, 774
  d'une matrice inversible complexe, 277
  d'une série formelle, 45
  détermination du —, 775
loi(s)
  de Kepler, 871
  de Moore, 932
  relativiste d'addition des vitesses, 310
longueur, 564
Lorentz (groupe de —), 295, 306, 310

Lotka-Volterra (équation de —), 873
LU (décomposition (—), 121
LU (décomposition —), 937

M

mantisse, 885
Maschke (théorème de —), 269
matrice(s)
adjointe, 187
antihermitienne, 187
bandes, 937
compagnon d'un polynôme, 228
congruentes, 111
conjuguée, 187
de Gram, 154, 198
de Hilbert, 934
de Householder, 173
diagonale par blocs, 218
diagonalisable, 231
facilement inversible, 942
hermitienne, 187
définie positive, 195
positive, 195
normale, 204
orthogonale —, 168
orthogonalement semblables, 280
scindée, 220
séparables, 335
symétriques réelles
définies positives, 133
positives, 133
transconjuguée, 187
triangulaire par blocs, 218
tridiagonales, 946
trigonalisable, 238
unitaire, 201
unitairement semblables, 280
matricielle (norme —), 443
maximal (idéal —), 19
maximum, 729
Mc Naughton-Yamada (algorithme de —), 388
McCarthy
fonction de —, 495
mesure
de Jordan, 566
méthode(s)
d'Euler, 917
de Gauß-Seidel de résolution d'un système linéaire), 943
de Householder, 175, 939
de Jacobi
de relaxation d'un système linéaire, 945
de résolution d'un système linéaire, 942
de la sécante, 899
de quadrature de Gauß, 929
de quadrature de Newton-Cotes, 923
de quadrature de Simpson, 924
de quadrature des rectangles à droite, 921
de quadrature des rectangles à gauche, 920
de quadrature du point milieu, 921
directes, 889
dites directes de relaxation d'un système linéaire), 932
erreur de troncature dans une — itérative, 889
itérative, 889
itératives de relaxation d'un système linéaire), 941
performances des — itératives, 889
milieu (méthode de quadrature du point —), 921
minimal (polynôme —), 29
minimum, 729
Minkowski (inégalité de —), 131, 458, 459
moment cinétique, 871
Monge (notations de —), 741
monoïde, 355
commutatif, 9
groupe des fractions d'un —, 9
libre sur un alphabet, 355
monômes en n indéterminées, 315
monotone (convergence —-), 630
Montel (théorème de —), 793

Moore (loi de —), 932
morphisme de K-algèbres associé à la substitution $X_i \mapsto x_i$, 316
Morris (fonction de —), 390
mot(s), 355
  concaténation de deux —, 355
  conjugués, 385
  facteur d'un —, 357
  facteur droit d'un —, 357
  facteur gauche d'un —, 357
  facteur itérant d'un —, 371
  longueur d'un —, 355
  miroir d'un —, 355
  palindrome, 385
  préfixe d'un —, 357
  recombinaison d'un —, 385
  suffixe d'un —, 357
  vide, 355
moyenne
  deuxième formule de la —, 609
  première formule de la —, 569
moyennisation, 268
multiindice, 315
  poids du —, 321
  poids pondéré du —, 330
multilinéaire (application —), 104
multiplicativité de l'indicatrice d'Euler, 54
multiplicité
  d'un zéro, 771
  d'une valeur propre, 224
musicale (notation —), 147, 148

N

négligeable
  sous-ensemble —, 596
Neville-Aitken (formule de —), 908
Newton
  formule de —, 907
  formules de —, 332
  méthode de —, 903
  somme de —, 329
Newton-Cotes (méthode de quadrature de —), 923
nilpotence (indice de —), 246
nilradical
  d'un idéal, 21
nombres
  caractéristiques, 846
  de Catalan, 42, 389
normal(e)
  convergence —, 492
  endomorphisme —, 203
  matrice —, 204
  sous-groupe —, 10
norme(s), 143
  de la convergence uniforme, 477
  de Schur, 193
  équivalentes, 421
  euclidienne, 142, 143
  hermitienne, 196
  matricielle, 443
  subordonnée, 443
  sur un espace vectoriel, 394
notations de Monge, 741
noyau
  d'une forme bilinéaire symétrique, 114
  d'une forme hermitienne, 194
  d'une forme sequilinéaire, 190
  de Dirichlet, 675
  de Féjer, 677
  de Peano associé à une méthode de quadrature, 926
  lemme des —x, 221

O

Ogden (lemme d'—), 371
orbites, 65
ordre
  d'un polynôme, 321
  d'une méthode de quadrature, 920
  d'une série entière, 513
  d'une série formelle, 31
  de multiplicté d'un zéro, 771
orientation
  canonique, 266
  d'un espace vectoriel réel de dimension finie, 266

orthogonal(e), 145
automorphisme —, 165
groupe —, 165, 166
matrice —, 169
projecteur —, 150, 198
représentation —, 285
somme directe —, 146
sous-espace —, 109, 146
supplémentaire —, 146
symétrie —, 150
système —, 144
orthogonalisation (de Gram-Schmidt), 155, 198
orthogonalité
de dualité, 98
forme et vecteur orthogonaux, 98
vecteurs φ -orthogonaux, 109
orthonormé(e)
base —, 145, 197
système —, 144
oscillateur harmonique, 800
ouverte (parties —), 403

P

pair (sommet —), 363
palindrome, 385
paramétrage, 723
Parseval
égalité de —, 465, 673, 685
partie
fermée, 403
multiplicative d'un anneau, 9
ouverte, 403
partiel (stratégie du pivot —), 938
pas d'une subdivision, 855
pattern matching, 382
pavé, 336
Peano (noyau de — associé à une méthode de quadrature), 926
pendule
non linéaire non amorti, 870
simple, 800
période propre, 815
permutation (nombre moyen d'inversions d'une —), 345
perpendiculaires (sous-espaces —), 146
phases (portrait de — d'un système différentiel autonome), 865
phénomène
de Gibbs, 688
de Runge, 913
pivot (stratégies de —), 937, 938
p-linéaire alternée
forme —, 106
poids, 919
Poincaré
demi-plan de —, 301
groupe de —, 297
théorème de —, 748
point(s)
à l'infini, 298
critique, 728, 739
d'accumulation, 772
de Tchebychev, 912
fixe
attractif, 902
pour une action, 68
répulsif, 902
superattractif, 902
théorème du —, 427
isolé, 772
selle, 741
polaire
forme — d'une forme quadratique, 116
polarisation (identités de —), 116
polynôme(s)
alterné, 349
anneau (ou K-algèbre) des — en les n indéterminées $X_1, \ldots, X_n$, 314
annulateur, 29
coefficient dominant d'un —, 23
composante homogène d'un —, 323
de Taylor, 735
degré d'un —, 23
forme initiale d'un —, 323
générateur d'une permutation, 345
homogènes, 322

minimal, 29
orthogonaux, 155
propriété universelle de l'algèbre des —, 29
substitution de —, 318
sur un anneau commutatif, 22
symétriques, 329
symétriques élémentaires, 314, 329
terme dominant d'un —, 23
trigonométrique, 660
polynomiale (fonction —), 317, 335
portrait de phases d'un système différentiel autonome, 865
positif
endomorphisme hermitien —, 202
endomorphisme symétrique —, 171
préhilbertien
complexe (espace vectoriel —), 196, 660
réel (espace vectoriel —), 142
premier (idéal —), 19
presque partout, 597
primitive, 605
généralisée, 605
principe
de prolongement des identités algébriques, 337
des zéros isolés, 771
du module maximum, 789
du prolongement analytique, 769
problème
d'extremum, 205
de Cauchy, 851
produit
de convolution, 670
de deux séries entières, 510
infini de série formelle, 56
scalaire, 142, 196
canonique de $\mathbb{R}^n$, 112
programmation
effets de bord en — impérative, 376
fonctionnelle, 376
itérative, 376
pile en — récursive, 377
récursive, 376
projecteur
orthogonal, 150, 198
spectral, 234, 243
projectif(ve)
espace —, 270
groupe —, 264, 270
transformation —, 270
projection canonique, 4
prolongement
principe de — des identités algébriques, 337
principe du — analytique, 769
propre (langage —), 358
Pythagore (théorème de —), 145, 197

Q

QR
algorithme —, 172
décomposition —, 161, 939
quadratique (forme —), 116
quadrature(s), 915
composées, 919
convergence des méthodes de —, 924
élémentaires, 919
erreur d'une formule de —, 921
erreur dans une méthode de —, 925
méthode de —
de Gauß, 929
de Newton-Cotes, 923
de Simpson, 924
des rectangles à droite, 921
des rectangles à gauche, 920
des trapèzes, 921
du point milieu, 921
méthodes de —, 919
noyau de Peano associé à une méthode de —, 926
ordre d'une méthode de —, 920
quaternions, 292

quotient(s)
anneaux —, 14
application linéaire induite par passage au —, 17
d'un ensemble par une relation d'équivalence, 3
de Rayleigh, 184, 206
espaces vectoriels —, 17
groupe —, 9
loi —, 6
passage au — d'une loi de composition interne, 6
passage au — d'une application, 5
propriétés héritées par la loi —, 7

R

racine(s)
carrée d'une série formelle, 42
d'un idéal, 54
d'un nombre complexe, 777
radial (théorème — d'Abel), 506
radical
d'un idéal, 21
rang
d'une famille d'hyperplans, 97
d'une forme bilinéaire, 111
d'une forme sesquilinéaire, 190
rationnel(le)
fonction —), 317
langage —), 361
Rayleigh (quotient de —), 184, 206
rayon de convergence, 501
réciproque d'une série formelle, 45
reconnaissable (langage —), 369
reconnu (langage — par un automate), 369
rectangles
méthode de quadrature des — à droite, 921
méthode de quadrature des — à gauche, 920
rectification, 915
récursif
analyse d'algorithme —, 380
prédicat héréditaire d'un programme —, 381
programme — élémentaire, 380
programme — bien fondé, 381
récursion, 375
et induction structurelle, 382
mutuelle, 378
récursive
filtrage par profils (*pattern matching*) en programmation —, 382
fonction, 375
fonction —, 380
programmation —, 376
réductible (représentation induite complètement —), 268
réduction modulo un idéal, 14
référentiel, 297
réflexion, 168
règle
d'Abel, 621
de dédoublement, 117
réglée
application —, 590
fonction —, 590, 680
regula falsi, 898
réguliers (langages —), 361
relation
algébrique, 316
compatible avec une application, 5
compatible avec une loi de composition interne, 6
compatible avec une loi externe, 17
de Chasles, 746
différentielle algébrique, 316
Relativité (théorie de la — restreinte), 297
relaxation (méthode de — de résolution d'un système linéaire), 945
renversement, 281
représentant, 4

représentation
contragrédiente, 308
induite, 268
complètement réductible, 268
irréductible, 268
simple, 268
linéaire, 267
orthogonale, 285
unitaire, 285
répulsif (point fixe —), 902
résiduel d'un langage, 359
résiduels (automate des —), 370
résolution
approchée, 892
méthode de Gauß-Seidel de — d'un système linéaire, 943
méthode de Jacobi de — d'un système linéaire, 942
méthodes de relaxation de — d'un système linéaire, 945
méthodes dites directes de — d'un système linéaire, 932
méthodes itératives de — d'un système linéaire, 941
résolvante d'un système différentiel, 828
résonance, 814
paramétrique, 849
reste intégral (formule de Taylor avec —), 608, 736
résultant, 338, 339
et élimination, 342
Riccati (équation de —), 862
Riemann
fonction zêta de —, 497, 792
sommes de —, 577, 582, 590
sphère de —, 301
Riesz (théorème de —), 463
Runge
phénomène de —, 913

S

saturé, 51
scalaire
produit —, 142, 196
Schur
lemme de —, 308
norme de —, 193
Schwarz (théorème de —), 731
sécante (méthode de la), 899
segments emboîtés, 426
selle (point —), 741
semblables (matrices —), 213
semi-linéaire (application —), 188
séparable (espace métrique —), 407
série
absolument convergente, 431
convergente, 429
de Fourier, 662, 783
de Laurent
développement à l'infini, 35
développement d'une fraction rationnelle, 33
formelle, 33
de Taylor, 522
entière, 500, 762
lacunaire, 696
somme d'une —, 429
somme d'une — entière, 501
trigonométrique, 660
série(s) formelle(s), 30
algorithme de Babylone pour les —, 41
composition des —, 44
dérivation d'une —, 36
développement d'une fraction rationnelle, 34
équation différentielle portant sur des —, 37
exponentielle d'une —, 45
formule du binôme généralisée pour les —, 46
limite d'une suite de —, 39
logarithme d'une —, 45
ordre d'une —, 31
produit infini de —, 56
réciproque d'une —, 45
racine carrée d'une —, 42
terme constant d'une —, 31

théorème des fonctions implicites pour les —, 43
théorème du point fixe pour les —, 43
valuation d'une —, 31
sesquilinéaire (forme —), 187
signature
d'une forme hermitienne, 194, 205
d'une forme quadratique réelle, 132
simple
convergence —, 475, 489
courbe —, 744
groupe —, 292
représentation induite —, 268
simplicité de $SO_3(\mathbb{R})$, 292
Simpson (méthode de quadrature de —), 924
Smith-Volterra-Cantor
ensemble de —, 565, 598, 651
somme(s)
d'une série, 429
d'une série entière, 501
de Darboux, 570
de Gauß, 259
de Riemann, 577, 582, 590
directe de sous-espaces, 214
directe orthogonale, 146
sommet
impair, 363
pair, 363
sommets d'un graphe, 363
sortie (lemme de — des compacts), 861
sous-ensemble
négligeable, 596
sous-espace(s)
orthogonal, 109, 146
perpendiculaires, 146
vectoriel stable, 267
sous-groupe
distingué, 9
invariant, 10
normal, 10
sous-variété, 718
spectral (théorème —), 181, 204
spectre, 224
sphère
de Riemann, 301
unité d'un espace euclidien, 284
spirale de Cornu, 833
stabilisateur, 68, 69
stabilité, 909, 910, 912, 913, 933
constante de —, 918
des solutions d'équations différentielles, 843
stable
méthode —, 918
partie —, 69
sous-espace —, 224
Stieltjes
intégrale de —, 578
Stirling (formule de —), 890
Stokes (théorème de —), 743
stratégie du pivot
partiel, 938
total, 938
Sturm (suite de —), 895
subdivision, 549
subordonnée (norme —), 443
suite
de carré sommable, 142
de Cauchy, 422
de fonctions analytiques, 791
superattractif (point fixe —), 902
supplémentaire orthogonal, 146, 198
support (d'une permutation), 78
surfaces, 712
Sylvester
matrice de —, 338
théorème d'inertie de —, 131, 195
symétrie orthogonale, 150
hyperplane, 168
symétriques
endomorphismes —, 169
formes bilinéaires —, 111

système différentiel
à coefficients constants, 823
admettant un cycle limite, 873
autonome, 865
autonome non dissipatifs à un degré de liberté, 867
conditions initiales d'un —, 823
flot d'un — autonome, 866
homogène, 820
intégrale première d'un — autonome, 866
linéaire de rang n, 820
méthode de variation des constantes pour résoudre un —, 823
période des petites oscillations d'un — autonome à un degré de liberté, 880
point critique — autonome, 866
portrait de phases d'un — autonome, 865
potentiel d'un — autonome à un degré de liberté, 867
relations de Chasles pour la résolvante d'un —, 831
résolvante d'un —, 828
solution matricielle d'un —, 822
solutions matricielles fondamentales d'un —, 822
système proie-prédateur, 874

T

tangent
espace —, 287, 724
vecteur —, 275
Taylor
formule de —, 608
formule de — avec reste intégral, 608, 736
formules de —, 326, 735
polynôme de —, 735
série de —, 522
Taylor-Lagrange (formule de —), 738
Taylor-Young (formule de —), 738
Tchebychev
points de —, 912
terme constant d'une série formelle, 31
tétraèdre régulier, 291
théorème
chinois, 21
d'Apollonius, 178
d'Arzela, 632
d'inertie de Sylvester, 131, 195
d'inversion locale, 706, 779
de Cauchy-Lipschitz, 859
de Cauchy-Lipschitz linéaire scalaire, 830
de Cauchy-Lipschitz linéaire vectoriel, 829
de Cayley, 61
de Cayley-Hamilton, 230
de d'Alembert-Gauß, 789
de Darboux, 572
de Dirichlet, 675, 686
de Féjer, 678, 687
de Frobenius, 87
de Fubini, 485, 486, 648
de Jordan, 253
de Kleene, 371
de Krull, 20
de Lebesgue, 597
de Liouville, 788
de Maschke, 269
de Montel, 793
de Poincaré, 748
de Pythagore, 145, 197
de Riesz, 463
de Schwarz, 731
de Stokes, 743
de Weierstraß, 487, 672
des accroissements finis, 602
des bouts, 858
des fonctions implicites, 708
pour les séries formelles, 43
du point fixe, 427
du point fixe pour les séries formelles, 43
ergodique, 693
fondamental, 606, 607
spectral, 181, 204

théorie de la Relativité restreinte, 297
total(e)
famille —, 463
stratégie du pivot —, 938
tours de Hanoï, 379
trace
d'un endomorphisme, 223
d'une matrice carrée, 222
transcendant (élément —), 29
transconjuguée (matrice —), 187
transformation
affine, 64
d'Abel, 490
projective, 270
transitive (action —), 68
translation, 64
transposée (application —), 99
transposition, 99
transvection, 262
triangle hyperbolique, 303
trigonalisable
endomorphisme —, 238
matrice —, 238
trigonométrique
polynôme —, 660
série —, 660
troisième cas d'égalité des triangles, 309

U

uniformément continue (application —), 414
uniforme
convergence —, 477, 489
critère de Cauchy —, 479, 490
unitaire
automorphisme —, 200
groupe —, 200
matrice —, 201
représentation —, 285
unité (approchée), 671

V

valeur d'adhérence, 402
valuation
d'un polynôme, 321
d'une série formelle, 31
Van der Pol (équation de —), 874
variation
bornée, 574
vecteur(s)
conjugués, 113
de genre espace, 296
de genre temps, 296
isotrope, 115, 296
orthogonaux, 144, 196
φ -orthogonaux, 109
tangent, 275
vectorialisation d'une équation différentielle, 819
vitesse aréolaire, 872
voisinage, 407

W

Wallis
formule de —, 951
intégrale de —, 891
Warshall (algorithme de —), 387
Weierstraß
théorème d'approximation de —, 487, 672
théorème de —, 487, 672
wronskien, 806, 822

Z

Zariski-dense, 337
zéros
principe des — isolés, 771
zêta
fonction —, 497
fonction — de Riemann, 497, 792